Università degli Studi di Milano-Bicocca

DEMS: Department of Economics, Quantitative Methods and Corporate Strategies

ISTEI: Sezione di Economia e Gestione delle Imprese

Doctor of Philosophy in Marketing and Management

VALERIO VEGLIO

PHD THESIS

Academic Year 2012-2013

Web Data Mining to monitoring Marketing Performance. Focus on Potential Customers Risk of Churn.

Tutor: Professor Marisa Civardi (University of Milano-Bicocca)

Co-Tutor: Dr. J. D Lamb (University of Aberdeen)

Research study conducted at Aberdeen University Business School

Management Department

Aberdeen – Scotland – United Kingdom

to my parents…

Abstract

The global economics crisis shows as business forecast are often inexact and as companies survival is only guaranteed by a continuous monitoring of the market players. Nowadays, the main goal for global organizations is to understand how cut down the gap between current and expected value for getting a competitive advantage.

In this scenario, due to the large amount of data available, traditional statistical models could be not enough to predict future events. In fact, the advent of globalization and the dramatic advance in information technology have completely changed the way of business is carried on. They have made the process of collecting data easier adding to volume of data available to business. The accessibility of large volume of data on customers has created new opportunities besides challenges for organizations to leverage data and gain competitive advantage. As consequence, marketing decision makers feel increased pressure because competition between companies has intensified and customers can obtain information more easily.

Global companies have realized that the knowledge collect in huge databases is the key to supporting the various business decisions, in particular for marketing function. Special attention is paid on the Web Marketing field. But, much of this useful knowledge is hidden and untapped. There are two main reasons why organisations have not tapped the information. First, information is sometimes poorly managed. Second, many organisations are either not aware that it is possible to develop powerful statistical data processing tools or do not know how to use them. Appropriate data mining tools, which are good at extracting and identifying valuable information and knowledge from enormous databases, is one of the best supporting approach to make different marketing decision. In fact, analyzing and understanding in advance the customer behavior represents the foundation for developing winning marketing strategies.

This PhD thesis shows the strategic magnitude of the predictive data mining models in today competitive landscape for discovering hidden knowledge collected in huge databases in order to maximize the probability of customer conversion and minimize their risk of churn. The main challenge for decision makers is to discover those customer are likely to churn. In particular, the attention has been paid on the main data mining techniques helpful to forecast the potential customers risk of churn within global organizations with an outside-in perspective in the web marketing field. The database analyzed was provided by a global company which develop web analytics services all over the world.

The main objective of this research is show that the hierarchical logistic regressions and classification trees joined with an opinion mining approach are accurate predictive data mining model useful to identify the main marketing drivers generated by potential customers before purchasing an online service. The criteria base on the loss functions such as the percentage correctly classified at the ‘economically optimal’ cutoff purchase probability (PPC) and the receiver operating characteristic (ROC) curve prove that hierarchical logistic regression models outperform decision tree models in predicting the potential customer risk of churn. Finally, a what-if scenario related to the probability of the customer conversion conclude this dissertation.

Acknowledgments

My sincere gratitude and deepest thanks go to Professors Silvio M. Brondoni to achieve me the opportunity to attend the Doctoral of Philosophy in Marketing and Management at Milano-Bicocca University. I also wish to thank Professor Margherita Corniani for her amazing market-driven management lectures.

A particular thanks go to my Supervisor Professor Marisa Civardi. She has guided me during these long and arduous but lovely experience. She taught me excellent practices and skills. She also gave me an optimal statistical background that I will use in my academic carrier.

A special thanks go to Dr. John D Lamb. I am very thankful to him for his insightful comments and advices during my visiting period at the Business School of Aberdeen. I really appreciate his help to improve the quality of my thesis.

Last but not least, my heartfelt thanks and greatest gratitude go to my parents. I am very grateful to them for their warm support and encouragement along the doctoral journey and always. I hope will make them proud of my achievements as I am proud of them. Their love will always be with me wherever I go!

Table of Contents

List of Tables

List of Figures

List of Graphics

'Data analysis is a tool for extracting the jewel of truth from the slurry of data'

[Jean-Paul Benzécri]

'All models are wrong but some are useful'

[George E.P. Box]

Chapter 1
Introduction

1.1. Research problem and research contributions to the knowledge

In global markets the main challenge for companies is to identify accurate models and methods to predict precise business performance. In fact, an essential part of managing any organization is planning for the future. Nowadays, the long-run success of global firms is closely related to how well management is able to anticipate the future and draw up appropriate strategies. Good judgement, intuition, and an awareness of the state of the economy may give a manager a rough idea or 'feeling' of what is likely to happen in the future.

According to Daniel Kahneman (Nobel in Economics for the work in Prospect Theory, 2002) it is obvious as qualitative approaches are not enough for predicting accurate and significant corporate performance in today global business. For this reason, data mining analysis has became an astonishing approach is so far the meaningful knowledge is often hidden in enormous databases and most traditional statistical methods could fail to uncover such knowledge.

Bueren, Schierholz, Kolbe and Brenner (2004) define the knowledge as a strategic intangible resource at the base of competitive advantages. Besides, the most important type of knowledge would be appearing to be customer knowledge. In fact, an efficient utilization of customer knowledge determines the development of global firms. This is particularly true in marketing area because of the proliferation of e-customer data collected in huge databases. Indeed, 'in over-supply markets, companies have acknowledged that their marketing strategies should focus on identifying those customers who are likely to churn' (Hadden, Tiwari, Roi and Ruta, 2005 p. 2903).

In literature, churn management is defined as a set of techniques that enable firms to keep their profitable customers and it aims at increasing customer loyalty (Lejeune, 2001). Additionally, enterprises can be identifying two groups of churners: voluntary or attrition and non-voluntary or forced churn. Non-voluntary churners are easier to detect because are the customers who have had their service or product withdrawn by the company. Instead, voluntary is more difficult to determinate because this type of churn occurs when customers make a conscious decision to terminate their service with the company (Hadden, Tiwary, Roy and Ruta, 2005).

This PhD dissertation proposes an extension of the customer churn definition focusing the attention on the concept of voluntary churners in digital contests. In this case a churner is a potential customer that did not buy the service online. In addition, this research investigates the effectiveness of churn methods in web marketing database without sampling the data. With this approach decision makers should be able to first predict the churn rate and to design appropriate marketing strategies for preventing it.

Nowadays, one of the main tools able to help marketers in the mentioned approach is data mining (Khak Abi and Glolamain, 2010). Prediction of behaviour, customer value,

customer satisfaction and customer loyalty are example of some of the information that can be extracted from the data, that should already be stored within a company's database. However, to perform such a complex analysis of the information is necessary to either purchase commercial software or implement a solution based on the many data mining techniques that have been developed for this purpose.

The key goal for global corporations is to discover knowledge in vast database and to make sense out of the data in order to improve the future accuracy of marketing performance. In particular, an appropriate use of data mining models combined with a subjective human judgment could help to improve the future accuracy of marketing performance in order to raise the number of customers.

The current economic crisis has showed as the market predictions are inaccurate. In fact, the business intelligence tools are one of the main priorities for the Chief Information Office Director in every industry. But, despite this, due to a limited knowledge in this field, corporations in order to forecast business performance fall into huge traps obtaining catastrophic results. Businesses often consider mathematical and statistical models too complex, inaccurate, expensive and very hard to elaborate the final results. On the contrary, some predictive data mining models are simple to deduce and really accurate. Today, as never before, there is an emerging need to create a "bridge" between academic and marketing managerial reality.

According to the lack of a comprehensive literature review about the application of data mining techniques in churn prediction, an overview of the existing literature about this topic is provided in this PhD thesis. It analyses 100 or so papers related to the most common data mining techniques in managing the customer churn. In details, we have tried to adapt the customer churn model proposed in literature to potential customers for estimating both the risk of churn and the customer conversion probability. On the basis of these theoretical perspectives, two wider research questions drive this PhD dissertation:

1. Why marketing forecasts are often inaccurate in today global business?
2. Data mining models are better than traditional statistical models in predicting both the churn and customer conversion probability?

These broad research questions can be split up in four assumptions:

1. Today, there is an emerging need to create a link between marketing and statistic area for improving the accuracy of marketing performance thus to draw up a new marketing research field.
2. The communication in terms of knowledge related to the development of new predictive methodologies between academic and managerial reality is almost zero.
3. Computational Methods for Data Mining such as Hierarchical Logistic Regression and Classification Decision Trees are the main models to predict the probability of customer conversion within global companies.

4. The Criteria based on the Loss Functions are the best methods for evaluating the Computational Data Mining Models in today's competitive landscape.

After presenting the research problems and the respective assumptions and before showing the research methodology, the following sections clarify the most important theoretical stance of this dissertation.

1.2. Background of the PhD thesis

The following sub-sections aim at presenting some theoretical underpinnings on which this PhD dissertation is based. Specifically, they briefly present a general overview of global markets with a particular attention on the structural changes as a consequence of the globalization and on the role of the data mining techniques able for predicting future marketing performance. Secondly, it focuses the attention on the market-driven management strategy underling the importance of intangible assets in today global business. Thirdly, it argues the relevance of the knowledge management in global companies with an outside-in perspective focusing the attention on the management of the competitive customer knowledge. Finally, it defines the main differences between statistical data analysis and data mining analysis. Special attention is paid on the origins of the Data Mining and on its strategic importance from a managerial standpoint.

1.2.1. Global Markets: A General Overview

The phenomenon of globalization pervades every aspect of the business, breaking down the barriers that hinder the transfer of goods, services, capitals, resources, information and technologies between countries. In particular, it raises the level of interdependence between markets and increases financial, commercial and cultural exchanges.

The globalization has led to a radical exchange the market features. Brondoni (2008) observes that in global business the market-space is not a given variable of the decision-making process but a factor of competition, whose profile is determined by the action/reactions of businesses and governments (market-space competition). In addition, Rancati (2010 p. 77) notes that 'this transforms the organizations within the competition space of the relations and transactions of any company that focuses on time as a competitive driver (time-based competition)'.Globalization has changed the traditional time and space relations that hold up the competition in the market and has led to the development of a competitive convergence through the fusion between different and really distant competitive environment.

The origins of convergence related to Chamberlin's study. 'Competitive convergence does not cancel companies original area of activities, but incorporates them into a broader competitive environment with a dual goal: to reduce costs (builds on a common inventory, information and logistic base, reduces cost of contacting customers and other interaction costs; virtual communities can help reduce the cost of company supplied information and support) and to achieve a potential increase in profits (reaches both cyber and traditional segments; achieves synergies between the online and offline business that promote sales; increase access to the business, anytime, anywhere, premium services such as customization, home delivery and choice to increase the stickiness and allow companies to charge a premium price)' (Rosen, 2009 p. 25).

In this scenario, global companies compete directly or indirectly into a competitive ecosystem (Moore, 1996; Lambin, 2008; Rancati, 2010) that is defined as 'a complex group of companies and consumers, suppliers, competitors, distributors, specifies and partners who benefit mutually from each other' (Manning & Thorne, 2003 p.76) in which the geographical or bureaucratic boundaries of the companies (if they exist) are overlapped and confused, fluid and dynamic. In other words, it is clear as new competitive boundaries among companies are established from the globalization and from global economies.

'In global economies, the role of strategic marketing is more important than ever. It remains the best mechanism to balance demand and supply and it also triggers a virtuous circle of economic and social development, strengthened today by the social, cultural and technological changes observed on the market' (Lambin 2002, p.15). Also, in order to improve business performance it becomes really important to join the concept of strategic marketing with the concept of data mining analysis so to draw a new research field.

According to Brondoni (2009), in global managerial economies coexist three competitive market conditions:

1. Scarcity of Supply (D>S). The managerial economics are focused on price competition in local markets (monopoly markets). In this contest the availability of data is limited thus decision makers, in order to provide accurate business forecast, can adopt basic analysis tools based on query and reporting system.

2. Demand and Supply in Dynamic Balance (D≈S). This economic contest reveals widespread internationalisation and non-price competition policies (static oligopoly markets and controlled competition). In this case, the volume of data tends to rise and decision makers to predict accurate business performance can be implement statistical data analysis tools based only on the past value data of variable. For instance, to forecast the quantity of sale, they can adopt predict models based on time series methods (single and double moving average, and exponential smoothing). This methods are based solely on past values of the variable and/or on past forecast errors. Tools based on query and reporting system are just used to support the statistical data analysis. Adopting only these static tools could seem too much limited in order for forecasting future market performance. Also, the exogenous variables-factors generated by the market environment does not affecting the analysis. Besides, businesses could be use sampling techniques and inferential methods to simulate market predictions.

3. Oversupply (D<S): the managerial economics emphasises the central role of corporate intangible assets. Particular attention is paid on the knowledge management, especially on the hidden knowledge. Due to the globalization and the dramatic advances in technological sector and in e-business sector, the availability of the data for companies it becomes inestimable. For this reason, the statistical data analysis is not adequate to predict accurate business performance. Additionally, marketers cannot use sampling techniques and inferential methods to simulate future scenario. These tools are too much limited. In this scenario to draw a sample of data is too much restricting because a lot of information will be missing. In fact, the

> World Wide Web (WWW) is the main huge database for global organizations and it can defined as a dynamic and interactive database able to contain infinite volumes of data. As consequence, data mining analysis are one of the main approaches in order to predict the future business events. This approach is really useful in marketing area because of the current global business is dominated by disloyalty customer behaviour thus it is really difficult to predict future markets events. In additions, the business forecast could be influenced by exogenous variables. In most case these variables must be treatment with qualitative methods because they cannot be quantified and historical data are either not applicable or unavailable. For this reason, data mining models become an astonishing approach; it allows marketers to integrate both quantitative and qualitative methods such as regression analysis and textual analysis. Data mining helpful and supports decision makers in order to get the best choices for the company's success. In particular, data mining draws upon ideas, such as: (1) sampling, estimation, and hypothesis testing from statistics and (2) search algorithms, modelling techniques, and learning theories from artificial intelligence, pattern recognition, and machine learning. Also, data mining has been quick to adopt ideas from other areas, including optimization, evolutionary computing, information theory, signal processing, visualization, and information retrieval.

According to Brondoni (2009, p. 4) 'the globalization of the markets make it necessary to abandon the competitive reference to a close domain, coinciding with particular physical or administrative context (a product category, a country region, a geographical area, a sector, etc.)'. In fact, before the development of globalization, corporations competed in a given and stable over time sector of activity characterized by well-defined and delimited administrative and territorial boundaries, in which whose structure influences business strategies (Rancati, 2010).

Due to the total homogeneity of the product insight the sector of activity and on the radical difference in the products from two different sectors the space of competition is very restricted, controlled and limited. Competition only exists if there is rivalry between companies. Organizations for achieving a business advantage must develop long run corporate strategies and to invest in tangible assets that predominate on intangible assets. Also, because of the limited amount of data and the competition space well-delimited, in order to implement winner marketing strategies and to increase the level of competitiveness, businesses were used statistical data analysis and query and reporting tools. In particular, to estimate the competitiveness of the sector firms can use measures of concentration such as Gini, Herfindahl, and Linda, the Lorenz Curve, Vertical Integration and Economies of Scale (Rancati, 2010).

It is clear as in global markets these statistical measures are inadequate and inaccurate to predict business performance and the organizations competitiveness level.

In fact, due to the high level of market dynamism, it is impossible to take a real picture of the market in a given period. In close domains, companies can achieve a leading position in the market especially through two main strategies: the cost leadership or product differentiation strategies. In the first case companies offer identical products at low costs than competitors while in the second case organizations offer products with peculiar characteristics that diverge from those proposed by a competitors. On the

contrary, with the current wave of globalization – started in the 1980s – the business scenario is completely changed.

The actual wave of globalization has been driven in many cases by significant technological advances in computing, communication and transportation, and perhaps also the internationalization of business corporations in search for new customers, cheaper research and skilled labour.

The development of information technologies as network, database and e-business (Gordijin, Akkermans and Villet, 2001; Kakousis, Paspallis and Papadopoulous, 2010; Piao, Hanc and Wu, 2010) has rapidly modified the competitive scenario. In fact, the advent of globalization, the virtualization of network economics and the e-business market have become larger and more complicated than ever.

The major e-business pressures is an effective way to use business intelligence (Gong, Muyeba and Guo, 2010; Luhn, 1958; Petrini and Pozzebon, 2009; Trikman, McCormack, Oliveira and Ladeira, 2010), which requires enterprises to use data mining tools (Han and Kamber, 2001; Hastie, Tibshirani and Friedman, 2001; Hsu and Wallace, 2007; Jaamour, 2005) to analyze and to manage business data.

Data Mining is an interactive and dynamic data-driven approach that can discover useful hidden knowledge from enormous databases. It has been considered a very useful and relevance methodology for data analysis in every sector and industry (Chen, Wei, Liu and Wets, 2002). In fact, the wide availability of data and the dramatic advances in information and communications technology have made it easy for organizations to collect a wealth of data about their customers. The data can be collected through information systems and stored in vast corporate databases. But it must be turned into knowledge to be useful an important tools for company's growth thus to get a competitive value.

Global organizations have realized that the knowledge available from their huge databases is the key to supporting various business decisions and remaining competitive. But much of this knowledge remains untapped.

There are two main reasons why organisations have not tapped the knowledge. First, knowledge is sometimes poorly managed. Second, organisations are either not aware that it is possible to use powerful statistical data processing tools or do not know how to use them (Giudici, 2009). This latter reason emphasizes the importance of the first research question proposed in this PhD thesis. Also, decision makers feel increased pressure because competition between companies has intensified and customers can obtain information more easily. Global companies can respond to this by developing market-driven strategies with a particular focus on the competitive customer value. In fact, knowledge about customers is a critical intangible asset for marketing functions in today's global business. But, global firms need to know how to turn data into knowledge. The need to do this has led to a rapid expansion in the use of business intelligence tools, especially predictive data mining models. A significant and relevant data mining analysis is the first step in order to predict accurate marketing performance in global markets.

1.2.2. Market-Driven Management Strategies

In this highly competitive arena the way marketing is done is completely different from the past. Lambin (2007) identifies the main elements that differentiate the marketing approach from market-driven management. Firstly, it is clear that the main marketing focus is on customers while market-driven management focus the attention on the market competitiveness and on every market players with an outside perspective. Secondly, 'marketing is based just on a simple 'pull market model' (strategic response marketing) so the need to launch a new product come from the market and not from the company while market-driven management is based both on the requests of market and on innovative models linked to a technological impulse' (Gordini, 2010, p. 7). In other words, this means that in a logic market oriented data mining models represent a great solution for to capture competitive information from the market. Thirdly, marketing is focused on the paradigm of 4 Ps instead market-driven management overcome this point of view and its centre of attention is on intangible assets as primary source of the competitive advantage. As a consequence, it is obvious as the main traditional marketing strategies as product differentiation, cost leadership and market segmentation are not sufficient for obtaining a business advantages and to break down the competitors. In conclusion 'the concept of marketing is only limited to the marketing function, while market-driven management is based on a culture that pervades every level and every function of the company, striving to achieve complete functional interaction' (Gordini, 2010 p. 7). In global market market-driven strategies represent the best approach for corporations for gaining a leading position in the market and to break down the competitors. This approach come out in the late 1980s with the publication of acknowledge paper (Shapiro, 1998; Webster, 1988, 1992; Deshpandè and Webster 1989; Kohli and Jaworsky, 1990), poses the question of the relationship between markets and behaviour whose goal is to obtain competitive and suitable advantages.

Market-driven management is one of the most innovative approaches to the understanding of successful companies' behaviour in global market (Day 1994; Brondoni, 2007). It is a business strategy based on direct, continuously benchmarking with competitors, in a contest of customer value management (Brondoni, 2008). In other words, market-driven management can be defined as 'a new competitive market-oriented management philosophy', in which competitive customer value management predominates. The market-driven approach (Day, 1994, 1998; Jaworski et al. 2000; Narver et al. 2004) is based on the silent assumption that the market is somehow a given (Vallini and Simoni, 2009). Stanley and Narver (1994 p. 22) observe that 'a business is market oriented when its culture is systematically and entirely committed to the continuous creating of competitive customer value'. This means that to building competitive customer value, companies have to collect, coordinating and analyse customer and competitor data. It is important underline that the heart of a market orientation is its customer focus. In fact, the marketing activities have to provide more concrete elements to increase the competitive advantages and achieve a 'breakthrough' of sales. Customer loyalty, the identification of customer segment or better the detection of bubble demand with high profitability (Corniani, 2002 and 2005), the discovery of new market sectors that offer high profit margins and the continuous monitoring of competitor are the new challenges that marketing sector have to able to deal.

Global companies before adopting a market-driven strategy, should implement a customer relationship management programme in order to generate long run relations with the main profitable customers. In fact, 'to create customer superior value for buyers continuously requires that a seller understand a buyer's entire value chain, not only as it is today but also as it evolves over time' (Stanley and Narver, 1994 p. 22).

Kohli and Javorsky (1990) said that market orientation provides leading superior performance through a unifying focus on the efforts and projects of individuals.

Recently, a lot of empirical research has showed a strong relationship between market orientation and several measures of business performance, including profitability, customer retention, sales growth and new product success. It is clear as a market-oriented approach requires that global organizations know and understand their markets (market orientation) and customers (customer orientation). In fact, customer relationship management can be defined as a strategic tool that collects, integrates, manages and analyzes various customer data from different operation system in departments within an enterprise (Kwok et al., 2007). Also, (Panagiotis et. all, 2007) define customer relationship management as a proactive system that optimizes values (revenue and customer satisfaction) through the identification of the high profitability customer segments in order to develop only with them long run relationships.

According to (Parvatiyar and Sheth, 2001, p.5) 'customer relationship management can be defined as a 'comprehensive strategy and process of acquiring, retaining and partnering with selective customers to create superior value for the company and the customer. It involves the integration of marketing, sales, customer service, and the supply chain function of the organization to achieve greater efficiencies and effectiveness in delivering customer value'. Finally, a market-driven company is therefore one that not only reveals a superior ability to understand, attract and keep valuable customer (Day, 2000, 2001), but one that is also able to organise and exploit resources and skills in order to act before and better than competitors. As a consequence of this, for market-driven companies is fundamental developing customer relationship programs joined with a data mining analysis in order to monitor continuous the customer behaviour for getting a suitable competitive advantage in the market.

According to Brondoni (2009 p.2) this new strategy is really striking because it favours: 'activities focus on the profitability of competition, rather than on simple customer satisfaction, market policies based on innovation and competitive pricing, to stimulate uncertain and unstable customers to purchase and performance metrics even with very short timeframe'. In open market the winner firms are only market-driven organizations because of they are able to organize and discover resources and capabilities (Hult and Ketchen, 2001). In fact, the market-oriented approach foresees the demand in a 'circular relationship' with trade and manufacturers, with the aim to create new purchasing models based on-loyal behaviour, which joins the well-known loyal mechanism (Lambin, 2007). Also, these global organizations focus strongly on communication and are permeable to information; they assume that all business functions are aware of the behaviour of the competition, anticipate demand expectations, and finally, are determined to propose solutions that go beyond the roles of individual functions and the physical space of natural competition (Webster, 2002).

In this contest, global companies must focus their attention on the competitive space for identifying bubble demand and choosing the product features closest to demand expectations in order to plan substantial differential advantages of supply (Brondoni, 2009; Best, 2004, Brondoni and Lambin, 2001; Golinelli, 2000; Day, 1999 and 1990). It is obvious that in global scenario the need for corporations to identify new metrics or better new models for evaluating the impact of intangible factors that influence corporate performance is stronger than ever. As a result of this, an adequate knowledge management becomes a crucial resource for the survival of the global companies.

According to Bueren, Schierholz, Kolbe and Brenner (2004) knowledge is a strategic intangible asset at the base of competitive advantage in global organizations. Decision makers must analysis and manage with accuracy all knowledge available. In fact, in market-driven organizations emerge the need to develop data mining models with the purpose to analyze the hidden knowledge collected in huge database for obtaining competitive knowledge. Also, global corporations should invest a lot of resources in data mining in order to make sense out of data. Only a significant analysis of these data it would allow companies to gain a leading position in markets and break down the competitors.

In conclusion, in this hypercompetitive arena, to catch and analyzing secreted knowledge it represents the main source for achieving a suitable competitive and for increasing the competitiveness level of the companies. Indeed, one of the main purposes of this PhD thesis is to focus the attention on the main data mining methods fit to analyze this immense wealth.

1.2.3. Knowledge Management and Competitive Customer Value

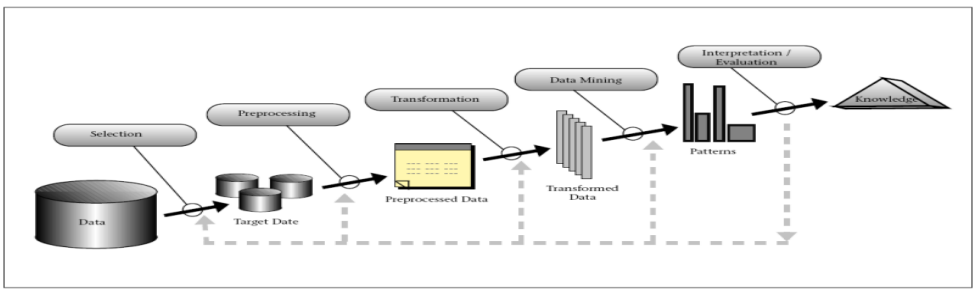
Knowledge management has emerged in recent times as a wide-ranging phenomenon's for organizational innovation and competitiveness. In the recent years, in particular from 2002, the knowledge management is became one of the top theme for business and management. This interest is linked to a numerous reasons. First, researches show that there is already evidence to indicate that global organizations are "flagging" due to information overload (KPMG, 2000 and BravoSolution, 2012). Secondly, the increasingly uncertainty of global business environment is stretching organization capabilities to the limit, requiring ever-higher level of information and intelligence to respond. Obviously, in this competitive arena 'the effective management of knowledge may not be a passing fad but potentially an essential tool to survival, and possibly the base for developing the ability to thrive through creating a new source for competitive advantage' (Lee and Carter, 2005 p.489). If global organizations understand the value of knowledge management, they have the opportunity to establish long term internal strengths, which will lead to external competitive advantages. Statistical data analysis tools must evolve beyond traditional reporting models in order to support the decision makers in business decisions. These tools have a real 360 degree view of the enterprise or business, but they analyze only historical data about what has already happened. Models based on time series analysis and seasonal forecasts help gain insight for what was right and what went wrong in decision-making providing a rear view analyses. However, one cannot change the past, but one can prepare better for the future and decision makers want to see the predictable future, control it, and take action today to

attain tomorrow's goals. Corporations for gaining a suitable competitive advantage on the market, must implement new data analysis approaches in order to create superior customer value and to increase the level of business performance. The knowledge generated by data mining models seem appear to be prominent than traditional statistical models. Data mining models allow changing raw data into business knowledge.

Current literature indicates that knowledge management can be implemented in every organizational discipline in particular in marketing area and in global companies. A number of authors have tried to define the concept of knowledge management. For example, Van Der Spek and Splijkervet (1997) define knowledge management as 'the explicit control and management of knowledge within an organization aimed at achieving the organization's objectives'. Also, Rowley (2000) considers knowledge management as 'concerned with the exploitation and development of the knowledge assets of an organization with view to furthering the organization's objectives'. Backman (1997) posits that the management of knowledge is essentially for the 'formalisation and access to experience, knowledge and expertise that create capabilities, enable superior performance, encourage innovation and enhance customer value'. Bassie (1997) considers the process of knowledge management as the system of creating, capturing, and using knowledge to increase business performance. Tan, Steinbach and Kumar (2011, p.3) define 'data mining as an integral part of knowledge discovery in databases, which is the overall process of converting raw data into useful information'. In other words, this process consists of a series of transformation steps, from data pre-processing to post-processing of data mining results.

The following figure provides an example of knowledge discovery in databases process adopted from Fayyad 1996.

Figure 1.1 The process of knowledge discovery in databases



More precisely, the input data can be collected in many formats (flat files, spreadsheets or relational tables) and many reside in a centralized databases or be distributed across multiple sites. After, during the pre-processing phase, the input data must be converted into an appropriate format for subsequent analysis. This step includes fusing data from multiple sources, cleaning data to remove noise and duplicate observations, and selecting records and features that are relevant to the data mining task at hand. Then, the a data mining analysis can be developed. For instance, in marketing application, the insights offered by data mining results can be integrated with campaign management tools thus that effective marketing promotions can be conducted and tested (for major details please see Chapter 5). Finally, a post-processing phase is conducted. Only valid and useful results are incorporated into the decision support system. An example of this

step is visualization, which achieve analyst to investigate the data and the data mining findings from a variety points of view. In this case, in order to eliminate spurious data mining results statistical measures or hypothesis testing methods can be implemented.

According to KPMG (2000), one of the big four consultancy companies, knowledge management can be described as 'the systematic and organised attempt to use knowledge within an organization to improve performance'. Carneiro (2000) observes that knowledge management is a fundamental strategic tool, because it can represent the mainly key for the formulation and evaluation of alternative strategies in the decision making process. Additionally, Meso et al. (2002) argue that is well-known as for companies the knowledge has a strategic role in obtaining sustainable competitive advantages.

Skyrme and Amindon (1997) note and identify the main success factors that global companies could be able to achieve through successful knowledge management process. These factors are: competitive advantages, customer focus, employee relations and development, innovation and cost. Finally, Lee and Carter (2005, p.494) define knowledge management as 'the process by which we gather, create, share, use information, experience, learning, information system, to add value, increase organizational wealth for competitive advantage and personal development'. In recent years, as said before, much research have established that in global business the most important type of knowledge would be appear to be customer knowledge. In fact, according to Zanjani, Rouzbehani and Dabbagh (2008) it is possible observe that traditional knowledge management is related to the efficiency gains while customer knowledge is connected to innovation and growth.

1.2.4. Competitive Customer Knowledge

The growing importance of the customer knowledge is demonstrated by numerous publications and is supported by many empirical studies (Bueren, Schierholz, Kolbe and Brenner, 2004; Rowley, 2002). In a survey carried out by Ernest Young, customer knowledge was quoted as the most important type of knowledge (97%) for global companies to act in an efficiently and effectively way. Daneshgar and Bosanquet (2010) paid the attention on the knowledge about of the best practises and effective process (87%), and on the knowledge about competencies and capabilities (86%).

According to Smith and Mckeen (2005) and Shammari (2009) global companies, in order to maximize the competitive customer value, need a large variety of knowledge about customers, such as: who are their customer? How can they use knowledge to retain and support customer? How can companies use customer knowledge to continuously improve product and service? How companies use customer knowledge to understand markets better? How can knowledge help companies acquire new customers? How can companies use customer knowledge to create new products and services? 'To reach these levels of knowledge, most global organizations have focused on collecting vast amount of data about customers' (Buchnowska, 2011 p.25).

Zanjani, Rouzbehani and Dabbaghl (2008) define customer knowledge as a kind of knowledge in the area of customer relationship management with direct or indirect effect on organizational performances. In addition, global companies have to understand data or information that can be analyzed, interpreted and eventually converted into

competitive knowledge. In literature there is a distinction between the following terms: customer data, customer information, and customer knowledge (Rowley, 2002; Rollins and Halimen, 2005).

Contact data, interaction data, purchasing data and customer feedback are examples of customer data (Wrycza, 2010). Data is collected in the company's databases, paper, and report and in the minds of the employees. Due to the introduction of marketing tools such as fidelity cards, dial into Voice Response Units and order off the Web, global corporation are saturated in customer data.

Most organisations have implemented information technology tools as Customer Relationship Management, Contact Center, Enterprise Resource System and E-Business System in order to collect customer data at every possible customer contact point (Rollins and Halinen, 2005). These touch point includes customer purchases, sales force contacts, service and support calls, web site visits, customers purchase behaviour, satisfaction surveys, credit and payments interactions and market research studies (Kotler and Armstrong, 2010).

Today, Customer Relationship Management is the main tool to capture data related to customer transactions (Ogunde, Folorunso, Adewale, Ogunleye and Ajavi, 2010).

Sometimes, such activities are insufficient because global companies cannot transform customer transactions into customer knowledge. In fact, to collect terabytes of customer transaction data does not guarantee business value.

It is fundamental to convert raw data into information and to combine this information throughout the corporation to develop knowledge useful for decision makers (Belbaly, Benbya and Meissonier, 2007) So, in order to make sense out of the data and to discover knowledge in databases, global companies must implement a new generation of computation techniques and tools for supporting the extraction of useful knowledge from the rapidly growing volumes of data. 'Customer data is obtained through filtering, integrating, extracting or formatting customer data' (Buchnowska, 2011, p.27). After collecting data, companies transform customer data into customer information, through various information systems. The main tools to extract and to analyze huge volume of data are the following: Customer Relationship Management, Business Intelligence and Customer Intelligence System. All this information systems are based on Data Mining Applications and Statistical Algorithms. An important remark is that the term customer knowledge is frequently incorrectly confused with the concept of customer relationship management. This latter it is been defined as the business strategies, process, culture and technology that achieve firms to maximize their revenue and to rise their value through understanding and satisfying the individual customer's needs (Reynolds, 2002).

According to Goldemberg (2003) it is possible notes that this process integrates people, process and technology in order to maximize relationships with all customers. The main difference between customer relationship management and customer knowledge is that the first is broadly focused on the management of customer knowledge, while the second is completely focused on knowledge from customers (Wilde, 2011; Gibbert, Leibold and Probst, 2002).

‘Knowledge from the customer is the knowledge that organizations receive from customers’ (Buchnowska, 2011 p. 27). This knowledge category includes the following types of knowledge: customer knowledge of products, supplier and markets (Gebert, Geib, Kolbe and Riempp, 2002), the customer ideas and suggestions about the improvement of the product or the service (Triki and Zouaoui, 2011), ideas thoughts and information related to the preferences, creativity or experience with products, services, processes or expectations (Peng, Lawrence and Lihua, 2011). Customer knowledge can be consider as an organized and analyzed competitive customer information so that it becomes understandable and applicable in solving problems and making decisions in the area of relations between an organization and its customers. Often, information technology tools facilitate the gathering of customer data and its transformation into customer information. Unfortunately, these tools cannot convert customer information into customer knowledge, because knowledge is always related to a person or group of people (Rollins and Halinen, 2005; Ziemba and Minich, 2005). Once again it emerges that the data mining approach and the knowledge discovery in databases are relevant for global companies for discovering tapped knowledge.

Finally, given the dynamism of the data mining models any process, from biotechnology to customer service, can be analyzed using data mining. Foley and Russuel (1998) and K Pal (2011) said that the top three ends of uses of data mining are in business, especially, in marketing area.

1.2.5. Statistical Data Analysis and Data Mining Analysis: what are the differences?

Sato (2000) observes that the data mining analysis differs from the statistical data analysis. In fact, statisticians use sample observations to study the population parameters by estimation, testing and predictions. According to the author is possible to consider the main features that differentiate these two different analyses. First, data mining analysis is governed by the need to uncover, in a timely manner, emerging trends, whereas statistical data analysis is related to historical fact and it is based on observed data. Second, statistical data analysis focused on finding and explaining the major source of variation in the data, instead data mining analysis endeavours to discover, not the obvious source of variation, but rather the meaningful, although currently overlooked information. Third, statistical data analysis manages data related to a specific research questions while data mining analysis explores data collected for different purposes other than the aim of the research.

Giudici (2003) observes that data mining is not just about analysing data; it is a much more complex process where data analysis is just one aspect. Turban, Aronson, Liang and Sharda (2007, p. 305) define data mining as ‘the process that uses statistical, mathematical, artificial intelligence and machine-learning techniques to extract and identify useful information and subsequently gain knowledge in databases’.

‘To apply a data mining methodology means following and integrated methodological process that involves translating the business needs into a problem which has to be analysed, retrieving the databases needed to carry out the analysis, and applying a statistical technique implemented in a computer algorithm with the final aim of achieving important results for taking a strategic decision’ (Giudici, 2003 p. 6).

The strategic decisions will itself create new measurement needs and consequently new business needs, setting off what has been called "the virtuous circle of knowledge" induced by data mining (Berry and Linoff, 2004).

Exploring databases for hidden competitive information through sophisticated data mining tools is the matter of concern for the business organizations. Effective use of information technology is the first step to achieve the data sufficiency that helps the companies in making time to time business decisions.

Rapid advances in information and sensor technologies along with the availability of database management technologies combined with the breakthroughs in computing technologies, computational methods and processing speeds, have opened the floodgates to data dictated models and pattern matching (Fayyad and Uthurusamy, 2002; Hand et al., 2001).

Today, it is clear as the use of sophisticated and computationally intensive analytical methods are expected to become even more common place with recent research breakthroughs in computational methods and their commercialization by leading vendors (Grossman et al., 2002).

'In the recent years computing environment search engines plays a crucial role in digging out the hidden facts that are processed through well define decision support system' (Murthy, 2010 p. 2). These decision systems utilize available information and through data mining models and human interaction to provide a decision making tools to analysis the data and information.

Decision support system and data mining models combined together can be appearing as 'the spectrum of analytical information technologies' providing a unifying platform in order to obtain an optimal combination of data dictated and human driven analytics. In other words, data mining process cover many activities: from the identification of business problem to the visualization of the results up to the interpretation and evaluation of the findings. In contrast, statistics is only the science of learning from the data. In fact, the main goal of statistics is just to interpret the data and not to understand the causes of the results obtained. In terms of data description and inference, about the parameters under study, it includes everything from the data collection to the data processing. 'The intersection between statistics and data mining enables the user to make effective utilization of the available data, to gain a better understating of the past, and predict the future through better decision making' (Murthy, 2010 p. 2). Analyzing this point of view emerge that statistical data analysis and data mining analysis are complementary.

The statistical data analysis emphasizes and removes the major part of data variation before that data mining analysis is used. This explains why the data warehousing tools not only stores data but also contains and executes some statistical analysis programs. Additionally, statistical theory and methods are central to the classification, clustering, and modelling issues involved in most data mining applications especially for driving the quality of the variables used and to test final results.

Statistical data mining applications, which nearly always involve making use of information draw from multiple databases, are particularly subject to limitations of data

and methods. Also, the procedures used to combine the individual data sources may themselves introduce error and uncertainly.

The different features of the datasets involved can give rise to multiple sources of error that may interact with one another in unknown ways. The underlying sources of error that may include, among others, coverage and content errors, the possibly different time references of individual datasets, and the additionally uncertainly introduced when some of the datasets are based on samples. These sources of error and uncertainly emphasize the importance of ensuring that the necessary statistical expertise is involved in data mining application.

Data mining tools and applications are proved to be an asset to the organization. In fact, according to Murthy (2010) global companies using the data mining techniques and tools to answer at the following questions:

- Which segment of population is most likely to respond to a particular advertising campaign?
- How many clusters or bubbles demand it is possible to find in huge databases?
- How many products are bought at the same time and in contemporary?
- How many customers could become churners in a given time?
- How many customers are in an insolvency state in a given period?
- What is the level of satisfaction related to a given product or service?
- What are the ideal conditions for launching a new product or service?

On the other hand, statistics can help greatly in this process by helping to answer several important questions about the data:

- What hidden patterns are there in the databases?
- What is the probability that an event will occur?
- What patterns are significant?
- What is the high level summary of the data that gives some idea of what is contained in the database?

Finally, it is clear as statistics and data mining are complementary. Data mining is useful with large quantities of data while inferential statistics when there are small quantities of data. Indeed, the domain knowledge is useful in either case.

1.2.6. From Traditional Statistical Analysis to Data Mining Analysis

Traditional statistical methods have often encountered many problems in meeting the challenges posed by new databases. According to Tan, Steinbach and Kumar (2011) the following bulleted list shows the main challenges that motivated the development of data mining.

- **Scalability.** A new generation of digital databases are becoming common in today's global business. Traditional statistical models and methods become inadequate to analyze this huge amount of data. In fact, if data mining algorithms are able to investigate these massive data sets, then they must be scalable. Additionally, scalability require the implementation of new data structures in order to access individual records in an efficient manner. Finally, scalability could also be improved by using sampling or developing parallel and distributed algorithms.

- **High Dimensionality.** From medicine to business fields it is common to encounter enormous databases, especially in web marketing contests. Statistical data analysis that is fine for low-dimensional is definitely inadequate for such huge databases. Besides, the computational complexity increase rapidly as the dimensionality (in terms of features) increases.

- **Heterogeneous and Complex Data.** Traditional statistical data analysis is helpful for decision makers with databases containing attributes of the same type and with categorical or continuous data. Instead, Data Mining Analysis is able to investigate and explore heterogeneous attributes. This approach is worthwhile in Web Marketing because of web pages containing semi-structured text and hyperlinks for instance. 'Techniques developed for mining such complex data should take into consideration relationships in the data, such as temporal and spatial autocorrelation, graph connectivity, and parent-child relationships between the element in semi-structured text and XML documents' (Tan, Steinbach and Kumar 2011).

- **Data Ownership and Distribution.** In global markets the data needed for an analysis is not stored in one location or owned by one organization. Today, data are geographically distributed among resources belonging to multiple entities. As consequence, Traditional Data Analysis are not enough to investigate with high accuracy these amount of data. In fact, the development of data mining techniques is a key factor for global companies that must survival in a community based on the network logics.

- **Non-traditional Analysis.** The traditional statistical approach based on a hypothesize-and-test paradigm is too much complex and extremely laborious.

Current data analysis tasks require the generation and evaluation of thousand of hypothesis, and consequently, the development of some data mining techniques has been motivated by the desire to automate the process of hypothesis generation and evaluation. Notice that this new generation of databases frequently involve non-traditional types of data and data distributions.

From this discussion emerges as data mining methodology is the main approach that decision makers could adopt within global companies in order to get optimal marketing strategies. Predicting marketing performance with Traditional Statistical Methods such as Double Moving Average or Exponential Smoothing could lead global companies in huge traps because of this methods are based just on one independent variable. Finally, these models are inadequate and inapplicable in digital contest in particular in web marketing field.

1.2.7. Data Mining Process and Competitive Knowledge Discover in Database

In global markets a new generation of statistical models are required to analyse massive amount of data collected in huge databases. The main problem for global companies is to understand which raw data could be contains competitive information thus relevant knowledge for getting strategic decisions. But, often this potential competitive knowledge remains stored in dormant databases without to create new competitive customer value for enterprises. In fact 'databases are frequently a dormant potential resource that, tapped, can yield substantial benefits' (Fayyad, Piatetsky-Shapiro and Smyth, 1996 p. 28). In close and static markets and in scarcity of supply market condition, due to the limited volume of data, for turning raw data into knowledge, decision makers use traditional methods based on manual analysis and interpretations. This approach is too much slowing, expensive, and highly subjective and it makes no sense in today's global business. 'Databases continue to grow both in terms of the number *N* of records, or objects, and the number *d* of fields, or attributes, per object' (Fayyad, Piatetsky-Shapiro and Smyth, 1996 p. 28). As consequence, the knowledge discovery in database process and the data mining methodology are always more important both in managerial and academic contest in order to extract competitive knowledge from huge databases.

It is really important to discern data mining from online analytical processing. This latter are quite different from data mining because it provides only a really good view of what is happening but cannot predict what will happen in the future or why it is happening (K Pal, 2011). Besides, data mining is considered an emergency methodology that has made a revolutionary change in the information society (K Pal, 2011).

In literature, the knowledge discovery in database and the data mining could have the same meaning. These terms are interchangeably (Fayyad, 1996; K Pal, 2011). Data Mining could be seen as an integration of more disciplines as statistics, computer science and artificial intelligence, machine learning and data base management (Murthy, 2010).

According to Gartner Group, world leading information technology research and advisory company, and Larose (2010), data mining is the process of discovering meaningful new relations, models and trends through the analysis of a large amount of data collected in huge databases, using statistical and mathematical techniques, models and methods. In other words, the main role of data mining is to discover, in advance, unknown relations in enormous databases in order to predict actions, behaviours and outcomes related to the market players.

In global scenario data mining is one of the main methodologies able to forecasting with high accuracy business performance. Additionally, 'data mining is the process that uses statistical, mathematical, artificial intelligence and machine-learning techniques to extract and identify useful information and subsequently gain knowledge from large database' (Turban, Aronson, Liang and Sharda, 2007 p.305).

Data mining models are helpful in the decision making process because they surmount the limits of traditional statistical models. In fact, before getting a strategic business decision, data miners, also, evaluate the quantitative results in a qualitative manner.

Broadly, 'data mining could be useful to answer the queries on: forecasting with regard to what may happen in the future, classifying things into groups by recognizing patterns, associating similar events that are likely to occur together, clustering the peoples into group based on their attributes and making the sequence what events are likely to lead to whom' (K Pal, 2011 p.10). Finally, data mining is based on the multiplicity of applications especially in business and marketing area (Figini and Giudici, 2009). Common uses of data mining are for instance:

- **Market-Basket.** It identifies which products are jointly purchased with others in order to improve the layout of goods on the shelves and to increase, through promotions on items associated with them, sales of determinate product or group of products.
- **Web-Click-Stream**. The main goal of this application is to understand the dynamics of web navigation for improving the navigation on the sites and, in e-commerce sites, to accelerate the paths that promote the purchase.
- **Profiling**. It allows creating homogeneous profiles of customers, based on their past behaviour, in order for involving them in targeted promotions.
- **Retention/Churn.** It identifies homogeneous groups of customers, in terms of behaviour and personal characteristics, in order to retain existing customers and to identify new potential customers.
- **Scoring.** A set of statistical methods used to associate a score to each customer; this score have to reflect the behave interest propensity. For instance, in the business field, in order to evaluate the potential of a potential customer.
- **Risk Management**. A set of statistical methods used to calculate how much capital to allocate to risk hedge. The committee of central banks has established the need to measure: market, credit and operational risk.
- **Mining Unstructured Data, such as Text.** The text data is always unstructured. So data mining tools can help to mine the unstructured data to help the various organizations to get good out of the data.

Finally, a significant data mining approach demands within the global companies the interaction between three business people at least:

- Business Project Managers: it is the developer of the data mining project and it manages and oversees the progress of the project.
- Information Technology Managers: it provides the necessary technologies for extracting the data from huge databases and to manage the creation and migration phase.
- Data Mining Analysts and Managers: they represent the main key for the success of the project because they analyse and interpret the data.

1.3. Research Methodology

The purpose of this dissertation is to identify the main data mining models for predicting the churn probability in global scenario in order to make sense of the data collected in enormous databases and to increase the probability of customer conversion. Also, it shows that there is an emerging need to create a 'bridge' between academic contest and managerial reality so as to improve the marketing performance accuracy. For this reason in this PhD thesis a new business research is proposed. It provides a mixed methods research combining quantitative and qualitative research with the aim to breaking down the quantitative/qualitative divide. The main reason of this choice is that in business contests both approaches must be considered in estimating accurate marketing performance.

1.3.1. Research Contest

In global business relevant and significant marketing forecasts are really important tools in order to improve the level of competitive customer value and to increase company's competitiveness. After the rapidly development of globalization and the dramatic advanced in information technology traditional forecasting methods as Single and Double Moving Average Model and Exponential Smoothing Model become inadequate to predict future marketing trends because of many global corporations are owners of a large amount of data collected in huge databases.

According to Hua Bo-quan (1995) Double Moving Average is an average computing method based on Single Moving Average model. 'Firstly, it uses the single moving average twice to get the one moving average value and one moving average based moving average value, here we call it twice moving average value. Then the algorithm uses the two kinds of value to calculate the target data, according to its compute model. This model has a lag bias problem in particular when the time series have a linear trend because the moving average is always 'behind' the changes of the observed value' (Geng and Du, 2010). For this reasons, in order to predict marketing performance is common to use the Double Moving Average model. This model solves both the problem of lag bias and it is applicable to time series with a clear changing trend based market prediction. Additionally, Double Moving Average model smooth out the impact of sudden fluctuations on predicted result.

Thomas Robert B.M. (2008) defines the Exponential Smoothing Model as a model that incorporates seasonality and trend (features of the time series) and realise on a weighted average of past values of a time series to estimate future value. This model is useful in particular in short term predictions because it was built on the theory that the trend has the features of stability and regularity. Finally, Exponential Smoothing Model can be defined as a very effective marketing budget, statistical methods (Yu-shui Geng and Xin-wu Du, 2010).

However, these models, especially in global markets have some problems. First, with the double moving average model the predictive value of the data will be less sensitive to the actual changes and moving average is not always a good trends indicator. Second, the predicted value always remains at the level of the past and cannot be expected to predict a higher or lower volatility of future. Instead, the exponential smoothing model requires a more complete historical data before starting the prediction and if season

factors influences business sales a lot, times decomposition is more applicable than exponential smoothing. Despite this, the exponential smoothing model is one of the main approach used in many global organization in order to predict business and marketing performance (Hyndman, Koehler, Ord, and Snyder, 2008). But, unfortunately, the adoption of this model could bring down many global firms into huge traps. This fact underline the importance of the first[1] and second[2] research question proposed in this PhD dissertation.

In addition, it is obvious that there is an emerging need to create a "bridge" between academic and managerial contest in particular in marketing sector. So, it is clear as in global scenario these models are inaccurate and too complex. For this reason, in today global business, the predictive data mining models are the key to solve these problems (Geng and Du, 2010).

1.3.2 Research Method

The dataset analyzed was provided by a global company that offers digital data driven marketing solutions across all interactive channels: digital, direct response, relationship based media and design. Currently this company operates in most different countries across Europe, North America, South America, Asia, Africa and Australia. Unfortunately for privacy reasons it is impossible give other details about the company.

A data mining exploratory analysis is used to accomplish the research goals. The original whole database contained more than 1,463,199 potential customer and 42 variables related to them. In order to ensure maximum accuracy of the results the available population was not sampled.

The dataset analyzed is composed both quantitative and qualitative variables. The qualitative variables (for instance: the name of the advertiser, the type of banner, the timestamp for the activity on the advertiser's website, the keywords categories, etc.) have been treated as a quantitative because they were categorized into groups.

Given the huge size of the database, before aggregating the data by user id (code that identifies each potential customer) it has been necessary to perform a preliminary screening of the variables for detecting possible outliers and insignificance variables.

After the aggregation by 'user id' the database contains 1,463,199 potential customers and 276 variables. The number of variables increase dramatically due to the creation of dummy variables. Before starting an exploratory analysis it was fundamental a second preliminary screening with the aim to eliminate new irrelevant and redundant variables through correlation index and multiple linear regressions. After this second screening, the number of variables decrease at 19.

The quantitative nature of the research aims at conducting rigorous theory testing using predictive computational data mining models.

[1] Why in today global business marketing forecasts are often inaccurate?
[2] Predictive data mining models should be better than traditional statistical models to predict marketing performance in global markets?

The target variable is dichotomous thus it assumes only two values: 0 when the potential customer does not buy the service (churner) and 1 when the potential customer purchases the service. In fact, given the form of the data and the dynamism of the database analyzed predictive data mining models as Hierarchical Logistic Regression and Classification Decision Tree are appropriate tools in order to forecast the best marketing activities for increasing the probability of customer conversion so minimize the number of churners. In addition, to evaluate the results of these models, decision makers can used the Criteria based on the Loss Functions as Confusion Matrix and Roc Curve.

The data mining analysis starts with an exploratory investigation based on the correlation index in order to identify the relationship between the target variable and one or more variables collected in the database. Then, through multiple linear regressions we have identified the multicolinearity between variables. This stage is really important because it achieves to create an accuracy predictive model.

In conclusion, Hierarchical Logistic Regressions and Classification Decision Trees have been implemented so as identify the best marketing drivers that lead potential customers in a customer state.

1.4. Structure of the Thesis

This PhD dissertation is organized as follows:

Chapter 2 explains through a literature review the strategic importance of the data mining analysis within global companies with an outside-in perspective. Special focus is paid on the predictive data mining churn models. In fact, hundred scientific studies related to the most common data mining techniques helpful to minimize the probability of churn are provided.

Chapter 3 describes the main features of the database analyzed and the research methodology developed to accomplish our research goals. This PhD dissertation is quantitative in nature but a qualitative approach (opinion mining) is used in each steps of the analysis. Both univariate and multivariate techniques are used in this research. In order to accomplish the research questions have been used Linear and Logit Data Mining Models such as Hierarchical Logistic Regressions and Classification Decision Trees. The results obtained have been validated through Criteria Based on Loss Functions such as the Percentage Correctly Classified at the 'economically optimal' cut-off purchase probability (PCC) and the Receiver Operating Characteristic (ROC) curve.

Chapter 4 shows the exploratory analysis and the aggregation criteria established to draw the final database.

Chapter 5 presents the results of the data mining analysis and proposes a what-if scenario simulation to testing the effectiveness of the model developed. In this section, particular attention must be paid on qualitative approach in order to read the findings in an accurate manner.

Chapter 6 provides the conclusions and proposes suggestion for future research.

Chapter 2
Churn Risk and Data Mining: A Theoretical Framework

2.1 Introduction

From the last decade customer churn has become a critical business issue for many companies regardless their size. Predictive modeling based on knowledge discovery of data mining, are an innovative approach in predicting the risk of churn in today's competitive landscape. In fact, the development of organizational capital through databases and data mining models is a fundamental parts for building a market-driven organizations. A mixture of new techniques and methods is emerging to help sort through the data and find useful competitive data able to generate competitive knowledge. The main aim of this chapter is to provide a theoretical framework related to the main data mining techniques helpful in the estimation of the risk of churn. It starts explaining the origins of the data mining methodology emphasizing its strategic importance for global organizations with an outside-in perspective. Then, it shows 'a five stage model' which explains the customer churn management framework on the base of the current literature. After, it investigates on the churn management predictive models providing an analysis of hundred scientific studies. Finally, it identifies further areas for customer churn management research.

2.2 The Origins of Data Mining

The First International Conference on Knowledge Discovery and Data Mining held in Montreal in 1995 and still considered one of the main conference on this topic. It was used to refer a set of integrated analytical techniques divided into several phases with the aim of extrapolating previously unknown knowledge from massive databases of observed data that no appears to have any obvious regularity or important relationship (Giudici, 2003). Data mining is an interdisciplinary field that combine artificial intelligence, database management, data visualization, machine learning, mathematic algorithms, and statistics (Tsai, 2012). Also, this methodology provides different techniques easy to implement during the decision making process. In other words, it is the process of discovery interesting knowledge such as patterns, associations, changes, anomalies and significant structures from huge amount of data stored in databases, data warehouses, or other information repositories. In an effort to develop new insights into practice performance relationships, data mining was used to investigate improvement programs, strategic priorities, environmental factors, manufacturing performance dimensions and their interactions (Hajirezaie, Husseini, Barfourosh, et. al., 2010). Data mining looks for relations and associations between phenomena that are not known beforehand. Besides, data mining allows the effectiveness of a decision to be judged on the data, which provides a rational evaluation to be made, and on the objective data available. Data mining has also been quick to adopt ideas from other areas, including optimization, evolutionary computing, information theory, signal processing, visualization, and information retrieval (Tan, Steinbach and Kumar, 2011). More precisely, the domain of data mining have been prospered and posed into new area of human life with various integrations and advancements in the fields of Statistics,

Databases, Machine Learning, Pattern Recognition, Artificial Intelligence and Computation Capabilities (Venkatadri and Lokanatha, 2011). The core functionalities of data mining are applying various methods and algorithms in order to extract patterns of stored data (Fayadd, Piatesky-Shapiro and Smyth, 1996). In fact, data mining has a rich focus for managerial implication due to its significance in decision making and it has become an essential components in various organizations. Focusing the attention on global companies, data mining is the right methodology to make better business decisions to remain highly competitive in the marketplace. According to Berson, Smith, and Thearling (2000), Lejeune (2001), Ahmed (2004), Giudici (2010), and Berry and Linoff (2011) data mining is the process of extracting or detecting hidden relationships or information from huge datasets. With an enormous amount of data, this methodology can provide business intelligence to generate new business opportunities (Bortiz and Kennedy, 1995; Fletcher and Goss, 1993; Langley and Simon, 1995; Lau, Wong, Hui and Pun, 2003; Salchenberger, Cinar and Lash, 1992; Su, Hsu, and Tsai, 2002; Tan and Kiang, 1992; Zhang, Hu, Patuwo and Indro, 1999). A number of data mining applications and prototypes have been developed for a variety of domains (Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro, and Simoudis, 1996) belong to at the following areas: marketing, banking, finance, manufacturing and health care. Today, a special attention is paid on time-series, spatial, telecommunications, web and multimedia data. According to Venkatadri and Lokanatha (2011) is obvious that data collected from different applications require proper mechanism of extracting information/knowledge from large repositories for better decision making. Finally, an important remark is that both data mining process and techniques to be apply depend very much on the domain application and the nature of the data available.

2.3 Customer Churn Management: A Theoretical Framework

In over supply market economic condition, global companies have realized that their business strategies should focus on identifying those customers who are likely to churn. 'Prediction of behavior, customer value, customer satisfaction and customer loyalty are examples of some of information that can be extracted from the data that should already be stored within a company's database' (Hadden, Tiwari, Roy and Ruta, 2005).

Focusing the attention on the churn definition is clear that it identifies those customers who are intending to move their customers to a competitive service provider. In other words, customer churn could be defined as the propensity of customers to cease doing business with a company in a given time period. It also became a serious problem for many firms. Particular attention must be paid on global companies with an outside-in perspective. The domain of the churn defection include several sectors such as publishing, investment service, insurance, electric utilities, health care, credit cards, banking, internet and telecommunication. Annual reports show that in the internet service industry the annual churn rates range is from 21% to 63,2% (Network World, 2001; Kolko and Jennifer, 2002). Indeed, in the wireless telephony industry, annual churns rates have been reported to range from 23.4% (Wirless Review, 2000) to 46% (Telephony Online, 2002). From a managerial standpoint, there are two basic approach to manage customer churn: untargeted and targeted. Untargeted approaches rely on superior product and mass advertising to increase brand loyalty and retain customers. A good example of this is AOL's recent efforts to decrease churn through better software and content (Yang, 2003). Targeted approach rely on identifying customer who are

likely to churn, and then either providing them with a direct incentive or customizing a service plan to stay. Once acknowledged, these customers can be targeted with proactive marketing campaign for retention efforts. Customer retention is become a essential aspect in the marketing field in today's global business. Remark that if the number of customer belonging to a business reaches its pick, finding and securing new customers becomes increasingly difficult and costly. At this point of the business lifecycle it should be higher priority to retain the most valuable, existing customers, than trying to win new ones (Hadden, Tiwari, Roy and Ruta, 2005). An example of this could be provide customers with market competitive service plans by segmenting their telecommunications calling behaviors. In specific, we can identify two types of targeted approaches: reactive and proactive. In the first case, companies tries to identify in advance customers who are likely to churn at some later data. Then, they target these customers with special programs or incentives to forestall the customer from churning. On the other hand, with the targeted proactive program, companies have potential advantages of lower incentive costs and not training customers to negotiate for better deals under the threat of churning. Decision makers must shed light the attention on these approaches because of these systems can be really wasteful when churn forecasting is inaccurate. As consequence, companies are wasting incentive money of customers who would have stayed anyway. Thus, the main goal for companies is to forecast customer churn as accurately as possible.

According to Yang and Chiu (2006) churn can be classified as following:

- Involuntary Churn: it occurs when customers fail to pay for service and as a result companies terminate service or when termination of the service is due to theft or fraudulent usage. For instance, there are several reasons why a company could revoke a customer's service, including abuse of service and non-payment of service.

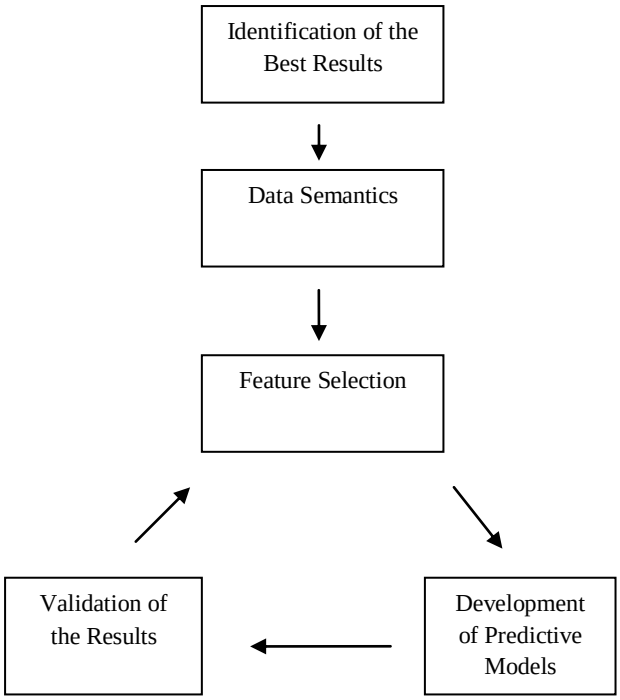
- Unavoidable Churn: it emerges when customer die, move or are otherwise permanently removed from the market place.

- Voluntary Churn: when customers make a conscious decision to terminate his/her service with the company. This kind of churn can be sub-divided into two main categories: incidental and deliberate churn. Incidental churn occurs when changes in circumstances prevent the customer from further requiring the provided service. It includes changes in the customer's financial circumstances, so that the customer can no longer afford the service, or when customer moves into new different geographical location in which the company's service is unavailable. On the other hand, particular attention must be paid on the deliberate churn because of its percentage is quite high. It occurs when customers move to a competitive company for technological or economical motivation. Customers can discover that a competitor is offering the latest products present on the market with an optimal price unlike his company. Examples of other reasons for deliberate churn include quality factors such as poor coverage, or possibly negative experiences with call center, etc. (Kim and Yoon, 2004).

Churn management efforts should not focus across the entire customer-base because not all customers are worth retaining, customer retention costs money and attempting to retain customers that have no intention of churning will waste resource (Hadden,

Tiwari, Roy and Ruta, 2005). According to Liu and Shih (2004) decision makers must develop news marketing strategies able to capture customer needs, and improving satisfaction and retention in order to capture the customer voice. Canning (1982) states that selling more to everyone is no longer a profitable sales strategy and a market place that continually growth more competitive requires an approach that focuses on the most efficient use of sales resources. Firms should take great care if they decide to purchase a CRM product 'off-the-shelf'. More precisely, Chen and Popovich (2003) state 'CRM vendors might entice organizations with promises of all powerful applications , to data there is no 100% solution'. As consequence, this research want to propose a 100% solution due to the uncertainty involved in churn prediction. The following figure explains the main stage related to the churn management problem, according to Hadden, Tiwari, Roy, and Ruta (2005) and Datta, Masand, and Mani (2001).

Figure 2.1 A five stage model for developing a customer churn framework



Adapting from Datta, Masand and Nami (2001)

The previously figure shows the main stage related to the churn management process. This model is very useful in global companies with an outside-in perspective. The followings subparagraphs explain in details these steps.

2.3.1 Identification of the Best Results

The first step concerns the identification of the best data in developing a customer churn management framework. It is fundamental to identify the data that best suits the type of the analysis that is being performed. Notice that different combination of data hold different analytical powers. In addition, different sets of data provide better indicators

for different problems and the service sectors. For instance, usage data has also been used for understanding e-customer behavior of website users (Jenamani, Mohapatra and Ghose, 2003) and predicting mail order repeat buying (Van den Poel, 2003). Ng and Liu (2001) suggest that usage data should be mined for identifying customer churn in the Internet Service Provider and telecommunication industry. Verhoef and Donkers (2001) state that the purchasing of products and services is best predicted using the historical purchasing data. This note could be false in today's global business especially in over supply market situation because of many exogenous factors affect the customer purchase behavior while, it is absolute true in close and static markets. These example confirms the strategic importance of this first step. In fact, to select the best data is the first point to develop adequate predictive models able to predict accurate marketing performance. In other words, the quality of the data determine the power and the accuracy of the overall model.

2.3.2 Data Semantics and Feature Selection

According to Ram (1995) data semantics can be describes as 'object, relationships among objects, and properties of objects'. In other words, it is the process of understandings the context of the data in a database. Sometimes the data collected in the datasets can be viewed as a collection of words difficult to interpret. Volz, Handschuh, Staab and Stojanovic (2003) suggests that there are more than 500 free bio-informatics databases available over the internet in which the data are difficult to understand because they have an inconsistent data definition. Today, this problem is diffuse not only in scientific fields but also in business contests. In fact, due to the development of the globalization and the rapid advances in technologies many databases contains web data. In literature they are various models able to capture the meaning and the structure of the data collected in the datasets (Ram and Khatry, 2003). Different types of tools could be used for communication between the designer of a database and the end-users such as the entity relationship model[3], the relational model[4], and the unifying semantic model[5]. Establishing data semantics is the most difficult phase of dealing within global organizations which manage huge amounts of data. But, despite some authors such as Datta, Masand, Mani, and Li (2001) that consider data semantics one of the most difficult phases in the churn management process especially with enormous datasets. The analysis of the literature review provided in this dissertation has discovered that the phase has not been documented in most research. There could be several reasons for this, including data sensitivity resulting in researchers being unable to publish any examples based on companies data, or again due to sensitivity, researchers are forced to use what they have been given, and are not given the chance to explore a company's data warehouse for themselves (Hadden, Tiwari, Roy and Ruta, 2005). However, data semantics is the most difficult phase in the churn management process, especially with vast databases that collect different types of data.

Feature selection is a critical process (Sun, Bebis, Miller, 2004). This process is helpful for identifying the fields which are the best for prediction. According to Yan,

[3] The entire relationship model represents data using entities, relationships, and attributes (Lee, 1999).
[4] The relational model supports a data sub-language for data definition, but can be seen as a complete database model, supporting all aspects of data management (Apenyo, 1999).
[5] The unifying semantic model is a formal specification for providing an accurate means of documentation and communication between users (Ram, 1995).

Wolniewwicz and Dodier (2004) it helps both data cleaning and data reduction, by including the important features and excluding the redundant, noisy and less informative ones. In literature are identified two steps related to features selection: the search strategy and the evaluation methods. The first concerns the identification of the future subsets while, the second refers to the patterns for testing their integrity, based on some criteria such as learning algorithm. Remark that once features that have been extracted from a dataset they need to validated. In addition, the data features extraction was based on should be completely independent from the validation data; otherwise could be a risk of over fitting.

2.3.3 Data Mining Predictive Models: Main Scientific Studies

A predictive model is defined as one that takes patterns that we have been discovered in the database, and predicts the future (Rygielski, Wang and Yen, 2002). In the current literature the most important predictive modeling techniques include decision trees and neural networks (Crespo and Weber, 2004). In addition, Baesens, Varstraeten, Van Den Poel, Egmont-Peterson, Van Kenhove, and Vanthienen (2004) defying neural networks and decision trees typical classification technologies. On the other hand, past research shows that decision trees, neural networks, logistic regression are well suited to study the customer churn management problem (Hadden, Tiwari, Roy and Ruta, 2005).

Kitayama, Matsubara and Izui (2002) used a decision tree approach to propose a model for customer profile analysis. The decision tree was applied to the segments in order to determine the necessary measures to take for both the preferred and regular divisions aim to prevent customer from switching to new companies. Decision trees performs accurately the research questions. Instead, Han, Lui, Leung (2011) propose a novel customer segmentation method based on customer lifecycle, which includes five decision models. Due to the difficulty of quantitative computation of long-term value, a decision tree method is used to extract important parameters related to long-term value. The results reached was adequate. Ng and Liu (2001) suggest that the for the purpose of customer retention the decision tree approach based on C4.5 algorithm is one of the main accurate techniques able to implement by decision makers. Wang, Chiang, Hsu, Lin and Lin (2009) use the decision tree algorithm to analyze data of over 60,000 transactions and of more than 4000 members, over a period of three months, in order to propose a recommender system for wireless network companies. The results of the experiment was useful for making strategy recommendations to avoid customer churn.

Risselada, Verhoef and Bijimolt (2010) test the predictive power of prediction methods such as regression analysis and decision trees in the internet service provider industry and insurance markets. In this case decisions trees outperforms the regression analysis. Buckinx, Verstraeten, and Van den Poel (2007) compare multiple linear regression with two state-of-the-art machine learning techniques in order to measure the loyalty of each single customer. As a result emerges that multiple linear regression model significantly outperforms the other models.

Zbkowski and Szczesny (2012) test several models such decision tree and neural networks to predict customer insolvency at one of the cellular telecommunication operator in Poland. Neural Networks perform well when modelling the customer's insolvency and the best model can capture significant amount of money owed by

insolvent customers. Indeed, decision trees get old quickly and their performance decrease over the time in the top percentiles of the score.

Guangli, Rowe, Zhang, Tian and Shi (2011) show that logistic regression analysis performs a little better than decision trees in building a churn prediction model using credit card data collected from a real Chinese Bank.

Hwang and Euiio Suh (2004) conducted experiment implementing decision trees, neural networks and logistic regressions. In this case decision trees showed slightly better accuracy over the other technologies but the authors state that this results do not prove decision tree to the best choice in all case. Mozer, Wolniewicz, Grimes, Johnson an Kaushansky (2000) confirms the opinion argue by the previously authors.

Mihelis, Grigoroudis, Siskos, Politis, Malandrakis (2001) developed a method to determine customer satisfaction using an ordinal regression based approach obtaining accurate results.

Rust and Zahorik (1993) used logistic regression to link satisfaction with attributes of customer retention.

Samimi and Aghaie (2011) explored the effect of heterogeneity across different classes of customers as well as their time dependent usage behaviour on the purchase rate of multiple services supplied by a subscription-based service provider. They show that a suitable model based on the logistic regression can effectively be employed to represent both the cross correlation and serially correlation of purchase rate for different kinds of services.

Kim and Yoon (2004) used a logit model to determine subscriber churn in the telecommunication industry, based on discrete choice theory.

Au, Chan and Yao (2003) noted that regression analysis is fine for determining a probability for a prediction, however, it is unable to explicitly express the hidden patterns in a symbolic and easily understandable form.

Keramati and Ardabili (2011) used a binomial logistic regression for identifying factors that affect customer churn. The results of this research indicated that a customer's dissatisfaction have the most influence on their decision to remain or churn. Also it implied that customer status (active-no active) mediates the relationship between churn and the cause of it.

Migueis, Van den Poel, Camanho and Cunha (2012) implemented a logistic regression analysis as the classification techniques in order to include in partial churn detection models the succession of first product categories purchased as a proxy of the state of trust and demand maturity of a customer towards a company in grocery retailing. The study confirmed that the logistic regression analysis outperformed the business expectations.

Datta, Masand, Mani and Li (2001) used simple regression to initially predict churn. Then, nearest neighbour, decision trees and neural networks have been developed. Their research could not established a best method, and they have stated future direction as including an explanation of customer behaviour, because their model could predict

customer churn, rather than it was unable to provide an explanation as to why a customer might churn. They suggest to include in the database exogenous information such as the state of the telecommunication market, and competing offers, etc. This model does not distinguished among loyal customers, valuable customer and less profitable customers.

Au, Guangquin and Rensheng (2011) proposed an iterative procedure to model multiple responses prediction into correlated multivariate predicting scheme combing partial least squares method and logistic regressions. Numerical results shown that the proposed scheme can improve the conventional regression models significantly.

Coussement, Benoit and Van den Poel (2010) focused their attention on how better supporting decision makers in identifying risky customers comparing logistic regression with generalized additive models. The study revealed that generalized additive models increase the business value in churn prediction context. Besides, Hwang and Euhiio Suh (2004) discovered that logistic regression better performed for forecasting customer churn when compared with neural networks and decision trees. They also suggests that logistic regression is the best model for their purpose.

Tsai and Lu (2009) consider two hybrid models by combining two different neural networks techniques for churn prediction, which are back-propagation artificial neural networks and self-organizing maps. In particular, the first technique of the two hybrid models performs the data reduction task by filtering out unrepresentative training data. Then, the outputs as representative data are used to create the prediction model based on the second technique. To evaluate the performance of these models, three different kinds of testing sets are considered. They are developed the general testing set and two fuzzy testing sets based on the filtered out data by the first technique of the two hybrid models. The experimental results show that the two hybrid models outperform the single neural networks baseline model in terms of prediction accuracy.

Rygielski, Wang and Yen (2002) discussed neural networks as data mining techniques for customer relationship management. Neural networks provide a more powerful and predictive model than other techniques in this research. They are also documented to be applicable to a wider area of applications. However, other disadvantages should be taken into account, like clarity of output, implementation and construction of model.

Tsai and Chen (2010) argued about the important processes of developing customer churn prediction models by data mining techniques. They contain the pre-processing stage for selecting important variables by association rules, which have not been applied before. Then they implement a neural networks and decision trees models, which are widely adapted in the literature, and four evaluation measures including prediction accuracy, precision, recall, and *F*-measure, all of which have not been considered to examine the model performance. Association rules allowed to the decision trees and neural networks models to provide better prediction performances over a chosen validation dataset. In particular, the decision trees models performs better than the neural networks models. Notice that some useful and important rules in the decision tree model, which show the factors affecting a high proportion of customer churn, are also discussed for the marketing and managerial purpose.

Boone and Roehm (2002) have applied neural networks to segmentation of customer databases in the retail service sector obtaining relevant findings.

Vellido, Lisboa and Meehan (1999) used neural networks to segment the online shopping market. In particular, they implemented a self organism map, which is an unsupervised neural networks. This approach is also analyzed by Shin and Sohn (2004) for segmenting stock trading customers according to potential value.

Kisioglu and Topcu (2011) paid the attention on the bayesian belief network to identify the behaviours of customers with a propensity to churn. The data used are collected from one of the telecommunication providers in Turkey. The results of the model performed the expectations.

To better clarify, table 2.1 provides an analysis of the main hundred scientific studies related to the main data mining techniques to manage the customer churn risk from 1978 to 2012 sorted by both the year of the publication and the kind of the data mining technique studied. Some general information such as: name of the author/s, name and sector of the scientific journal in which the study is published, the name of the country where the author/s work, and the year of the publications have been following provided.

Table 2.1 Scientific Studies Analyzed

Author/s	Author/s Location	Scientific Journal	Sector of the Journal	Publication Year	Data Mining Technique
Morgan	United States	Journal of Business Research	Management	1978	Cluster Analysis
Ahn, Han, and Lee	Korea	Journal of Systems and Software	Statistical	2004	Cluster Analysis
Duen-Ren Liua and Ya-Yueh Shih	Taiwan	Journal of Information and Management	Operational Management	2005	Cluster Analysis
Nie, Chen, Zhang and Guo	China-United States	Procedia Computer Science	Operational Management	2010	Cluster Analysis
Karahoca	Turkey	Expert System with Applications	Statistical	2011	Cluster Analysis
Wang, Hao, Ma, and Huang	China	Expert System with Applications	Statistical	2010	Cluster Analysis Neural Networks
Prinzie and Van den Poel	Belgium	Decision Support System	Operational Management	2007	Cluster Analysis SAM
Ng and Liu	China	Journal of Artificial Intelligence Review	Statistical	2001	Decision Tree
Wei and Chiu	Taiwan	Expert System with Applications	Statistical	2002	Decision Tree
Ho Ha, Min Bae and Chan Park	Korea	Journal of Computers and Industrial Engineering	Engineering	2002	Decision Tree
Shin and Sohn	Korea	Journal of Expert Systems with Applications	Operational Management	2004	Decision Tree
Wei and Chiu	Taiwan	Expert System with Applications	Statistical	2002	Decision Tree
Van Den Poel and Lariviere	Belgium	European Journal of Operational Research	Operational Management	2004	Decision Tree
Baesen, Verstraeten, Van den Poel, Kenhove, and Vanthiemen	Belgium-Netherlands	European Journal of Operational Research	Operational Management	2004	Decision Tree
Nie, Zhang, Li and Shi	China	IEEE International Conference on Data Mining	Statistical	2006	Decision Tree
Bryson	United States	Expert System with Applications	Statistical	2008	Decision Tree

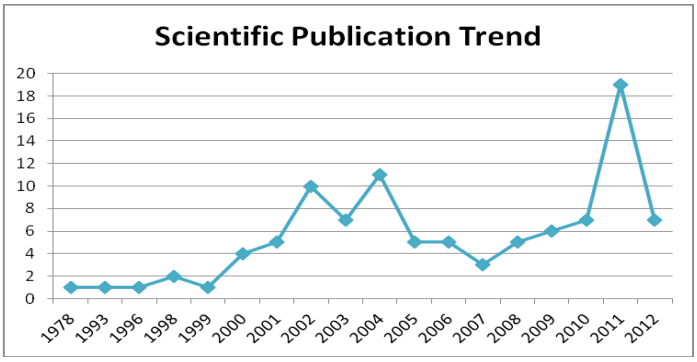
Wang, Chiang, Hsu, Lin,Cj. and Lin Il.	Taiwan	Expert System with Applications	Statistical	2009	Decision Tree
Risselada, Verhoef and Bijmolt	Netherlands	Journal of Interactive Marketing	Marketing	2010	Decision Tree
Han, Lu and Leung	China	Expert System with Applications	Statistical	2012	Decision Tree
Yu-Bao, Bao-Sheng and Xin-Quan	China	IEEE Computer Society	Operational Management	2011	Decision Tree
Hung, Yen and Wang	Taiwan-United States	Expert System with Applications	Statistical	2006	Decision Tree Neural Network
Mozer, Walniewicz, Grimes, Johnson and Kaushansky	United States	IEEE Transaction on Neural Networks	Statistical	2000	Decision Tree Regression Neural Network
Hwang and Euiiho Suh	Korea	Expert System with Applications	Statistical	2004	Decision Tree Regression Neural Network
Shiraishi and Fukumizu	Japan	Neurocomputing	Statistical	2011	Decision Tree Regression Neural Network
Shim, Choi and Suh	Korea	Expert System with Applications	Statistical	2012	Decision Tree Regression Neural Network
Sun, Bebis and Miller	United States	Patter Recognition Letters	Statistical	2004	Genetic Algorithm
Meyer-Base and Watzel	Germany	Patter Recognition Letters	Statistical	1998	Genetic Algorithm Decision Tree
Kavzoglu and Mather	Turkey-United Kingdom	Tailor & Francis	Statistical	2002	Genetic Algorithm Decision Tree
Datta, Masand, Mani and Li B.	United States	Artificial Intelligence Review	Statistical	2001	Genetic Algorithm Neural Network Decision Tree
Chen and Rosenthal	United States	International Journal of Industrial Organization	Operational Management	1996	Markov Model
Wei, Wang and Towsley	United States	Performance Evaluation	Management	2002	Markov Model
Avrachenkov and Sanchez	France	Fuzzy Optimization and Decision Making	Operational Management	2002	Markov Model
Jenamani, Mohapatra and Ghose	India	Electronic Commerce Research and Applications	Marketing	2003	Markov Model
Fleming	United States	Reliability Engineering and System Safety	Engineering	2003	Markov Model
Jonker, Piersma and Van del Poem	Belgium-Netherlands	Expert System with Applications	Statistical	2004	Markov Model
Slotnick and Sobel	United States	European Journal of Operational Research	Operational Management	2005	Markov Model
Dierkes, Bichler and Krishnan	United States	Decision Support System	Operational Management	2011	Markov Model
Au, Chan and Yao	China	IEEE Transaction on Evolutionary Computation	Statistical	2003	Neural Network Decision Tree
Zabkowski and Szczesny	Poland	Expert System with Applications	Statistical	2012	Neural Network Decision Tree
Bartleet	Australia	IEEE transaction of Information theory	Statistical	1998	Neural Network
Vellido, Lisboa and Meehan	United Kingdom	Expert System with Applications	Statistical	1999	Neural Network
Smith and Gupta	Australia-United States	Computer & Operations Research	Operational Management	2000	Neural Network
Bose and Mahapatra	United States	Information & Management	Operational Management	2001	Neural Network
Boone and Roehm	United States	International Journal of Research in Marketing	Marketing	2002	Neural Network
Hsieh	Taiwan	Expert System with Applications	Statistical	2005	Neural Network

Tsai and Lu	Taiwan	Expert System with Applications	Statistical	2009	Neural Network
Tsai and Chen	Taiwan	Expert System with Applications	Statistical	2010	Neural Network
Kisioglu and Topcu	Turkey	Expert System with Applications	Statistical	2011	Neural Network
Huang, Kechadi, Buckley, Kiernan, Keogh and Rashid	Ireland	Expert System with Applications	Statistical	2010	Neural Network Decision Tree Support Vector Machine
Buckinx and Van den Poel	Belgium	European Journal of Operational Research	Operational Management	2005	Neural Network Random Forests
Zhang, X., Zhu, J., Xu, S., Wan, Y.	China-United States	Knowledge-Based System	Operational Management	2012	Opinion Mining
Subrananiam, Faruquie, Ikbal, Godbole and Mohania	India	IEEE International Conference on Data Engineering	Statistical	2009	Opinion Mining
Coussement and Van den Poel	Belgium	Information & Management	Operational Management	2008	Opinion Mining
Burez and Van den Poel	Belgium	Expert System with Applications	Statistical	2007	Other
Chiang, Wang, Lee and Lin	Taiwan	Expert System with Applications	Statistical	2003	Other
Lertworasirikul, Fang, Joines, and Nuttle,	United States-Thailand	Journal of Operation Research	Operational Management	2003	Other
Cho and Kim	Korea	Fuzzy Sets and Systems	Statistical	2005	Other
Wang and Hong	Taiwan	Industrial Marketing Management	Marketing	2006	Other
Coussement and Van den Poel	Belgium	Decision Support System	Operational Management	2008	Other
Chiang, Wang and Chen	Taiwan	Knowledge-Based System	Operational Management	2010	Other
Chueh	Taiwan	African Journal of Business Management	Management	2011	Other
De Bock and Van den Poel	Belgium-France	Expert System with Applications	Statistical	2011	Other
Lin, Tzeng and Chin	Taiwan	Expert System with Applications	Statistical	2011	Other
Kim, N., Jung, Kim, Y., and Lee	Korea-United States	Expert System with Applications	Statistical	2012	Other
Rust and Zahorik	United States	Journal of Retail	Management	1993	Regression
Mihelis, Grigoroudis, Siskos, Politis, and Malandrakis	Greece	European Journal of Operational Research	Operational Management	2001	Regression
Athanassopoulos	United Kingdom	Journal of Business Research	Management	2000	Regression
Viaene, Baesens, Gestel, Suykens, Van del Poel, Vanthiemen and Dedene	Belgium	International Journal of Intelligent Systems	Operational Management	2001	Regression
Rygielski, Wang and Yen	Taiwan-United States	Technology in Society	Operational Management	2002	Regression
Bloemer, Brijs, Vanhoof and Swinnen	Belgium and Netherlands	International Journal of Research in Marketing	Marketing	2003	Regression
Van Den Poel	Belgium	Working Paper-Leuven University	Operational Management	2003	Regression
Kim and Yoon	Korea	Telecommunication Policy	Management	2004	Regression
Verhoef and Donkers	Netherlands	Decision Support System	Operational Management	2001	Regression
Yan, Walniewicz and Dodier	United States	IEEE Intelligence Systems	Statistical	2004	Regression
Auh and Johnson	Australia-United States	Journal of International Psychology	Psychology	2005	Regression
Ahn, Han and Lee	Korea	Telecommunication Policy	Management	2006	Regression
Coussement, Benoit and Van den Poel	Belgium	Expert System with Applications	Statistical	2010	Regression

Keramati and Ardabili	Iran	Telecommunication Policy	Management	2011	Regression
Samimi and Aghaie	Iran	Computer and Industrial Engineering	Engineering	2011	Regression
Lee, H., Lee, Y., Cho, Im and Kim	Korea-United States	Decision Support System	Operational Management	2011	Regression
Migueis, Van den Poel, Camanho, Falcao and Cunha	Portugal-Belgium	Expert System with Applications	Statistical	2012	Regression
Vasu and Ravi	India	International Journal of Data Mining, Modelling and Management	Operational Management	2011	Regression
Au, Guangqin and Wang	United States	IEEE transaction of Information theory	Statistical	2011	Regression
Guangli Nie, Wei Rowe, Lingling Zhang, Yingjie Tian and Yong Shi	China-United States	Expert System with Applications	Statistical	2011	Regression Decision Tree
Verbeke, Dejaeger, Martend, and Baesens	Korea-Belgium-United Kingdom	European Journal of Operational Research	Operational Management	2012	Regression Decision Tree
Huang, Kechadi, and Buckle	Ireland	Expert System with Applications	Statistical	2012	Regression Decision Tree Neural Network Vector Machine
Verhoef, Spring, Hoekstra and Leefland	Netherlands	Decision Support System	Operational Management	2002	Regression Neural Network
Buckinx, Verstraeten and Van den Poel	Belgium	Expert System with Applications	Statistical	2007	Regression Neural Network
Berne, Mugica, and Yesus	Spain	Journal of Retailing and Consumer Services	Management	2001	Regression Opinion Mining
Coussement and Van den Poel	Belgium	Expert System with Applications	Statistical	2009	Regression Support Vector Machine
Chen, Hsu, and Hsu	Taiwan	Expert System with Applications	Statistical	2011	Regression Support Vector Machine
Coussement and Van den Poel	Belgium	Expert System with Applications	Statistical	2008	Regression Support Vector Machine
Quian, He, and Wang	China	Journal of Management Science	Operational Management	2007	Support Vector Machine
Lessmann and Vob	Germany	European Journal of Operational Research	Operational Management	2009	Support Vector Machine
Yu, Guo, Guo, and Huan	China	Expert System with Applications	Statistical	2011	Support Vector Machine
Chen, Fan and Sun	China-United States	European Journal of Operational Research	Operational Management	2012	Support Vector Machine
Farvaresh, Mohammad and Sepehri	Iran	Engineering Application of Artificial Intelligence	Statistical and Engineering	2011	Support Vector Machine Neural Network Decision Tree
Xia and Jin	China	System Engineering - Theory and Practice	Engineering	2008	Support Vector Machines
Xie, Li, Ngai and Ying	China	Expert System with Applications	Statistical	2009	Support Vector Machines

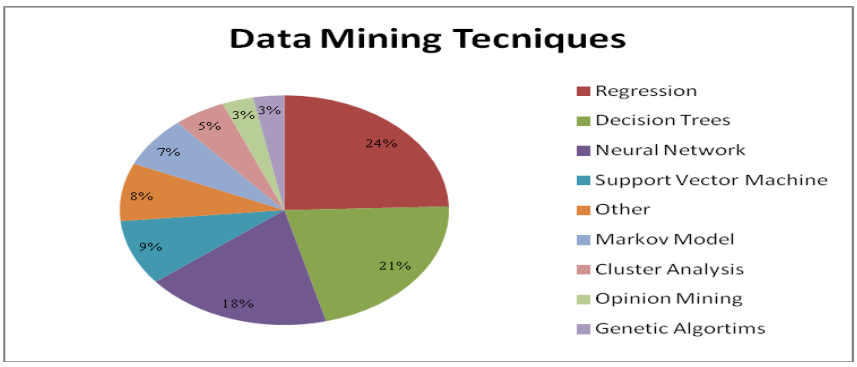
The following graph emphasizes the trend of the main churn management studies between 1978 and 2012.

Graph 2.1 Trend of the main churn management studies from 1978 to 2012



From the graph emerges an irregular trend of the year of the publication. Also, it is clear a considerable lack of studies from 1978 to 1993. Indeed, between 1999 and 2002 the number of contributions related this topic substantially increase rather than a slight reduction in 2003 followed by a rise in 2004. Additionally, from 2004 to 2007 a large decrease whit a slight growing in 2006 was occurred. Finally, in the range 2007-2010 the publication in the churn management field grow fast with a peak in 2011 and a consequence drop in 2012. Instead, Graph 2.2 provides the ordered percentage value related to the predictive data mining techniques analysed to manage the risk of churn.

Graph 2.2 Development of the Data Mining Techniques



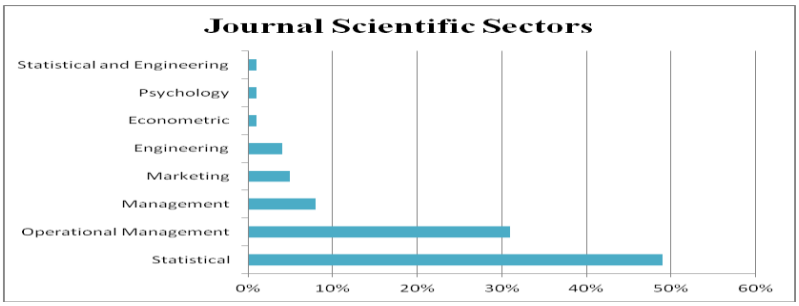
The graph underlines that the most applied data mining technique in managing customer churn is the regression analysis. In fact, for 24% of cases authors have studied this techniques to forecast the risk of churn. No less important are decision trees and neural networks analysis because of their development in analysing the churn defection is respectively to 21% and 18%. Besides, support vector machine techniques have a good implementation in forecasting the level of the risk of churn. Indeed, their development rate is equal to 9%. On the other hand, the percentage related to the opinion mining strategies in estimating the risk of churn is really alarming because of equal to 3%. In other words, from an academic point of view an opinion mining approach is irrelevant to improve the quality of the marketing predictions even though fundamental from a

managerial point of view. Particular attention must be paid on the group called 'Others[6]' because of includes different several statistical techniques used in managing the customer churn phenomena but not belonging to the data mining field. Finally, an analysis of the literature review shows that Markov models, cluster analysis, and genetic algorithm techniques seem to be underutilized to manage the customer churn. Additionally, Table 2.2 underlines the name of the first three scientific journals that contains a considerable number of studies related to the customer churn management. Due to the interdisciplinary of the customer churn management problem, Graph 2.3 provides a percentage distribution about the scientific sectors argued in the journals. Finally, Graph 2.4 shows in which geographical area this topic is more studied and developed.

Table 2.2 Scientific Journal: Descriptive Statistics

Name of Scientific Journal	Publication Percentage Value
Expert System with Applications	34%
European Journal of Operational Research	8%
Decision Support System	6%
Others	52%

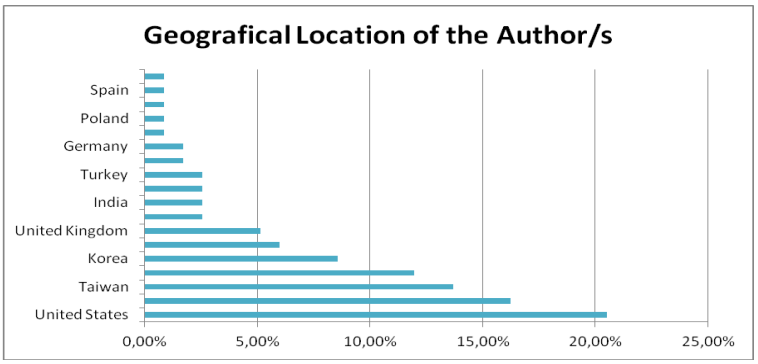
Graph 2.3 Journal Scientific Sector: Percentage Distribution



From the Graph 2.3 emerges that the most scientific sectors in which the customer churn defection is studied concerns the statistical field. Indeed, a considerable studies have been developed in the operational management area. On the other hand, the marketing field seem to be irrelevant for the analysis of this topic.

[6] Belong to this group techniques such as association rules, sequential pattern, retention analysis, fuzzy correlation analysis, logarithmic series distribution and model based on the Naïve Bays algorithms.

Graph 2.4 Geographical Location of the Author/s



Finally, research work in customer churn management are widely studied in United States followed by Belgium and Taiwan.

2.3.4 Validation of Results

Depending on the amount of data available there are several methods documented for testing the effectiveness of a customer churn model. The main methods argued in literature are the following:

- Cross-fold validation: it is based on the principle of using the available for both training and validation. Hwang and Euchiio Suh (2004) performed validation by creating a 70/30 divide of the data. Others cross-validation methods such as V-fold cross validation and Monte Carlo cross validation have been proposed in the literature. With the first model the learning set is randomly portioned into limited databases of equal size. Then, each set is used as validation set. Instead, in the second case, the learning set is repeatedly divided into two random sets for training and validation. Remark that the Cross-fold validation is an accurate method in presence of a scarcity of data.

- Using separated databases: it is helpful in presence of large amount of data and is more suitable in those cases in which data availability is not an issue. Bloemer, Brijis, Vanhoof and Swinnen (2002) and Prinzie and Van Den Poel (2004) used a separated datasets to evaluate their customer churn model. For instance, Datta, Masand and Mani (2001) validated their model by comparing their results against a simple regression model and decision trees with a validation composed by 17,000 records.

- Criteria based on the loss function: it is applicable when the target variable is dichotomous (Giudici, 2010) and it is very suitable in presence of huge databases for the validation of the customer churn model built. Van den Poel (2003) used a criteria based on the loss function in order to test the effectiveness and the accuracy of the customer-oriented conceptual model of segmentation variables for mail-order repeat buying behaviour. More precisely, the authors validated the model in term of the area under the ROC curve (AUC) obtaining a moderate performance, according to Swets (1988).

Chapter 3
Research Methodology

3.1 Introduction

The current literature distinguishes mainly between two types of research: quantitative and qualitative. In general, quantitative research are based on the traditional statistical models and methods, while qualitative research are build on interpretive and interactive strategies. Additionally, some researchers argue about a mixed research method based on triangulation theory[7]. But, despite its importance, it was never widely used in research for reasons not well defined. Probably, it has never been clear what its strengths are in the research context.

Today, due to the large amounts of data available within companies, traditional research based on quantitative or qualitative methodology done inadequate results in the development of accurate business and marketing strategies. Since, global companies have an emergent need to implement mixed research methods able to manage huge amounts of data in a short time. A modern quantitative methodology research is proposed in this thesis. Data mining is an astonishing methodology for data analysis inasmuch as the meaningful knowledge is often hidden in enormous databases and most traditional statistical methods could fail to uncover such knowledge.

This chapter follows a straightforward structure. Firstly, it explains the main differences between quantitative and qualitative research compared to data mining methodology. Secondly, it presents the research design proposed in this dissertation. Thirdly, it provides a description of the data mining process implemented in this work. Finally, it describes the forecasting models developed in this research.

3.2 Quantitative and Qualitative Research: A Theoretical Framework

Despite the considerable interest by many writers to integrate quantitative and qualitative research, in the current literature there is a clear distinction between quantitative and qualitative research (Bryman and Bell, 2003; Cooper and Shindler, 2005; Cameron and Price, 2009). Nowadays, these two different types of research should be viewed as complementary rather than as rival camps (Jick, 1979; Todd, 1979 and Tashakkari and Teddle, 2003). The division seems to be very restrictive because the same phenomena can be analyzed at the same time in different ways. For instance, global companies need to combine quantitative and qualitative research in order to get more accurate predictions and so better performance. In this case a good starting point could be to develop quantitative research based on data mining methodology in order to combine both quantitative and qualitative analysis.

Bryman and Bell (2011), Cameron and Price (2009), and Cooper and Schindler (2005) note many differences between quantitative and qualitative research. First, quantitative

[7] Triangulation is defined by Denzin (1978 p.291) as 'the combination of methodologies in the study of the same phenomenon'.

research focuses on numbers and graphs, while qualitative research focuses on words, sentences and narratives. Qualitative is the essential character or nature of something and qualitative research explores definition, analogy, model or metaphor characterizing something, while quantity is the amount assumes the meaning in known and needs only a measure of it. Third, both quantitative and qualitative researches have different epistemological foundations. Fourth, quantitative research incorporates the practices and norms of the natural scientific model and embodies a view of social reality as an external objective reality. Finally, quantitative and qualitative research form two different clusters of research strategy.

Today there is an emerging need to reduce the distance between quantitative and qualitative research and implement mixed methodologies in order to provide accurate marketing research. For instance, with a data mining perspective both quantitative and qualitative data can be scientifically treated and they can provide significant strategic information really helpful for decision makers. In addition, the latest developments in data mining have focused the attention on the text mining analysis with the main aim to manage faithfully qualitative data, in particular textual data from web marketing context. A large amount of qualitative data can be converted to quantitative information through appropriate statistical techniques without losing significant information. This qualitative data could represent the key drivers for global companies with an outside-in perspective for getting strategic decisions, especially in over-supply market situation. In fact, due to the large amount of data available, businesses must combine both quantitative and qualitative research in order to get a competitive advantage.

The following table shows the differences between quantitative and qualitative research identified by some authors (Halfpenny, 1979; Bryman, 1988*a*; Hammersley, 1992) and proposes a new type of research called data mining research.

Table 3.1 Quantitative and Qualitative Research versus Data Mining Research

Quantitative Research	Qualitative Research	Data Mining Research
Numbers	Words	Numbers and Words
Point of view researchers	Point of view of participants	Point of view of researchers and participants
Researcher distant	Researcher close	No distance between researchers and the population under study
Theory testing	Theory emergent	Scientific and Interpretive Approach
Static	Process	Dynamic Process
Structured	Mainly Unstructured	Structured and Unstructured
Generalization	Contextual understanding	Results always generalizable
Hard, reliable data	Rich, deep data	Any type of data

Usually Macro	Micro	Macro and Micro
Behavior	Meaning	Behavior and Meaning
Artificial settings	Natural settings	Artificial and Natural settings

Source: personal author elaboration adapting from Bryman and Bell, 2011

According to Bryman and Bell (2011) given the importance of these different researches a brief description of these is provided below.

- **Numbers versus Words:** quantitative research analyze number (quantitative data) and are based on a scientific process, while qualitative research are built on verbal data (words, text, narrative, etc.) and concern personal opinions related to a given phenomena. Instead, in a data mining perspective researchers can analyze jointly numbers, words, sentences and text by apposite techniques. For instance, it can combine text mining and opinion mining with quantitative methods.

- **Point of view of researchers versus Point of view of participants:** quantitative researchers as a first stage explore the data; instead in qualitative research the perspective of those being studied provides the point of view. However, data mining research considers as a starting point the results of the exploratory analysis combined with the opinion of the participants.

- **Researcher is distant versus Researcher is close:** quantitative researchers believe that too many relationships with the people studied could lead to reduce the objectivity of the results obtained. But, qualitative researchers suggest that a close communication with the people studied leads to better in results. Data mining research there is less distance between the researcher and the population studied because of the quality of the research is always evaluated through many different perspectives (quantitative and qualitative point of view). Also, the relationship between researchers and the population studied is very important because the personal opinion (opinion mining) could represents a critical success factor for researchers.

- **Theory and concepts tested in research versus Theory and concepts emergent from data:** in quantitative research the theoretical work precedes the collection of data, while in qualitative research concepts and theoretical elaboration emerge out of data collection. In a data mining perspective these two aspects are carried on at the same time.

- **Static versus Process:** quantitative research is focused on the relationship between variables and it is usually seen as a static image of social reality. However, qualitative research is described as 'attuned to the unfolding of event over time and to the interconnections between the actions of participants of social settings' (Bryman and Bell, 2003 p. 426). Data mining research can seem static image of the social reality but automatically they incorporate the interconnections between the actions of participant to social settings.

- **Structured versus Unstructured:** in a quantitative research it is possible to analyze with high accuracy the precise concepts and issues that are the focus of the study, while in a qualitative research the approach is invariably unstructured, thus it is improbable to analyze with high accuracy the precise concepts and issues that are the focus of the study. On the contrary, data mining research achieve to convert unstructured approach in structured so analyzing with high accuracy each element of the research.

- **Generalization versus Contextual understanding:** the results obtained from a quantitative analysis are extensive to the relevant population; instead the results related to the qualitative research are just valid in the field of the research. The results of data mining research can extend to the population studied regardless of the data source.

- **Hard, reliable data versus Rich, deep data:** due to the precision provided by measurement, quantitative data are defined as 'hard' in the sense of robustness and no ambiguity. On the contrary, qualitative researchers claim that their contextual approach engender rich data. In the data mining research there is not distinction among the features of the data because of any type of data can be analyzed.

- **Macro versus Micro:** quantitative research are usually developed in a macro perspective because of the main aim of these is to discover in a large scale trends and relationships between variables, by contrast, qualitative research refers small-scale aspects of social reality such as interaction between events. Conversely, research based on data mining methodology cover either small and large social trends or aspect of social reality.

- **Behavior versus Meaning:** quantitative research are normally concerned with people behavior while, qualitative research are based on the meaning of action. Data mining research foresees both behavior and meaning.

- **Artificial settings versus Natural settings:** while quantitative research is based on the artificial context, qualitative research is focused in natural environments. Research based on data mining methodology refers both artificial and natural settings.

In conclusion, quantitative research based on data mining methodology can integrate both quantitative and qualitative information because of qualitative information can be converted in quantitative information through specific data mining application (for instance: text mining). Finally, researchers must pose particular attention on qualitative information because they could be representing a critical success factor for decision makers in today's global business.

3.3 Research Design

In management area an adequate choice of research design it is fundamental for companies because it reflects decisions about the priority being given to a range of

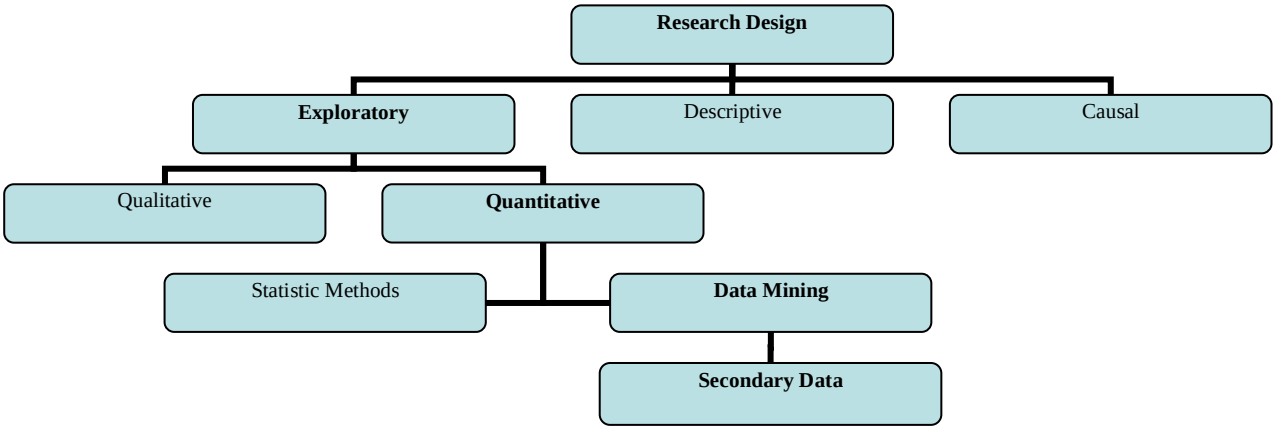
dimensions of the research process. For example some significant dimensions are provided below:

- Causal connection between variables.
- Understanding behavior and the meaning of that behavior in its specific social context.
- Having a temporal appreciation of social phenomena and their interconnections.

A good research design ensures the consistency of data gathered and the accuracy about the procedures used for collecting data.

According to Cooper and Schindler (2005), the following graph shows the research design developed in this PhD thesis.

Graph 3.1 Research Design



3.4 Research Purpose

According to Saunders et al. (2000) research can be classified on the base of their purpose. In fact in the current literature are described three types of quantitative research:

- **Explorative**. It sheds light on the nature of a situation and identifies any specific data needs to be addressed through additional and news research. This research is fundamental in management area, especially, when marketers implemented a data mining analysis. More precisely, exploratory research is most useful when a decision maker wishes to better understand a situation and/or identify decision alternatives. Also, exploration is particularly useful when researchers lack a clear

idea of the problem they will meet during the study. Finally, exploratory studies establish causal relationship between variables.

- **Descriptive.** It describes the market features or functions. This research could be connect with the exploratory research because of researchers might have started off by wanting gain insight to a problem and after having started in their research becomes descriptive (Saunders et al. 2000). In contrast to exploratory studies more formalized studies are typically structured with clearly stated hypothesis or investigate questions (Cooper and Schidler, 2005). In particular descriptive studies serve a variety of research objectives such as (1)description of phenomena or features associated with a subject population, (2)estimates the proportions of a population that have the previously characteristics, and (3)discovery of associations among different variables.

- **Explanatory or Causal.** Marketers should discover casual relationships between variables. They use this approach primarily for purposes of prediction and to test hypotheses, though it can also be used to a lesser extent for discovery and explanatory purposes. In marketing, causal research is used for many types of research including testing marketing scenarios, such as what might happen to product sales if changes are made to a product's design or if advertising is changed. If causal research is performed well marketers may be able to use results for forecasting what might happen if the changes are made.

According to Giudici (2003) 'Data mining is the process of selection, exploration, and modeling of large quantitative of data to discover regularities of relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database'. Data Mining is a useful tool that joins exploration and discovery with confirmation analysis. Since the focus of this research is data mining so the purpose of this research is exploratory.

3.5 Research Approach

There are two main research approaches to choose from when conducting research in social science: quantitative and qualitative methods (Yin, 1994). As said before, there are a lot of differences between these two types of methods. The main difference is that quantitative methods are based on numbers and statistics while qualitative methods refer to personal opinion without consider a scientific process.

Exploration relies more heavily on qualitative techniques but in some cases there exist sufficient data to allow data mining or exploration of individual measurements to take place (Malhatra and Birks, 2003).

The focus of this study is data mining thus it is clear that the research approach of this research is quantitative.

3.6 Research Strategy

The research strategy can be defined as a particular way to collect data by researchers. The way of collecting data depends on the features of the research question. According

to Yin (1994) the researchers for gathering data can choose among many approaches such as a survey, history, secondary data analysis, case study, etc.
In general, primary or secondary data conduct a business research. Primary data are originated by a research for a specific purpose of addressing the problem at hand. These data cannot divide in internal or external data. Indeed, secondary are data that have been collected for purposes other than the problem at hand. For instance, these data include data generated within a company and information made available by business and private or public sources. These data can be dividing in internal and external. Internal data are generated insight the corporation in relation to specific business problems, by contrast, external data are generated by external sources outside the firms (Malhatra and Birks, 2003). In our case we will analyze external secondary data because of our data come from by an external databases that matches digital campaign exposure data with web site sales and customer lifetime value metrics for Fortune 100 advertiser.

In conclusion, the purpose and approach of this thesis is an explorative quantitative research and the research strategy is the analysis of the secondary data.

3.7 Research Process

Berry and Linoff (2011) define data mining as a business process for exploring large amounts of data to discover meaningful patterns and rules. According to Giudici (2009) from an operational point of view data mining can be defined as a process composed by the following phases:

- **Identification of the business problem**: decision makers must be identifying the business problem. In this phase data miners and decision makers work closely in order to define the project objectives and the requirements from a business perspective. The project objective is then translated into a data mining problem definition. In the problem definition phase, data mining tools are not yet required.

- **Selection, organization and pre-treatment of the data**: at the first step, it is fundamental to define the data sources. Usually, companies take data from internal sources because are the most cheaper and reliable data channel. After the identification of the database and the creation of a data matrix is often necessary to carry out a process of preliminary cleaning of the data. This is a formal process used to find possible outliers so variable that exist but are not suitable for analysis. Finally, before starting the data exploration phase, it is necessary to check the contents of each variable in order to identify the possible presence of missing or incorrect data**.**

- **Data exploration and transformation**: data miners collect, cleanse, and format the data because some of the mining functions accept data only in a given format. Also, they identify quality problems of the data. Traditional data analysis tools, for instance position and dispersion index, are used to explore the data. This phase is essential because it allows at the data miners and business analysts to select the most appropriate statistical methods for the next phase of the analysis. This choice must consider the features and the quality of the data available. If the data collected are insufficient for the purpose of the analysis, data miners can be make a new data extraction.

- **Identification of statistical methods**: there are various statistical methods and many algorithms available for data mining analysis. The choice of which method to use in the analysis is related to the problem being studied or on the type of data available. The data mining process is guided by more applications. For this reason, the classification of the statistical methods depends on the analysis's aim. In general it is possible to divide the main methods in two groups: descriptive and predictive. The main aim of the descriptive methods is to describe groups of data in different ways. In addition, in descriptive methods there are no hypotheses of causality among the available variables. Instead, the main objective of the predictive methods is to describe the relation between one or more variables. In this case, in order to predict or to classify the future events data miners can be adopt classification rules or predictive models based on the data available. The main patterns for this purpose are developed in the field of machine learning, such as neural networks and decision trees, or classical statistical models as linear and logistic regression models. This dissertation focuses the attention on the predictive models, in particular on hierarchical logistic regression and classification decision trees models contextualized in a data mining point of view.

- **Modeling**: data miners select and apply various mining functions because they can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data. Also, they must evaluate each model. In the modeling phase, a frequent exchange with the domain experts from the data preparation phase is required. The modeling phase and the evaluation phase are coupled. They can be repeated several times to change parameters until optimal values are achieved. When the final modeling phase is completed, a model of high quality has been built.

- **Evaluation and comparison of the methods used and final choice**: data miners evaluate the model. If the model does not satisfy their expectations, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved. In fact, to produce a final decision is necessary to choose the best model among the various statistical methods available. The choice of the model is based on the comparison of the results obtained. Sometimes for optimizing the results obtained should be better to develop more than one method.

- **Interpretation of the model chosen and its use in the decision process:** in this phase analysts decide how to use the data mining results and the marketers, after a significant interpretation of these, draw new strategic business decision. In fact, data mining is more than a simple statistical data analysis. This phase is really important for global companies because decision makers must be complete the analysis with an opinion mining approach integrating the results with a text-mining analysis. Sometimes the opinion mining oversets the results of quantitative analysis. 'Qualitative forecasts are often based on the judgment of a single expert or represent the consensus of a group of experts' (Williams and Shoesmith, 2009 p. 716). In this process, the experts individually consider information that they believe will influence the analysis results, then they aggregate their conclusions into a forecast. No formal model is used, and no two experts are likely to consider the same information in the same way. This method could provide good forecasts in many situations. Additionally, decision makers can be considering others qualitative

methods such as Delphi method[8], Scenario writing[9] and Intuitive approaches[10], according to Anderson, Sweeney, Williams, Freeman, and Shoesmith (2011) definitions. After this qualitative evaluation, decision makers must choose and test the best model and to generalize the classification rules. Obviously, the inclusion of the data mining process within global organizations must be done gradually, setting out realistic aims and looking at the result along the way. The final aim for data mining is to be complete integrated with the other activities that are used to support organization decisions.

During the process of integration it is possible to identify four stages:

- Strategic phase. Data miners study the business procedures for identifying where data mining could be more helpful. At the end of this phase decision makers can determinate the business objective and to set up the data mining pilot project. Also, they can establish the criteria for evaluating the project activated.

- Training phase. This phase allows evaluating the data mining activity with more accuracy. A fundamental aspect of the implementation of a data mining procedure is the choice of the pilot project. In fact, in order to generate success, the data mining project must provide interest against company's workers.

- Creation phase. In this phase it is necessary to reorganize the business database and if possible to create a data warehouse. Then, it is important to develop the previous data mining prototype until the migration phase and to allocate personnel and time to follow the project.

- Migration phase. At this stage all people must to be ready for change. The data mining process must be implemented with success.

Finally, in order to develop an optimal data mining process decision makers must embroil minimum three different people or better three business functions:

- Marketing and Finance Department: marketing and business analysts and experts must set the objective of the analysis and they must interpret the data mining findings.

- Information Technology Department: information technology experts must implement the necessary technologies for developing an accurate data mining analysis.

[8] The goal of the Delphi method is not to produce a single answer as output, but instead to generate a relatively narrow spread of opinions within which the majority of experts concur.

[9] Scenario writing consists of developing a conceptual scenario of the future based on a well-defined set of assumption. In fact, decision makers decide how likely each scenario is and then to make decision accordingly.

[10] Intuitive approaches are based on the ability of the human mind to process a variety of information that, in most case, is difficult to quantify. Individual are freed from the usual group restrictions of peer pressure and criticism because they can present any idea or opinion without regards to its relevancy and without fear of criticism.

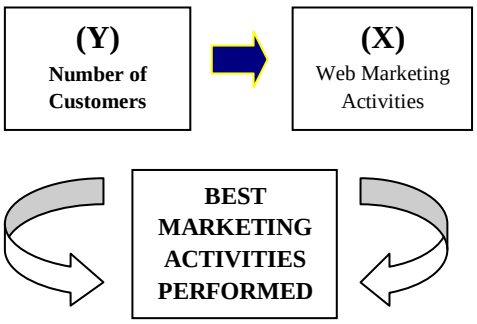
3.7.1 Linking Data Mining Process to Research Questions

This subsection shows the linking between the phases of data mining process developed in this dissertation.

- **Identification of the business problem:** by indentify a predictive model able to estimate with high accuracy the probability of customer conversion in web marketing sector.
- **Selection, organisation and pre-treatment of the data:** definition of the aggregations criteria to obtain the final database and develop of the first investigation of the data in order to discover anomalies in the data.
- **Data Exploration and Transformation:** multicollinearity analysis based on the Pearson correlation and multiple linear regression analyses to detect outliers in the data.
- **Identification of Statistical Methods:** choice of the best predictive methods for predicting accurate business performance in competitive landscape.
- **Modelling:** hierarchical logistic regressions and classification decision trees because of are the main models able to predict with high accuracy the main marketing activities which could lead potential customers to purchase the service proposed by the web marketing campaign.
- **Interpretation of the model chosen and its use in the decision process:** statistical interpretation joined with opinion mining and expert judgement.

The main objective of this PhD thesis is to identify the best data mining model for global companies with an outside-in perspective able to predict accurate web marketing performance in today competitive landscape. Special attention is paid on the estimation of the potential customer churn risk and on the customer conversion probability. The goal is to identify, through independent variables, the main web marketing activities that have been performed by a potential customer in order to increase the number of sales in the next web marketing campaign. Figure 3.1 proposes a general overview of the predictive model used in this research.

Figure 3.1 General overview of the Predictive Model



Before explaining the results of the analysis it is useful to provide an overview of computational and data mining forecasting models proposed subsequently.

3.8 Predictive Data Mining Models

The main purpose of this section is to explain the strategic importance of the predictive data mining models for global companies in today's competitive landscape for minimizing the number of churners and maximize the sales level. Particular attention is paid on the General Linear Models (GLM) such as Simple and Multiple Linear Regression, Hierarchical Logistic Regression and Decision Trees Models because of, as shown in the previously chapter, they seem to be accurate tools to estimate the risk of churn and the probability of customer conversion within global organizations. According to Berry and Linoff (2011), Figini and Giudici (2009) and Tan, Steinbach and Kumar (2005) these models and their evaluation criteria can be described as follows.

3.9 Linear Regression

The Linear Regression Model is adopted in business contexts when decision makers want study the dependence of the target (dependent) variable on one or more independent (exploratory) variables. The target variable must be quantitative continuous. This subsection is dividing in two parts. Firstly, it considers the relationship between the target variable and one independent variable. After, it explains the multivariate case with more independent variables. Finally, it describes the criteria used to evaluate the goodness of fit of these models.

3.9.1 Simple Linear Regression and the Correlation Index

The Bivariate Linear Regression could be useful into the decision making process for estimating the force of the relationship between one target or dependent variable and one independent variable. In this case the linear regression relationship is the following:

$$y_i = a + b_1 x_{i1} + e_i \qquad (i = 1,2,...,n)$$

where:

a = intercept of the regression function;
b = slope coefficient of the regression function or regression coefficient;
e_i = random error of the regression function related to the i observations

In details, the first part of the linear regression function is called the regression line, while the second part is denominated the error term. This latter indicate the level of approximation of the regression line to the target variables. Indeed, the regression line is the linear function and can be express in this way:

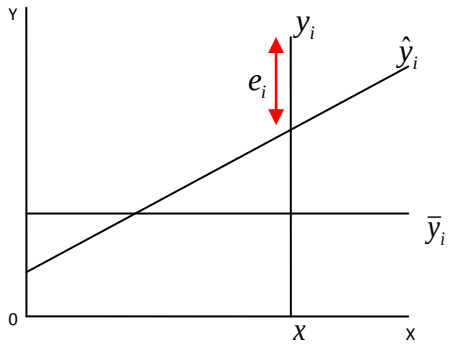
$$\hat{y}_i = a + bx_i \qquad (i = 1,2,\ldots,n)$$

where:

$\hat{y}_i$ = estimate value of the dependent variable

e_i = residual for each observations. It is equal to the difference between the observed response value y_i, and the corresponding values fitted with the regression line ($\hat{y}_i$). In a mathematical form the error term can be write as $e_i = y_i - \hat{y}_i$. Each residual can be considered as the part of the value that the regression model does not explain.

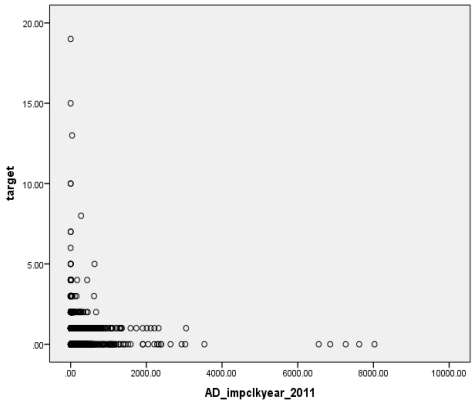
Graph 3.2 Description of the Regression Line



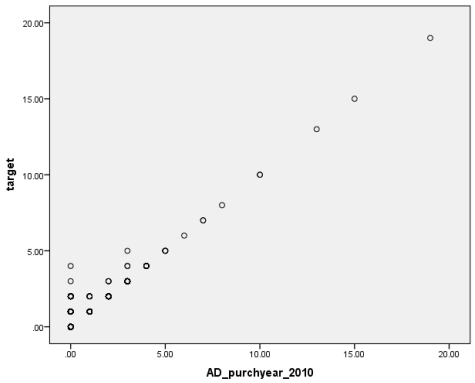
Source: Adaptation from Applied Data Mining, Giudici 2003

In business contexts usually decision makers tend to used only graphic representation (for instance: scatter plot) for understanding the possible relationship between two variables. The following graphs show cases of low correlation and strong correlation between two variables. Graph 3.3 shows the relation between the number of potential customers who purchased the service online and the number of impressions that a potential customer made. Indeed, Graph 3.4 provides an example of a strong correlation between the target variable and the variable related to the number of customer conversions in the first part of the launch of the marketing campaign.

Graph 3.3 No Correlation between variables



Graph 3.4 Strong Correlation between variables



These graphical representations provide an important subjective assessment about the strength of the linear relationship between two variables but it is not enough in order to predict accurate business performance.

In order to improve the accuracy of the predictions decision makers should develop bivariate statistical indexes that summarize the frequency distribution. Since, it is useful to explain the concept of covariance measure in order to better understand the importance of the correlation index. According to Giudici (2010) the concordance between two variables is the tendency of observing high or low values of a variable together with high or low values of the other. However, discordance is the tendency of observing low or high values of a variable together with high or low value of the other.

This measure is defined as follows:

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}[x_i - \mu(X)][y_i - \mu(Y)]$$

where:

$\mu(X)$ = mean of the variable X

$\mu(Y)$ = mean of the variable Y

The covariate is an explorative index that identifies the presence of a relationship between two quantities without explain the level (degree) of relationship between them. It gets positive value if the variables are concordant and negative values if they are discordant. In particular its maximum value is equal to the product of the two standard deviations of the variables $(\sigma_x\sigma_y)$, while its minimum value is equal to $(-\sigma_x\sigma_y)$.

Additionally, the covariance measure takes its maximum value when the line where lie the observed data has positive value, instead the measure gets its minimum level when the line where lie the observed data has negative slope. For this reason it is worthwhile define the Pearson Correlation Coefficient (PCC) between two variables as follows:

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

where:

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}[x_i - \mu(X)][y_i - \mu(Y)]$$

$\sigma(X)\sigma(Y) =$ product of the two standard deviation of the two variables

The PCC is used as a measure of the strength of the association between two variables. It assumes value in the range -1 and 1. It is equal to -1 when there is a strong inverse correlation between two variables. Also, if the coefficient is 0 means that there is no linear correlation between the two variables, while if it is 1 implies perfect correlation. In this case all data points lie exactly on the regression line. In other words, it means that there is not dispersion around the regression line indicating that the dependent variables are perfectly forecast by the independent variables. Indeed, if the correlation coefficient assumes value in the range 0 and 1 decision makers can interpret this value as shown by the following table:

Table 3.2 Interpretation of the Correlation Coefficient

Pearson Coefficient Value	Value Interpretation
0.10 < r < 0.29	Small Relationship
0.30 < r < 0.49	Medium Relationship
0.50 < r < 1.00	Large Relationship

In conclusion, the PCC measures the degree of the relationship between two variables and not causes and effects. Also, it detects only the linear relationships between two variables. Finally, it is good to estimate the linear relationship between two variables but in many cases it is not useful because of many variables could have a non linear relationship.

3.9.2 Estimate of the Best Line Fit

The parameters *a* and b must be calculate by decision makers in order to find the line of the best fit through the least square methods. This method is one of the most common techniques used for this purpose especially in management area. It chooses the straight line that minimizes the sum of the squares of the errors of the fit (SSE), defined by:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_{i-} - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

To calculate the minimum of SSE statisticians must calculate the first partial derivates of the SSE function with respect to *a* and *b* then equate them to zero. These parameters can be found through these normal system of equations:

$$\frac{\partial \sum (y_i - a - bx_i)^2}{\partial a} = -2\sum_i (y_i - a - bx_i) = 0$$

$$\frac{\partial \sum (y_i - a - bx_i)^2}{\partial b} = -2\sum_i (y_i - a - bx_i) = 0$$

It is clear that the results of the first equation are the following:

$$a = \sum \frac{y_i}{n} - b \sum \frac{x_i}{n} = \mu_x - b\mu_x$$

Since substituting the factorization of the term *a* in the second equation it is possible estimate the *b* value that is equal to

$$b = \left(\frac{\sum x_i y_i / n - \sum y_i \sum x_i / n^2}{\sum x_1^2 / n - \left(\sum x_i / n \right)^2} \right) = \frac{Cov(X,Y)}{Var(X)} = r(X,Y) \frac{\sigma_y}{\sigma_x}$$

where μ_y and μ_x are the means, σ_y and σ_x the standard deviations of the variables Y and X, and $r(X,Y)$ is the correlation coefficient between X and Y.

At this stage, it is meaningful to introduce and explain the analysis of the variance of the dependence variables so as to evaluate the results of the regression model previously described.

3.9.3 Evaluation of the Goodness of Fit

The aim of this section is to describe the main index able to evaluate the degree of approximation of a regression line. The index shows following is based on a decomposition of the variance of the dependent variable. By applying Pythagoras' theorem it is possible obtain:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

where:

$\sum (y_i - \bar{y})^2 =$ total sum of square (SST)
$\sum (\hat{y}_i - \bar{y})^2 =$ total sum of square explained by the regression (SSR)
$\sum (y_i - \hat{y}_i)^2 =$ total sum of square of the error (SSE)

This identity establishes that the total sum of square (SST) equals the sum of squares explained by the regression (SSR) plus the sum of squares of the errors (SSE). In other words, the previously identity can be rewrite as follows:

$$SST = SSR + SSE$$

These three quantities are called deviances. Dividing deviances by the number of observations and indicating the statistical variables with the corresponding capital letters it is possible obtained that $Var(Y) = Var(\hat{Y}) + Var(E)$. In this case a decomposition of the variance of the target variable in two components is carried on.

The variance has been divided in the variance 'explained' by the regression line, and the 'residual' variance. As a consequence of this the index of determination R^2 becomes the main index able to evaluate the goodness of fit of the regression line. This index is defined as follows:

$$R^2 = \frac{Var(\hat{Y})}{Var(Y)} = 1 - \frac{Var(E)}{Var(Y)}$$

R^2 coefficient is the square of the correlation coefficient and it takes value between 0 and 1. If the coefficient assumes value equal to 0 means that the regression line is constant (the slope of the regression line is 0 thus $Y = \bar{y}$) while, if the coefficient is equivalent to 1 the fit of the regression line is perfect. In this case all residual ($(y_i - \bar{y}_i)$) are null. A good value of R^2 coefficient means that the Y variable can be well forecasted by a linear function of X.

Statisticians can define R^2 as a summary index able to forecast the degree of the relationship among variables. Sometimes decision makers in order to understand where the regression line approximates the observed data well and where the approximation is poorer can develop diagnostic graphical measures. If the linear regression model is exactly the Y points will be distributed close to the fitted line in a random way and they will not indicate particular trends. Finally, it is important remember that the determination of the regression line could be strongly affect by the anomalous values called outliers. In this case decision makers must be better exploring the data to eliminate anomalous observations.
For a major accuracy of the analysis before getting a strategic business decision marketers must to evaluate the statistical significance of R^2 coefficient. In statistic field there are a variety of ways for evaluating the value of the determination coefficient. One of the most common methods is to develop a test t on b parameters (slope of the regression line) so as to understand in which measure the independent variable affects the target variable. For instance, decision makers could set up a hypothesis test for determining whether the b term in the equation could actually have come from a statistical population where the slope (β) was actually zero. If this supposition is true it could indicate that there is not statistically significant relationship between the two variables. On the contrary, if decision makers reject this hypothesis it means that b was not zero thus there was a significant relationship between the two variables. In a statistical and mathematical way the t test can explain in the following form:

$$t_{test} = \frac{b - \beta}{SE_b}$$

In other words, the t_{test} is equal to the difference between the sample value (b) and the hypothesis population value (β) dividing by the standard error of b. In details, the null hypothesis is formulated such that decision makers assume there is no relationship between the two variables (X and Y) thus $\beta = 0$. The significance level is set at 0.05

and decision makers must use t distribution with $(n-2)$ degree of freedom. Decision makers must understand that the use of this test is fundamental before developing a regression models in business forecasting process because of it is a formal hypothesis test able to evaluate the statistically sound of the regression equation (Wisniewski, 2008). Even then, this no guarantee that the regression line equation will be reliable in a forecasting sense in today's global business. Since, once again, is underlined the crucial importance of the qualitative approaches as opinion mining and expert judgments, especially in global companies.

The next subsection introduces the multiple regression models in order to shed light its strategic importance in real and complex business situations.

3.10 Multiple Linear Regressions

Multiple regression models are particularly complex both for their structure and for the statistical assumptions which underpin them. These models seem to be accurate and appropriate for forecasting phenomena in many business situations. Sometimes global companies use hasty these models lead enormous negative business consequences.

Decision makers can suppose that the database analyzed is composed by one target variable and many independent variables indicate with the parameter k. The multiple linear regression is defined by the following relationship:

$$y_i = a + b_1 x_{i1} + b_2 x_{i2} + b_l x_{il} + ... + b_k x_{ik} + e_i \qquad (i = 1,2,...,n) \text{ and } (l = 1,2,...,k)$$

where:

a = the intercept of the regression function;
b_l = the slope coefficient of the regression function of the independent variable x_l;
e_i = the random error of the regression function related to the i observations

This model is an extension of the previously model. Due to the very complicated calculations behind this model a substantial computer aid is required.

Today there are much free and paid statistical software able to perform these results. It is clear and obvious that the principle behind the method is the same of the simple linear model. In fact, for estimating the relationship that best fits the data across all X variables the method of least squares is used. Finally, the general results of the model are evaluated in principle in the same way described previously.
An important strengths of the regression model is that it is possible to determinate the partial net contribution of each independent variables. Decision makers must be evaluate statistically the overall goodness of fit the equation (R^2) through the F_{test}. In this case, decision makers must suppose that the b term in the equation are none statistically different from zero at every time.

Since if at the end of the test decision makers reject H_0 it implies that one of the X variables is statistically significant at least, by contrast, if H_0 is accepted it means that none of the X variables are statistically significant. In other words, Y is

statistically independent of the X variables and the estimated equation in not reliable. In a statistical and mathematical way the F_{test} is express as follows:

$$F = \frac{MSR}{MSE}$$

where:

MSR is equal to the Mean Square of Regression.

MSE is equal to the Mean Square of Error.

The F value must be assessing by decision makers on the bases of its critical value obtained from appropriate statistical tables. If the F value is greater than the critical value decision makers must be reject H_0, thus there is a significant relationship between Y and X variables. On the contrary, there is none relationship between Y and X variables when H_0 is true. In this case, the regression equation cannot be implemented in the forecasting process because it is not real from a business perspective. A real important evaluation in the multiple regression model concerns the examination of each b_l parameters. It may well be than even though the equation overall is statistically significant some parts of this could not be. For example, in a regression equation composed by five independent variables (X) some of these could not be statistically significant. In fact, in order to understand which variables are related to the target variable decision makers can develop a single t_{test} for each variable. The concept on the base of this test is the same explained for the simple linear regression. Anyway, the t_{test} is equal to $\frac{b_l}{SE_{bl}}$. In other words, the t_{test} can be calculated dividing the estimated coefficient b_l by its standard errors SE_{b_l}. Decision makers can be associated for parameter β_l the following hypothesis:

$$\begin{aligned} H_0 \beta_l &= 0 \\ H_1 \beta_l &\neq 0 \end{aligned}.$$

As said before, if the t_{test} value greater than the critical t value decision makers reject H_0 and they must conclude that β_1 is significantly different from zero. So, the variable considered is statistically significant in the regression model.

From a statistical point of view, the F_{test} is an important test to evaluate the general goodness of fit of the linear regression even if it does not describe the single contribution given from each variable to the regression equation. Since, for a major accuracy of the analysis, decision makers should calculate the t_{test} for each variable in order to estimate their single contribution to the regression equation. But, despite the accuracy of these tests, sometimes they could appear not useful in a managerial reality

for the following reasons. For instance, they seem to be expensive in term of time. Businesses must be optimize the forecasting process in order to monitor continuously each shake up in the market in order to arrive *before and better than competitors*. Then, the results obtained could be difficult to interpret by decision makers because of it requires high competence in statistical field.

Before introducing the Logistic Regression Model it is fundamental to describe the concept related to the association measures for qualitative variables focusing the attention on the concept of the Odds-Ratio (OR). This class of easily interpretable indices both for academic and managerial reality do not depend on the marginal distributions but it is based on probabilistic models. According to Figini and Giudici (2009) the Odds-Ratio Measure can be defined as follows.

3.11 Odds-Ratio Measure

Consider X and Y dichotomous variables respectively associated with the rows $(X = 0,1)$ and columns $(Y = 0,1)$ of a 2 × 2 contingency table. Let $\pi_{11}, \pi_{00}, \pi_{10}$ and π_{01} indicate the probabilities that one observation is classified in one of the four cells of the table. 'The odds ratio is a measure of association that constitutes a fundamental parameter in the statistical models for analysis of qualitative data' (Giudici, 2003 p.59). Let $\pi_{1|1}$ and $\pi_{0|1}$ indicate the conditional probabilities of a having 1 (success) and a 0 (a failure) in row 1; let $\pi_{1|0}$ and $\pi_{0|0}$ be the same probabilities for row 0. The odds of success for row 1 are defined by

$$odds_1 = \frac{\pi_{1|1}}{\pi_{0|1}} = \frac{P(Y = 1 \mid X = 1)}{P(Y = 0 \mid X = 1)}$$

Indeed, the odds of success for row 0 is defined by

$$odds_0 = \frac{\pi_{1|0}}{\pi_{0|0}} = \frac{P(Y = 1 \mid X = 0)}{P(Y = 0 \mid X = 0)}$$

The ratio between the two previous odds is the odds ratio:

$$\theta = \frac{odds_1}{odds_0} = \frac{\pi_{1|1} / \pi_{0|1}}{\pi_{1|0} / \pi_{0|0}}$$

From the definition of the odds and using the definition of joint probability, it can easily shown that $\theta = \frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}}$. In other words, this means that the odds ratio is a cross product ratio that is the product of probabilities on the main diagonal to the product of the probabilities on the secondary diagonal of a contingency table. In the actual

computation of the odds ratio, the probabilities will be replaced with the observed frequencies leading to the following expression: $\theta_{ij} = \frac{n_{11}n_{00}}{n_{10}n_{01}}$.

An important remark is that the odds are always a non-negative quantitative because of is defined in the interval $[0,+\infty]$. Also, when X and Y are independent $\pi_{1|1} = \pi_{1|0}$ thus $odds_1 = odds_0$ and $\theta = 1$. On the other hand, depending on whether the odds ratio is greater or less than 1, decision makers can assess the sign of the association. In fact, if $\theta > 1$ there is a positive association between the variables while, if $0 < \theta < 1$ it means that there is a negative association between them. The odds ratio deals with the variables in a symmetrical way thus it is not necessary to identify the dependent variable and the other as independent.

Finally, decision makers could estimate the odds ratios as well for larger contingency tables. In this case, the odds ratio for $I \times J$ tables can be defined with reference to each of the $\binom{I}{2} = I(I-2)/2$ pairs of rows in combination with each of the $\binom{J}{2} = J(J-2)/2$ pairs of columns. There are $\binom{I}{2}\binom{J}{2}$ odds ratios of this type. In the light of this, it is obvious that the number of odds ratios becomes enormous thus for a major accuracy of the analysis could be fine to choose parsimonious graphical representations of them.

3.12 Hierarchical Logistic Regression Model

Linear Regression Models are not a good shape for estimating probability (Berry and Linoff, 2011). One of the main reasons of this is that these models would be inappropriate to predict a binary response model because a linear function is unlimited. In fact, these models could predict values for the response variable outside the interval [0;1], which would be meaningless (Giudici, 2003).

Focusing the attention on the logistic regression modeling decision makers can underline that this technique is very appealing because it provides a closed-form solution for the posterior probabilities and it is easy to use and gives quick and robust results. Logistic regression, which is a widely used statistical modeling technique, could build a model with dichotomous outcome and has been proven as a powerful algorithm (Lee et al., 2006). This model has been well studied and used in a lot of applications, specially, in management and marketing area in order to estimate the relationship among variables.

Let $y_i (i = 1,2,...,n)$ be the observed value of a binary response variable, which can get only two values: 0 or 1. Normally, the value 1 refers the 'success' while, the value 0 concerns the 'no success'. In our case, if the target variable is equal to 1 means that the potential customer purchased the service on line. According to Giudici (2003) a logistic regression model is defined in terms of fitted values and is interpreted as probabilities that the event occurs in different subpopulations $\pi_i = P(Y_i = 1),$ for $i = (1,2,...,n)$.

The logistic regression model underlines that an appropriate function of the fitted probability of the event is a linear function of the observed values of the available independent variables. In fact, $\log\left[\frac{\pi_i}{1-\pi_i}\right] = a + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik}$. In other words, the left-hand side defines the logit function of the fitted probability, logit (π_i) as the logarithm of the odds for the event, namely the natural logarithm of the ratio between the probability of occurrence (success) and the probability of non-occurrence (failure):

$$\text{logit}\,(\pi_i)\log = \left[\frac{\pi_i}{1-\pi_i}\right]$$

Since π_i is calculated, on the basis of the data, a fitted value for each binary observations $\hat{y}_i$ can be obtained, introducing a threshold value of π_i above which $\hat{y}_i = 1$ and below which $\hat{y}_i = 0$.

By inverting the definition of the logit function it is possible obtain

$$\pi_i = \frac{\exp(a + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik})}{1 + \exp(a + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik})}$$

This relationship refers to the logistic regression function. It is very helpful in management and marketing area for identifying the main drivers or activities that could maximize the probability of customer conversion in order to minimize the level of churners. Particular attention must be paid on the coefficient β of the logistic function.

This parameter determines the rate of growth or increase of the function and the sign of β indicate whether the function increases or decreases and the magnitude of β determines the rate of that increase or decrease. In particular when $\beta > 0$ then (x) increases as x increases, by contrast, if $\beta < 0$ then $\pi(x)$ decreases as x increases. An important remark is that when $\beta \to 0$ the curve tends to became a horizontal straight line. In other words, if $\beta = 0$, Y and X are independent. It is important remember that even if the probability of success is a logistic function and therefore not linear exploratory variables, the logarithm of the odds is a linear function of the independent variables. In fact, $\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$, where positive log-odds favors $Y = 1$ whereas negative log-odds favors $Y = 0$. The log-odds expression establishes that the logit increases by β units for a unit increase in x. For the logistic regression model, the odds of success can be expressed by the following formula:

$$\frac{\pi(x)}{1-\pi(x)} = e^{\alpha+\beta x} = e^{\alpha}(e^{\beta})^{x}$$

This exponential relationship is an useful interpretation of the parameter β. A unit increase in x multiplies the odds by a factor e^{β}. Indeed, the odds at level $x+1$ is equal to the odds at the level x multiplied by e^{β}. Finally, if $\beta = 0$ it means that $e^{\beta} = 1$ thus the odds do not depend on X.

3.13 Decision Tree: A General Overview

While logistic regression is defined as a popular predictive techniques to link satisfaction with attributes of customer retention (Rust and Zahorik, 1993), decision tree have become an import knowledge structure, used for the classification of future events (Muata and Bryson, 2004). In addition, tree models are considered as non-parametric models, in fact they do not require assumptions about the probability distribution of the dependent variable (Giudici, 2010). Also, decision trees as a hierarchical collection of rules that describe how to divide a large collection of records into successively smaller group of records (Berry and Linoff, 2011). With each successive division, the member of the resulting segment became more and more similar to one another with respect to the dependent variable. In other words, decision trees produce a classification of observations into groups and obtain a score for each group obtained but they are predictive models rather than descriptive (Figini and Giudici, 2009).

Tree models can be defined as a recursive procedure, through which a set of n statistical units are progressively divided into groups, according to a division rule that aims to maximize a homogeneity or purity measure of the response variable in each of the obtained groups. In fact, at each step of the procedure, a division rule is specified by the choice of an explanatory variable to split and choice of a splitting rule for the variable, which establishes how to partition the observation. To achieve a final partition of the observation it is necessary to specify stopping criteria for the division process. In fact, supposing that a final partition has been reached, consisting of g groups $(g < n)$. After, for any given response variable observation y_i, a regression tree produce a fitted value $\hat{y}_i$ that is equal to the mean response value of the group to which the observation i belongs. Let m be such a group we can obtain that $\hat{y}_i = \frac{\sum_{l=1}^{n_m} y_{lm}}{n_m}$. On the other hand, for a classification tree, fitted values are given in terms of fitted probabilities of affiliation to a single group, which suppose only two classes are possible (binary classification); the fitted success probability is given by $\pi_i = \frac{\sum_{i=1}^{n_m} y_{lm}}{n_m}$ where the observations y_{lm} can take the value 0 or 1, and the fitted probability corresponds to the observation proportion of success in group m. An important remark is that both $\hat{y}_i$ and π_i are constant for all the observations in the group.

After the implementation on the decision tree decision makers must start with an accurate preliminary exploratory analysis. First of all, it is fundamental to control that the sample size is sufficiently large. Partitions with fewer observation will have fitted values with an high level of variance. After, it accurate to investigate the dependent variable in order to identify some anomalies in the data that could severely distort the

results of the analysis. Then, special attention to the shape of the dependent variable distribution. In other words, if the distribution is strongly asymmetrical, the procedure may lead to isolated group with few observations from the tail of the distribution. Additionally, when the dependent variable is qualitative, for an optimal accuracy of the model, the number of the levels should not be too large in order to improve the tree stability and the predictive performance.

After this step, it is really important to choose an appropriate model algorithm paying attention to how it performs. According to the current literature, the two main aspects are the division criteria and the methods employed to reduce the dimension of the tree. Indeed, the most popular algorithm in the statistical community is the CART algorithms (Breiman, Friedman, Olshen and Stone, 1984), which stands for 'classification and regression trees'. Other algorithms include CHAID (Kass, 1980), C4.5 and its later version, C5.0 (Quinlan, 1993). More precisely, C4.5 and C5.0 are widely used by computer scientists. Notice, that the first version of C4.5 and C5.0 were limited to categorical predictors, but the most recent versions are similar to CART. As consequence it is clear that in our case the C4.5 and C5.0 are inadequate for the data analysis due to the nature of the dependent variable. In addition, we have implemented a classification tree rather than a regression tree. In the following section division criteria and pruning modality to reduce the complexity of a tree are provided.

3.14 Decision Tree: Division Criteria

Identifying a division criteria means choosing a predictor from those available, and choosing the best partition of its levels. Generally, the choice is made using a goodness measure of the corresponding division rule. From a statistical standpoint, a goodness measure $\Phi(t)$ is a measure of the performance gain in subdividing a (parent) node t according to a segmentation into a number of (child) nodes. Let $t_r, r = 1,...,s,$ denote that child groups generated by the segmentation ($s = 2$ for a binary segmentation) and let p_r denote the proportion of observations, among those in t, that are allocated to each child node, with $\sum p_r = 1$. The criterion function is usually expressed as $\Phi(s,t) = 1(t) - \sum_{r=1}^{s} I(t_r)\, p_r$ where I denotes an impurity function. High values of the criterion function imply that the chosen partition is adequate. The concept of impurity is used to measure the variability of the dependent values of the observations. For instance, in a regression tree, a node will be pure if it has null variance and impure if the variance of the observations is high. In fact, for regression trees impurity is equal to the variance while, for classification trees alternative measures such as the Misclassification Impurity[11], the Gini Impurity[12] and Entropy Impurity[13] should be considered. Compared

[11] The Misclassification Impurity is given by $I_M m = \frac{\sum_{l=1}^{n_m} 1(y_{lm}, y_k)}{n_m} = 1 - \pi_k$ where y_k is the modal category of the node, with fitted probability π_k, and the function $1(\cdot,\cdot)$ denotes the indicator function, which take the value 1 if $y_{lm} = y_k$ and 0 otherwise.

to Misclassification impurity, Gini and Entropy are more sensitive to changes in the fitted probabilities because of the decrease faster than Misclassification rate as the tree grows. Despite this, to obtaining an accurate decision tree decision makers should choose the misclassification impurity especially for the goodness of fit of a classification tree. In addition, an impurity measures can be used to provide an overall assessment of a tree. Indeed, let $N(T)$ be the number of leaves (terminal node) of a tree T. The total impurity of T is given by $I(T) = \sum_{m=1}^{N(T)} I(t_m) p_m$ where p_m are the observed proportions of observations in the final classification. Finally, the impurity measure used by CHAID is the distance between the observed and expected frequencies. More precisely, the expected frequencies are calculated using the hypotheses for homogeneity for the observation in the considered node. This split criterion is the Pearson χ^2 index. If the decrease in χ^2 is significant (the *p*-value is lower than a specific level α) then a node is split; otherwise it remains unsplit and becomes a leaf.

3.15 Decision Tree: Pruning

When no stopping criteria are established, a tree model could grow until each node contains identical observation in terms of values or levels of the dependent variable. This approach is definitely inaccurate and inadequate because of does not constitute a optimal segmentation of the observation. Anyway, it fundamental to stop the growth of the tree at a reasonable dimension. Decision trees must have a small number of leaves because of the predictive rule can be easily implemented from a business standpoint. Also, decision tree must have a large number of leaves that are maximally pure. The final choice is bound to be a compromise between the two opposite strategies.

In the current literature we can find many contributes that explain the features of the many trees algorithms used to resolve business problems. For instance, some tree algorithms use stopping rules based on the number of the leaves, or on the maximum number of steps in the process. Other algorithms are based on probabilistic assumptions on the variable, allowing us to use suitable statistical tests. Notice that in absence of probabilistic assumptions, the growth is stopped when the decrease in impurity is too small. In this research, due to the nature of the data, particular attention is paid on the CART algorithm. In fact, a classification and regression tree (CART) is constructed by recursively splitting the instance space into smaller sub-groups until a specified criterion has been met (Bloemer, Brijs, Vanhoof and Swinnen, 2003; Hadden, Tiwari, Roy and Ruta, 2005). The tree is only allowed to grow until the decrease in impurity falls below

---

[12] The Gini Impurity is $I_G(m) = 1 - \sum_{i=1}^{k(m)} \pi_i^2$ where the π_i are the fitted probabilities of the levels present at node m, which are the most $k(m)$.

[13] The Entropy Impurity is $I_E(m) = -\sum_{i=1}^{k(m)} \pi_i \log \pi_i$ with π_i are the fitted probabilities of the levels present at node m, which are the most $k(m)$.

a user-defined threshold. At this time the node becomes a terminal, or leaf node (Giudici, 2003).

Let T be a tree, and let T_0 denote that tree of greatest size. From any tree a subtree can be obtained by collapsing any number of its internal (non-terminal) nodes. The idea of pruning is to find a subtree of T_0 in a optimal way, so as, to minimize a loss function. The loss function implemented in the CART algorithm depends on the total impurity of the tree T and the tree complexity: $C_\alpha = I(T) + I(T) + \alpha N(T)$ where, for a tree $T, I(T)$ is the total impurity function calculated at the leaves, and $N(T)$ is the number of leaves; whit α a constant that penalizes complexity linearly. On the other hand, in a regression tree the impurity is a variance, thus the total impurity is calculated as $I(T) = \sum_{m=1}^{N(T)} I_{v(m)n_m}$.
From a managerial point of view, the misclassification impurity is usually chosen in a practice. The minimization of the loss function leads to a compromise between choosing a complex model (low impurity but high complexity cost) and choosing a simple model (high impurity but low complexity cost). The choice depends on the chosen value of α. For each α it can be shown that there is a unique subtree of T_0 that minimizes $C_\alpha(T)$. A possible weakness of this loss function is that the performance or each tree configuration is evaluated with the same data used for building the classification rules, which can lead to optimistic estimates of the impurity especially for large trees because of the goodness of the fit to the data increases with the complexity of the number of leaves. Also, decision makers can consider alternatives pruning criteria based on the predictive misclassification errors. In this case, the database analyzed could be divide in two parts and they can use the second part for validating the model measuring the impurity in an unbiased fashion. The loss function is so evaluated by measuring the complexity of the model fitted on the training database, whose misclassification errors are measured on the validation dataset. Finally, the CHAID algorithm used chi-squared testing to produce an implicit stopping criterion based on testing the significance of the homogeneity hypothesis; the hypothesis is rejected for a large value of χ^2. If homogeneity is rejected for a determinate node, then splitting continues, otherwise the node becomes terminal.

3.16 Criteria based on the Loss Functions

Businesses need to evaluate the results obtained from the models previously implemented not only by comparing them among themselves but also by comparing the advantages to be had by using one model rather than another. The main objective for marketers and decision makers is to reduce uncertainties in the risks factors or loss factors within companies. In fact, in order to evaluate the accuracy of the logistic regression model decision makers can develop two criteria: the percentage correctly classified (accuracy) at the 'economically optimal' cutoff purchase probability (PCC) and the receiver operating characteristic (ROC) curve. Both criteria are predictive in nature, as oppose to more traditional evaluation criteria such as the hypothesis-testing approach, or the resubstitution rate. Notice that the hypothesis-testing approach investigate the statistical significance of parameters estimate, while the resubstitution rate or PCC based on the estimation sample.

3.16.1 Confusion Matrix

The classification method is used to identify the potential customer according to their purchase behavior. This measure is provided by the logit modeling techniques as the 'posteriori' probability. The result of a classification can be summarized in the following classification table called confusion matrix (Morrisson, 1969). It contains the number of elements that have been correctly or incorrectly classified for each class. The main diagonal shows the number of observations that have been correctly classified for each class; the off-diagonal elements indicate the number of observations that have been incorrectly classified. The proportion of incorrect classification over the total number of classifications is called misclassification error. Companies to achieve a competitive advantage in the market have to minimize this quantity.

Table 3.3 Theoretical Confusion Matrix

		Predicted status	
		Customer	Non-Customer
True Status	Customer	True Positive (TP)	False Negative (FN)
	Non-Customer	False Positive (FP)	True Negative (TN)

Personal adaption from Dirk Van den Poel, 2004

According to Bradley (1997) from the previously table we can extract these information:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Following these definitions, it is clear that sensitivity represents the proportion of event observations that the model predicts as events (number of true positives divided by the number of events) and specificity is defined as the proportion of non-event observations that the model predicts to be non-events (number of true negatives divided by the total number of non-events). A disadvantage of this measure is that it is not very robust concerning the chosen cut off value in the 'a posteriori' probabilities (Baesens et al., 2002; Giudici 2003; Van den Poel, 2003). These criteria could be useful for global organizations because of it indicate approximately the predictive accuracy of the model

previously developed. Finally, for a greater simplicity an example of confusion matrix is following proposed.

Table 3.4 Confusion Matrix

	ESTIMATED	ESTIMATED	
ACTUAL	0	1	TOTAL
0	**49**	6	55
1	25	**20**	45
TOTAL	74	26	**100**

Table 3.4 notes that the model has classified correctly 20 observations about the category of good individuals out of 26 and 49 observations related to the category of bad individuals. In addition, it shows that the model accuracy rate is equal to 69 per cent. In fact, (49+20)/100 = 69.

Despite the good accuracy of the predictive model, this criterion is not very robust concerning the chosen cut off value in the 'a posteriori' probabilities. For this reason, a more robust model is described below.

3.16.2 ROC Curve

Invariance of the performance criterion with respect to the selected cutoff value can be achieved by considering the curve which plots sensitivity (vertical axis) versus one minus specificity (horizontal axis) for all possible cutoff values (Van den Poel, 2003). The former is also called hit percentage instead, the latter is also called the 'false-alarm probability' (Green and Swets, 1966). This curve is named a Receiver Operating Characteristics (ROC) curve. We refer to Green and Swets (1966); Swets (1979;1988); Figini and Giudici (2009) and Giudici (2010) for more details. The authors have shows that the predictive accuracy of a classification procedure of the logit modeling can be measured by the Area Under the ROC curve (AUC).

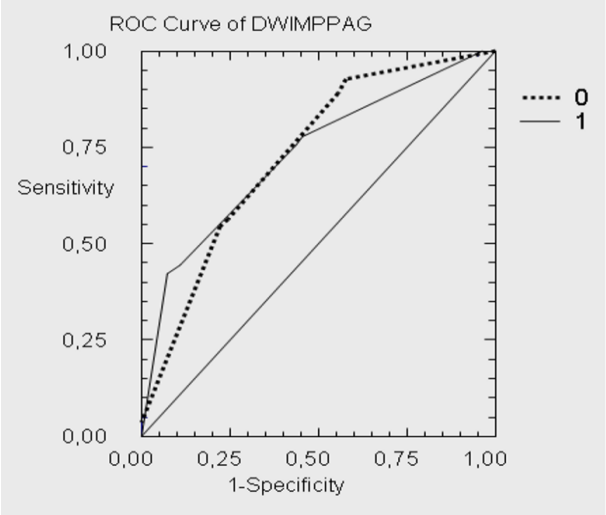
Hanley and McNeil (1982) provide an intuitive interpretation of the AUC, which is based on the equivalence of the AUC to the Mann-Whitney or Wilcoxon statistics. More precisely, they show that the AUC represents the probability that a randomly chosen positive example (rated as a buyer) is correctly rated (or ranked) higher than a randomly selected negative example (rated as a buyer). This again illustrates that this performance measure is not dependent on the choice of a cutoff value. The AUC statistic ranges from a lower limit of 0,5 for change (null model) performance to an upper limit of 1,0 for perfect performance (Green and Swets, 1966). Instead, in terms of comparison between models the best model is that in which the curve is more shifted to the left. The ideal curve coincides with the vertical axis (Giudici, 2010).

In order to interpret the AUC value, decision makers can consider the classification of discriminative capacity proposed by Swets (1988). This classification is based on subjective criteria and can be represented as follow:

- AUC = 0.5 Not informative test
- $0.5 < AUC \leq 0.7$ Inaccurate test
- $0.7 < AUC \leq 0.9$ Moderate test
- $0.9 < AUC \leq 1$ Highly accurate test
- AUC = 1 Perfect test

Since neither of the two criteria (AUC and PPC) the first is clearly superior, we will use both performance measures, and assess they convergent validity. One disadvantage is shared by both performance criteria, namely the assumption of equal opportunity costs of misclassification. Both PPC and AUC weigh the opportunity cost of misclassifying a buyer as a non-buyer and the cost of misclassifying a non-buyer as a buyer equally. It is easier to incorporate the issue of unequal misclassification costs into the PCC criterion than into AUC. For instance, the probability of a misclassification is multiplied by the cost of misclassification (for both buyers and non-buyers). Both performance criteria will be calculated on a test or holdout sample, which only consists of observations not used during model estimation, and which is half the size of the total sample (Van den Poel, 2003). Finally, the next graph shows an example of ROC curve and its interpretation.

Graph. 3.5 ROC curve



According to Swets (1988) the model is accurately moderate because the AUC is equal to 73.03%. Indeed, in term of models comparison the predictive model forecasts with major accuracy the observations belonging the group 0 because of the curve is more shifted to the y axis.

Chapter 4
Data Exploration and Aggregation

4.1 Introduction

Dramatic advances in data generation and collection are creating huge and massive databases in many scientific disciplines. The field of data mining grew out the limitation of the traditional statistical data analysis in handling the challenges posed by these new types of datasets. Traditional data analysis tools and techniques cannot be used to analyze these enormous amount of data. Indeed, the non-traditional nature of the data means that traditional approaches and methods cannot be applied even if the dataset is relatively small especially in today's global and competitive landscape. In fact, the challenges of analyzing new type of data cannot be met by simply applying data analysis techniques in isolation from those who understand the data and the domain in which it resides (Tan, Steinbach and Kumar, 2011). More precisely, point-of sale data collection such as credit card number, ratio frequency identification, and Web logs from e-commerce Web sites have allowed marketers to collect up-to-the-minute data about potential customer purchases at the online checkout counters. Marketers, can utilize this information, along business-critical data such as textual data from social networks or magazine, to help them better understand the desire of their potential customers and make more informed business and marketing decisions.

This chapter is structures as follows. It begins explaining the changeover from cookies to potential customer in order to draw an aggregate database. A special focus is paid on the irrelevant variables present in the dataset analyzed. After, it shows the results of both, Pearson Correlation Analysis (PCA) and Multicollinearity Analysis emphasizing the strategic importance of the qualitative approach for marketing decision. Finally, it underlines the linking research question to research strategy.

4.2 From cookies to potential customers

The original database contained more than 1,463,199 potential customers and 42 variables related to their purchase behavior. Initially, the dataset was not aggregated by potential customer. In fact, given the huge dimension of the database, in order to develop an accurate data mining analysis, it is fundamental to aggregate the dataset by 'user id' (unique code for each potential customer). This means that at each 'user id' will correspond a different potential customer. Before aggregating the database, a preliminary qualitative exploratory analysis has been conducted so as to eliminate some insignificant variables. In this case, a qualitative approach such as expert judgment and opinion mining overcomes quantitative approaches. Indeed, data miners must converge both quantitative and qualitative research strategies in the same direction. An important remark is that the data reduction and the establishment of the aggregation criteria have been established according with the Executive Vice President (EVP) Data Platforms of the company. The following tables show these first steps. Special attention must be paid on the Table 4.3 because of it underlines the variables aggregation criteria established.

Table 4.1 Variables description and measures

Variables	Description	Variables Measures
Activity Timestamp	Timestamp for the activity on the advertiser's website	Nominal
Activity Tag name	Activity tag name associated to the action that the potential customer performed on the client (company) site	Nominal
Advertisement Name	Advertisement name associated with the exposure	Nominal
Type of Banner	Type of banner proposed to the potential customer	Nominal
Name of the Advertiser	Name of the advertiser	Nominal
Amount of Model Conversion	Amount of conversion attributed to a specific potential customer in a journey based on the results of the model	Scale
Activity Quantity	Quantity associated with the activity. Quantity can typically be 1 representing the activity but in some cases this will be > 1	Scale
Revenue Activity	Revenue associated with the activity	Scale
Cost per Click	Cost per Click. It is an internet advertising model used to direct traffic to websites, where advertisers pay the publisher when the advertisement is clicked	Scale
Click Through Rate	Click Through Rate. It is a way of measuring the success of an online campaign for a particular product or service. The click through rate advertisement is defined as the number of clicks on an advertisements divided by the number of times the advertisement is shown	Scale
Average Position	The average position of a search term in the search engine	Scale
Brand Search	The search term includes any of the branded terms	Scale
Name of the Campaign	Name of the campaign	Nominal
Creative Height	Creative Height	Scale
Creative Type	Creative Type	Scale
Creative Width	Creative Width	Scale
Creative Name	Creative for display advertisement	Scale
Head Flag	Search flag used to mark high volume keywords	Nominal
Timestamp for impression and clicks	Timestamp for impressions and clicks	Nominal
Impression or Click	'Imp' if the event is an impression, 'Click' if the event is a click.	Nominal
Keywords Advertising Group	Group of Advertising Keyword.	Nominal
Keywords Campaign	Keywords Campaign	Nominal
Keywords Category	Keywords Category	Nominal
Keywords Name	Keywords on which search advertisements appear in the web	Nominal
Match Type	Type of key word in form users	Nominal
Max Search Click	If 1 then includes search if 0 then no search.	Scale
Price Paid	Price paid by the consumer for the purchase	Scale
Quantity Sold	Number of items sold	Scale
Purchases	Number of items purchased by potential customers	Scale
Min Search Click	If 1 is only search, if 0 then includes display	Scale
Single Conversion Activity	Flags journeys where the potential customer only has a single conversion/activity	Scale
Site Placement	Placement on the Site (homepage)	Scale
Rank0	Variable no identified from the company	Scale
Rank1	Auto incrementing value representing all touch points of a potential customer journey. Does not reset on a conversion/activity	Scale

Rank2	Auto incrementing value representing all touch points of a potential customer journey. Does reset on a conversion/activity	Scale
Rank3	Auto incrementing value representing the conversion number for the user. The same value will repeat across all exposures leading to an activity and then will increment and repeat for the next set of exposures leading to a conversion/activity	Scale
Record Number	Record number	Scale
Search Engine Name	Potential customers search on the search engine information related to the marketing campaign	Nominal
Search Click	Represents an exposure that is a search click	Nominal
Segment	Client specific field for this report. For instance shows a segment applied to users based on location/energy consumption.	Nominal
Site Name	Site (for display advertisement).	Nominal
User Id	ID of the potential customers.	Scale

Table 4.2 No-significant statistical variables

Variables	Description	Variables Measures	Notes
Activity Tag name	Activity tag name associated to the action that the potential customer performed on the client (company) site	Nominal	Unclear Variable
Amount of Model Conversion	Amount of conversion attributed to a specific potential customer in a journey based on the results of the model	Scale	Unclear Variable
Activity Quantity	Quantity associated with the activity Quantity can typically be 1 representing the activity but in some cases this will be > 1	Scale	Redundant Variable
Revenue Activity	Revenue associated with the activity	Scale	Redundant Variable
Creative Name	Creative for display advertisement, belong to an advertiser	Scale	Low Predictive Value
Price Paid	Price paid by the consumer for the purchase	Scale	Many Missing Data
Single Conversion Activity	Flags journeys where the potential customer only has a single conversion/activity.	Scale	Redundant
Site Placement	Placement on the Site (for instance: homepage)	Scale	Low Predictive Value
Rank 0	Variable no identified from the company	Scale	Unclear Variables
Record Number	Record number	Scale	No Predictive Variable

4.3 Aggregation Criteria

The following table provides the variables aggregation criteria established in according to the EVP of the company that provided the dataset. This step is really important because of shed light the attention on the strategic importance of both quantitative and qualitative methods in marketing performance monitoring. The quantitative results (Pearson Correlation) have been joined to the decision makers opinion mining and human judgment.

Table 4.3 Variables Aggregation Criteria

Variables	Aggregation Criteria	Variable Measures	New Variable Measures
Activity Timestamp	Categorization into groups and creation of dummy variables	Nominal	Scale
Advertisement Name	Categorization into groups and sum	Nominal	Scale
Type of Banner	Categorization into groups and sum	Nominal	Scale
Name of the Advertiser	Categorization into groups and sum	Nominal	Scale
Cost per Click	Average Value	Scale	Scale
Click Through Rate	Average Value	Scale	Scale
Average Position	Average Value	Scale	Scale
Brand Search	Categorization into groups and sum	Scale	Scale
Name of the Campaign	Categorization into groups and sum	Nominal	Scale
Creative Height	Categorization into groups	Scale	Scale
Creative Type	Categorization into groups	Scale	Scale
Creative Width	Categorization into groups	Scale	Scale
Head Flag	Categorization into groups and sum	Nominal	Scale
Timestamp for impression and click	Categorization into groups and creation of dummy variables	Nominal	Scale
Impression or Click	Categorization into groups and sum	Nominal	Scale
Keywords Advertising Group	Categorization into groups and sum	Nominal	Scale
Keywords Campaign	Categorization into groups and sum	Nominal	Scale
Keywords Category	Categorization into groups and sum	Nominal	Scale
Keywords Name	Categorization into groups and sum	Nominal	Scale
Match Type	Categorization into groups and sum	Scale	Scale
Max Search Click	Sum	Scale	Scale
Quantity Sold	Sum	Scale	Scale
Purchases	Sum	Scale	Scale
Min Search Click	Categorization into groups and sum	Scale	Scale
Rank 1	Maximum Value	Scale	Scale
Rank 2	Maximum Value	Scale	Scale
Rank 3	Maximum Value	Scale	Scale
Search Engine Name	Categorization into groups and sum	Nominal	Scale
Search Click	Categorization into groups and sum	Nominal	Scale
Segment	Categorization into groups and sum	Nominal	Scale
Site Name	Categorization into groups and sum	Nominal	Scale
User Id	Parameter of aggregation	Scale	Scale

4.4 Pearson Correlation Analysis

After the aggregation by ‘user id’, the number of potential customers decreased at 1,463,199. Instead, the number of variables increased at 276 because of the creation of

dummy variables[14]. Due to the huge increase number of potential customers, it could be inaccurate to investigate each variable with graphs and diagrams. In fact, first of all, decision makers must estimate the force of the relationship between the target variable (Purchases: OD) and each independent variable through the Correlation Analysis in order to obtain an accurate database for the analysis. In this case, it has been chosen the Person Correlation due to the nature of the data collected in the database. The data analyzed are quantitative continuous and articulate in interval or ratio scales. In addition, the target variable is dichotomous thus it assumes only two values: 0 (bad potential customer) and 1 (good potential customer). First, the following tables show some general descriptive statistics such as minimum and maximum value, arithmetic mean, standard deviation and coefficient of variation. Second, the Person Correlation Analysis (PCA) among the target variables and each independent variables are provided in the next tables. Finally, an analysis based on the mean correlation variables conclude this exploratory investigation. In other words, the following tables provide an estimation of the variables that could be more related to the purchase of the service offered by the company.

Table 4.4 PCA: Purchases and Advertisement Name associated with the exposure.

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	,02	,13	6,5	1,00	-	1,463,199
Advertisement Name (Brand Top)	[0;0]	-	-	-	.[a]		1,463,199
Advertisement Name (Brand Energy)	[0;23.157]	4,14	40,66	9,82	0,07		1,463,199
a. Cannot be computed because at least one of the variables is constant							

Table 4.5 PCA: Purchases and Average Position Best Five - Brand Search

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	,02	,13	6,5	1,00	**,39**	1,463,199
Average Position Best Five	[0;23]	,01	,15	15	**,38****		1,463,199
Brand Search	[0;22]	,01	,14	14	**,39****		1,463,199
**Correlation is significant at the 0.01 level							

Table 4.6 PCA: Purchases and Creative Height

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	0,02	0,13	6,5	1,00	**0,16**	1,463,199
Height_0	[0;30]	0,04	0,37	9,25	**,69****		1,463,199
Height_1	[0;273]	0,17	1,5	8,82	,02**		1,463,199
Height_60	[0;5.371]	0,41	8,19	19,98	,03**		1,463,199
Height_90	[0;11.919]	1,07	17,63	16,48	,05**		1,463,199
Height_250	[0;5.903]	0,88	10,9	12,39	,05**		1,463,199
Height_600	[0;4.446]	1,58	12,21	7,73	,09**		1,463,199
** Correlation is significant at the 0.01 level							

[14] Dummy variables can be defined as variables which, when it occurs in an expression, can be replaced with another variable without changing the meaning of the statement (Keisler and Robbin, 1996).

Table 4.7 PCA: Purchases and Creative Type

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	0,02	0,13	6,50	1,00	**0,04**	1,463,199
AOLBraOptV 1Eco2020	[0;21]	0	0,07	0,00	,01**		1,463,199
AOLBraOptV 2Eco2020	[0;26]	0	0,07	0,00	,01**		1,463,199
AOL Coin 160x600	[0;14]	0	0,05	0,00	,00**		1,463,199
AOL Coin 300x250	[0;12]	0	0,06	0,00	,00**		1,463,199
AOL Coin 468x60	[0;6]	0	0,03	0,00	,00**		1,463,199
AOL Coin 728x90	[0;31]	0	0,08	0,00	,01**		1,463,199
AOL Kite 728x90	[0;22]	0	0,07	0,00	,01**		1,463,199
AOL Neon 160x600	[0;16]	0	0,08	0,00	,00**		1,463,199
AOL Neon 300x250	[0;26]	0	0,06	0,00	,00**		1,463,199
AOL Neon 468x60	[0;17]	0	0,04	0,00	,00**		1,463,199
AOL Neon 728x90	[0;12]	0	0,07	0,00	,01**		1,463,199
DrvPMBraOpt V1Eco2	[0;478]	0,01	0,6	60,00	,00**		1,463,199
DrvPMBraOpt V2Eco2	[0;435]	0,01	0,58	58,00	,00**		1,463,199
DrvPMCoin 120x600	[0;46]	0	0,09	0,00	,02**		1,463,199
DrvPMCoin 300x250	[0;831]	0,02	1,12	56,00	,00**		1,463,199
DrvPMCoin 468x60	[0;85]	0	0,19	0,00	,00**		1,463,199
DrvPMCoin 728x90	[0;422]	0,01	0,57	0,00	,00**		1,463,199
DrvPMKite 120x600	[0;39]	0	0,09	0,00	,02**		1,463,199
DrvPMKite 234x60	[0;143]	0,03	0,33	11,00	,00**		1,463,199
DrvPMKite 728x90	[0;444]	0,01	0,6	60,00	,00**		1,463,199
DrvPMNeon 120x600	[0;33]	0	0,08	0,00	,00**		1,463,199
DrvPMNeon 300x250	[0;740]	0,02	1,1	55,00	,05**		1,463,199
DrvPMNeon 468x60	[0;84]	0	0,2	0,00	,00**		1,463,199
DrvPMNeon 728x90	[0;474]	0,01	0,59	59,00	,00**		1,463,199
EbayBraOptV 1Eco202	[0;517]	0,04	0,69	17,25	,06**		1,463,199
EbayBraOptV 2Eco202	[0;556]	0,04	0,71	17,75	,05**		1,463,199
EbayCoin 160x600	[0;769]	0,26	2,21	8,50	,02**		1,463,199
EbayCoin 300x250	[0;131]	0,01	0,25	25,00	,01**		1,463,199
EbayCoin 728x90	[0;48]	0	0,09	0,00	,00**		1,463,199
EbayKite 728x90	[0;61]	0	0,1	0,00	,00**		1,463,199
EbayNeon 160x600	[0;945]	0,05	2,13	42,60	,02**		1,463,199

EbayNeon 300x250	[0;970]	0,07	1,07	15,29	,01**		1,463,199
EbayNeon 728x90	[0;51]	0	0,1	0,00	,00**		1,463,199
Gif	[0;1188]	0,02	1,33	66,50	,02**		1,463,199
Long160x600	[0;118]	0,02	0,26	13,00	,00**		1,463,199
Long300x250	[0;16]	0	0,07	0,00	,00**		1,463,199
Long468x60	[0;4]	0	0,02	0,00	,00**		1,463,199
Long728x90	[0;11]	0	0,03	0,00	,00**		1,463,199
Magnify 160x600	[0;1059]	0,13	2,04	15,69	,06**		1,463,199
Magnify 300x250	[0;1515]	0,12	1,96	16,33	,06**		1,463,199
Magnify 468x60	[0;773]	0,03	1,11	37,00	,03**		1,463,199
Magnify 728x90	[0;2127]	0,14	2,69	19,21	,07**		1,463,199
Product 120x600	[0;194]	0,06	0,53	8,83	,06**		1,463,199
Product 728x90	[0;119]	0,06	0,6	10,00	,07**		1,463,199
Product 300x250	[0;648]	0,07	1,12	16,00	,03**		1,463,199
Product 468x60	[0;341]	0,02	0,49	24,50	,03**		1,463,199
Product 728x90	[0;825]	0,05	0,98	19,60	,06**		1,463,199
** Correlation is significant at the 0.01 level							

Table 4.8 PCA: Purchases and Creative Width

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	0,02	0,13	6,50	1,00	**0,14**	1,463,199
Width_0	[0;30]	0,04	0,37	9,25	**0,69****		1,463,199
Width_1	[0;273]	0,17	1,5	8,82	0,02**		1,463,199
Width_120	[0;3036]	0,58	5,72	9,86	0,07**		1,463,199
Width_160	[0;2821]	1	8,74	8,74	0,08**		1,463,199
Width_234	[0;676]	0,05	1,07	21,40	0,11**		1,463,199
Width_300	[0;5903]	0,88	10,9	12,39	0,05**		1,463,199
Width_468	[0;5046]	0,36	7,66	21,28	0,03**		1,463,199
Width_728	[0;11919]	1,07	17,63	16,48	0,05**		1,463,199
** Correlation is significant at the 0.01 level							

Table 4.9 PCA: Purchases and Head Flag

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	,02	,13	6,50	1,00	**,34**	1,463,199
Headfl_0	[0;22]	,00	,10	0,00	**,36****		1,463,199
Headfl_1	[0;14]	,01	,09	9,00	**,32****		1,463,199
**Correlation is significant at the 0.01 level							

Table 4.10 PCA: Purchases and Impression or Click

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	,02	,13	6,50	1,00	**,07**	1,463,199
Impression or Click 2010	[0;16300]	2,35	25,51	10,86	0,06**		1,463,199
Impression or Click 2011	[0;8028]	1,80	19,98	11,1	0,08**		1,463,199
** Correlation is significant at the 0.01 level							

Table 4.11 PCA: Purchases and Hour of Impression or Click

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	0,02	0,13	6,50	1,00	**0.05**	1,463,199
Impression or Click at 12:00 pm	[0;374]	0,02	0,78	39,00	,01**		1,463,199
Impression or Click at 1:00 am	[0;442]	0,03	0,87	29,00	,01**		1,463,199
Impression or Click at 2:00 am	[0;1074]	0,07	1,54	22,00	,02**		1,463,199
Impression or Click at 3:00 am	[0;1338]	0,12	2,31	19,25	,03**		1,463,199
Impression or Click at 4:00 am	[0;1694]	0,18	2,68	14,89	,05**		1,463,199
Impression or Click at 5:00 am	[0;1759]	0,21	2,93	13,95	,06**		1,463,199
Impression or Click at 6:00 am	[0;1835]	0,22	2,96	13,45	,06**		1,463,199
Impression or Click at 7:00 am	[0;3418]	0,23	4,14	18,00	,04**		1,463,199
Impression or Click at 8:00 am	[0;3945]	0,24	4,74	19,75	,04**		1,463,199
Impression or Click at 9:00 am	[0;2451]	0,23	3,49	15,17	,05**		1,463,199
Impression or Click at 10:00 am	[0;2201]	0,24	3,21	13,38	,06**		1,463,199
Impression or Click at 11:00 am	[0;2374]	0,27	3,45	12,78	,06**		1,463,199
Impression or Click at 12:00 am	[0;1208]	0,27	2,43	9,00	,08**		1,463,199
Impression or Click at 1:00 pm	[0;1189]	0,28	2,2	7,86	,01**		1,463,199
Impression or Click at 2:00 pm	[0;1071]	0,3	2,26	7,53	,01**		1,463,199
Impression or Click at 3:00 pm	[0;1161]	0,31	2,32	7,48	,10**		1,463,199
Impression or Click at 4:00 pm	[0;1215]	0,3	2,3	7,67	,10**		1,463,199
Impression or Click at 5:00 pm	[0;1053]	0,25	2,17	8,68	,09**		1,463,199
Impression or Click at 6:00 pm	[0;975]	0,16	1,71	10,69	,07**		1,463,199
Impression or Click at 7:00 pm	[0;667]	0,1	1,45	14,50	,04**		1,463,199
Impression or Click at 8:00 pm	[0;907]	0,05	1,29	25,80	,02**		1,463,199
Impression or Click at 9:00 pm	[0;383]	0,03	0,92	30,67	,01**		1,463,199
Impression or Click at 10:00 pm	[0;414]	0,02	0,85	42,50	,01**		1,463,199
Impression or Click at 11:00 pm	[0;381]	0,02	0,75	37,50	,09**		1,463,199
**Correlation is significant at the 0.01 level							

Table 4.12 PCA: Purchases and Day of Impression or Click

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	0,02	0,13	6,50	1,00	**0,06**	1,463,199
Impression or Click Day: 1	[0;567]	0,08	1,11	13,88	,06**		1,463,199
Impression or Click Day: 2	[0;1205]	0,15	1,97	13,13	,06**		1,463,199
Impression or Click Day: 3	[0;981]	0,15	1,94	12,93	,07**		1,463,199
Impression or Click Day: 4	[0;811]	0,16	1,76	11,00	,07**		1,463,199
Impression or Click Day: 5	[0;804]	0,16	1,83	11,44	,06**		1,463,199
Impression or Click Day: 6	[0;798]	0,16	1,97	12,31	,07**		1,463,199
Impression or Click Day: 7	[0;948]	0,16	2,2	13,75	,07**		1,463,199
Impression or Click Day: 8	[0;869]	0,12	1,92	16,00	,06**		1,463,199
Impression or Click Day: 9	[0;538]	0,11	1,51	13,73	,06**		1,463,199
Impression or Click Day: 10	[0;774]	0,14	1,72	12,29	,06**		1,463,199
Impression or Click Day: 11	[0;959]	0,15	2,07	13,80	,06**		1,463,199
Impression or Click Day: 12	[0;796]	0,15	1,71	11,40	,07**		1,463,199
Impression or Click Day: 13	[0;738]	0,3	2,28	7,60	,06**		1,463,199
Impression or Click Day: 14	[0;603]	0,12	1,5	12,50	,06**		1,463,199
Impression or Click Day: 15	[0;646]	0,1	1,51	15,10	,06**		1,463,199
Impression or Click Day: 16	[0;762]	0,11	1,54	14,00	,06**		1,463,199
Impression or Click Day: 17	[0;660]	0,13	1,76	13,54	,06**		1,463,199
Impression or Click Day: 18	[0;745]	0,12	1,66	13,83	,05**		1,463,199
Impression or Click Day: 19	[0;540]	0,12	1,38	11,50	,06**		1,463,199
Impression or Click Day: 20	[0;541]	0,11	1,34	12,18	,06**		1,463,199
Impression or Click Day: 21	[0;493]	0,1	1,27	12,70	,07**		1,463,199
Impression or Click Day: 22	[0;592]	0,1	1,43	14,30	,06**		1,463,199
Impression or Click Day: 23	[0;1083]	0,12	1,94	16,17	,05**		1,463,199
Impression or Click Day: 24	[0;714]	0,1	1,67	16,70	,05**		1,463,199
Impression or Click Day: 25	[0;1125]	0,1	1,9	19,00	,05**		1,463,199
Impression or Click Day: 26	[0;1306]	0,15	2,27	15,13	,04**		1,463,199
Impression or Click Day: 27	[0;2809]	0,13	3,05	23,46	,03**		1,463,199
Impression or Click Day: 28	[0;755]	0,13	1,61	12,38	,06**		1,463,199
Impression or Click Day: 29	[0;1570]	0,15	2,51	16,73	,04**		1,463,199
Impression or Click Day: 30	[0;4324]	0,16	4,66	29,13	,02**		1,463,199
Impression or Click Day: 31	[0;1029]	0,09	1,69	18,78	,01**		1,463,199
** Correlation is significant at the 0.01 level							

Table 4.13 PCA: Purchases and Keywords of Advertising Groups

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	0,02	0,13	6,50	1,00	**0,04**	1,463,199
Keywords Group: Bills	[0];2]	0	0,01	0,00	,00*		1,463,199
Keywords Group: Brand	[0;22]	0,01	0,14	14,00	**,39****		1,463,199
Keywords Group: Brand Energy (1)	[0;3]	0	0,01	0,00	,07**		1,463,199
Keywords Group: Brand Energy (2)	[0;2]	0	0,01	0,00	0		1,463,199
Keywords Group: Dual Fuel	[0;2]	0	0,01	0,00	,02**		1,463,199
Keywords Group: Eco 2020	[0;4]	0	0,01	0,00	,01**		1,463,199
Keywords Group: Eco Action Team	[0;1]	0	0,01	0,00	0		1,463,199
Keywords Group: Eco Manager	[0;2]	0	0,01	0,00	,01**		1,463,199
Keywords Group: Electricity	[0;4]	0	0,01	0,00	,01**		1,463,199
Keywords Group: Energy	[0;17]	0	0,02	0,00	,02**		1,463,199
Keywords Group: Fixed Price	[0;3]	0	0,01	0,00	,09**		1,463,199
Keywords Group: Fuel	[0;3]	0	0,01	0,00	0		1,463,199
Keywords Group: Gas	[0;3]	0	0,01	0,00	,01**		1,463,199
Keywords Group: Gas Electricity (1)	[0;3]	0	0,01	0,00	,07**		1,463,199
Keywords Group: Gas Electricity (2)	[0;2]	0	0,01	0,00	0		1,463,199
Keywords Group: Gas Electricity (3)	[0;2]	0	0,01	0,00	0		1,463,199
Keywords Group: Online Tarr	[0;2]	0	0,01	0,00	,05**		1,463,199
Keywords Group: Site Link	[0;4]	0	0,01	0,00	,04**		1,463,199
Keywords Group: Site Linkto	[0;8]	0	0,03	0,00	,17**		1,463,199
Keywords Group: Site Link Top Performer	[0;3]	0	0,02	0,00	0		1,463,199
Keywords Group: Stst	[0;5]	0	0,02	0,00	,08**		1,463,199
Keywords Group: Team Atob	[0;2]	0	0	0,00	0		1,463,199
Keywords Group: Top Performance (1)	[0;4]	0	0,02	0,00	,11**		1,463,199
Keywords Group: Top Performance (2)	[0;3]	0	0,02	0,00	(-),00*		1,463,199
Keywords Group: Utilities	[0;11]	0	0,03	0,00	(-),00*		1,463,199
*Correlation is significant at the 0.05 level **Correlation is significant at the 0.01 level							

Table 4.14 PCA: Purchases and Keywords Campaign

	Min-Max Value	Mean	Standard Deviation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	,02	,13	1,00	**,10**	1,463,199
Keywords Campaign: Brand	[0;9]	,00	,05	,18**		1,463,199

Keywords Campaign: Generic	[0;6]	,00	,02	,02**		1,463,199
Keywords Campaign: New Content Net	[0;11]	,00	,03	(-),00**		1,463,199
Keywords Campaign: Product	[0;4]	,00	,02	,04**		1,463,199
Keywords Campaign: T Green Britain	[0;2]	,00	,01	,00		1,463,199
Keywords Campaign: Top Performers	[0;23]	,01	,14	**,39****		1,463,199
** Correlation is significant at the 0.01 level						

Table 4.15 PCA: Purchases and Match Type

	Min-Max Value	Mean	Std. Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	,02	,13	7,52	1,00	**,18**	1,463,199
Match Type: Advanced	[0;11]	,00	,03	46,35	,09**		1463199
Match Type: Broad	[0;17]	,00	,06	24,76	**,16****		1463199
Match Type: Exact	[0;21]	,01	,13	15,12	**,38****		1463199
Match Type: Phrase	[0;6]	,00	,02	54,18	,09**		1463199
** Correlation is significant at the 0.01 level							

Table 4.16 PCA: Purchases and Max-Min Search Click

	Min-Max Value	Mean	Std. Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	0,02	0,13	6,50	1,00	**0,12**	1,463,199
Max Search Click	[0;22370]	0,24	26,25	109,38	,05**		1,463,199
Min Search Click	[0;11]	0	0,09	0,00	,20**		1,463,199
** Correlation is significant at the 0.01 level							

Table 4.17 PCA: Purchases and Quantity Sold

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	,02	,13	6,50	1,00	**,97**	1,463,199
Quantity Sold	[0;19]	,02	,14	7	,97**		1,463,199
** Correlation is significant at the 0.01 level							

Table 4.18 PCA: Purchases and Name of the Campaign

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	0,02	0,13	6,50	1,00	**0,17**	1,463,199
Name of the Campaign: BrandQ109STST	[0;218]	0,01	0,39	39,00	,00**		1,463,199
Name of the Campaign: Brand_Quidco_August_09	[0;30]	0,02	0,27	13,50	**,49****		1,463,199
Name of the Campaign: Dart Search	[0;26]	0,01	0,16	16,00	**,40****		1,463,199
Name of the Campaign: Merlin London Eye	[0;132]	0,01	0,42	42,00	,01**		1,463,199
Name of the Campaign: Other	[0;7]	0	0,01	0,00	0		1,463,199

Name of the Campaign: Winter Price	[0;273]	0,16	1,43	8,94	,02**		1,463,199
Name of the Campaign: Brand Affiliate	[0;18]	0,01	0,14	14,00	**,39****		1,463,199
Name of the Campaign: Brand Energy	[0;23046]	3,93	40,3	10,25	,07**		1,463,199
** Correlation is significant at the 0.01 level							

Table 4.19 PCA: Purchases and the Day of the Purchase

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases**:** OD	[0;19]	0,02	0,13	6,5	1,00	0,17	1,463,199
Purchase Day: 1	[0;3]	0,00	0,04	0,00	,27**		1,463,199
Purchase Day: 2	[0;3]	0,00	0,03	0,00	,23**		1,463,199
Purchase Day: 3	[0;3]	0,00	0,03	0,00	,19**		1,463,199
Purchase Day: 4	[0;3]	0,00	0,02	0,00	,18**		1,463,199
Purchase Day: 5	[0;2]	0,00	0,02	0,00	,18**		1,463,199
Purchase Day: 6	[0;3]	0,00	0,03	0,00	,18**		1,463,199
Purchase Day: 7	[0;2]	0,00	0,02	0,00	,17**		1,463,199
Purchase Day: 8	[0;3]	0,00	0,02	0,00	,16**		1,463,199
Purchase Day: 9	[0;2]	0,00	0,02	0,00	,15**		1,463,199
Purchase Day: 10	[0;3]	0,00	0,02	0,00	,14**		1,463,199
Purchase Day: 11	[0;2]	0,00	0,02	0,00	,13**		1,463,199
Purchase Day: 12	[0;3]	0,00	0,02	0,00	,15**		1,463,199
Purchase Day: 13	[0;3]	0,00	0,02	0,00	,15**		1,463,199
Purchase Day: 14	[0;2]	0,00	0,02	0,00	,14**		1,463,199
Purchase Day: 15	[0;2]	0,00	0,02	0,00	,15**		1,463,199
Purchase Day: 16	[0;4]	0,00	0,02	0,00	,14**		1,463,199
Purchase Day: 17	[0;3]	0,00	0,02	0,00	,16**		1,463,199
Purchase Day: 18	[0;2]	0,00	0,02	0,00	,14**		1,463,199
Purchase Day: 19	[0;2]	0,00	0,02	0,00	,14**		1,463,199
Purchase Day: 20	[0;3]	0,00	0,02	0,00	,15**		1,463,199
Purchase Day: 21	[0;6]	0,00	0,02	0,00	,16**		1,463,199
Purchase Day: 22	[0;5]	0,00	0,02	0,00	,17**		1,463,199
Purchase Day: 23	[0;2]	0,00	0,03	0,00	,19**		1,463,199
Purchase Day: 24	[0;4]	0,00	0,02	0,00	,18**		1,463,199
Purchase Day: 25	[0;2]	0,00	0,02	0,00	,17**		1,463,199
Purchase Day: 26	[0;3]	0,00	0,03	0,00	,19**		1,463,199
Purchase Day: 27	[0;2]	0,00	0,02	0,00	,18**		1,463,199
Purchase Day: 28	[0;5]	0,00	0,03	0,00	,20**		1,463,199
Purchase Day: 29	[0;3]	0,00	0,04	0,00	,26**		1,463,199
Purchase Day: 30	[0;7]	0,00	0,04	0,00	,29**		1,463,199
Purchase Day: 31	[0;2]	0,00	0,01	0,00	,09**		1,463,199
** Correlation is significant at the 0.01 level							

Table 4.20 PCA: Purchases and the Hour of the Purchase

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	,02	,13	6,50	1,00	**,18**	1,463,199
Purchase Hour: 12pm	[0;1]	,00	,00	0,00	,03**		1,463,199
Purchase Hour: 1:00 am	[0;2]	,00	,01	0,00	,05**		1,463,199
Purchase Hour: 2:00am	[0;1]	,00	,00	,00	,08**		1,463,199
Purchase Hour: 3:0am	[0;3]	,00	,02	0,00	,15**		1,463,199
Purchase Hour: 4:00am	[0;6]	,00	,03	0,00	**,20****		1,463,199
Purchase Hour: 5:00am	[0;4]	,00	,03	0,00	**,24****		1,463,199
Purchase Hour: 6:00am	[0;7]	,00	,04	0,00	**,26****		1,463,199
Purchase Hour: 7:00am	[0;4]	,00	,04	0,00	**,26****		1,463,199
Purchase Hour: 8:00am	[0;3]	,00	,04	0,00	**,26****		1,463,199
Purchase Hour: 9:00am	[0;2]	,00	,04	0,00	**,26****		1,463,199
Purchase Hour: 10:00am	[0;2]	,00	,03	0,00	**,25****		1,463,199
Purchase Hour: 11:00am	[0;2]	,00	,03	0,00	**,26****		1,463,199
Purchase Hour: 12:00am	[0;2]	,00	,04	0,00	**,26****		1,463,199
Purchase Hour: 1:00pm	[0;3]	,00	,03	0,00	**,25****		1,463,199
Purchase Hour: 2:00pm	[0;2]	,00	,04	0,00	**,27****		1,463,199
Purchase Hour: 3:00pm	[0;3]	,00	,04	0,00	**,29****		1,463,199
Purchase Hour: 4:00pm	[0;3]	,00	,04	0,00	**,27****		1,463,199
Purchase Hour: 5:00pm	[0;3]	,00	,03	0,00	**,24****		1,463,199
Purchase Hour: 6:00pm	[0;4]	,00	,03	0,00	,18**		1,463,199
Purchase Hour: 7:00pm	[0;3]	,00	,02	0,00	,11**		1,463,199
Purchase Hour: 8:00pm	[0;2]	,00	,01	0,00	,07**		1,463,199
Purchase Hour: 9:00pm	[0;2]	,00	,01	0,00	,05**		1,463,199
Purchase Hour: 10:00pm	[0;2]	,00	,00	0,00	,03**		1,463,199
Purchase Hour: 11:00pm	[0;1]	,00	,00	,00	,03**		1,463,199
** Correlation is significant at the 0.01 level							

Table 4.21 PCA: Purchases and the Number of Purchases

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	,02	,13	6,50	1,00	**,61**	1,463,199
Purchases 2010	[0;19]	,02	,13	6,50	**,92****		1,463,199
Purchases 2011	[0;4]	,00	,04	0,00	**,30****		1,463,199
** Correlation is significant at the 0.01 level							

Table 4.22 PCA: Purchases and Search Engine Name

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	,02	,13	6,50	1,00	**,24**	1,463,199
Search Engine GOOGLE	[0;26]	,01	,16	16,00	**,39****		1,463,199
Search Engine MSN	[0;5]	,00	,03	0,00	,07**		1,463,199
Search Engine YAHOO	[0;11]	,00	,03	0,00	,09**		1,463,199
Generic Search Engine	[0;26]	,01	,16	16,00	**,40****		1,463,199
** Correlation is significant at the 0.01 level							

Table 4.23 PCA: Purchases and Site Name

	Min-Max Value	Mean	Std. Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	,02	,13	6,50	1,00	**,05**	1,463,199
Aconion MCUK	[0;3882]	,22	7,15	31,97	,09**		1,463,199
Adjug6	[0;4194]	,41	7,14	17,21	,05**		1,463,199
Pepper MCUK	[0;52]	,06	,60	10,58	,00*		1,463,199
Vertisingcom MCU	[0;11]	,00	,01	1209,63	,00		1,463,199
Affiliate Window	[0;18]	,01	,14	18,05	,39**		1,463,199
Any Media	[0;555]	,41	8,48	20,66	,04**		1,463,199
AOLMCUK	[0;407]	,09	1,31	14,34	,03**		1,463,199
Audience Science	[0;286]	,03	,56	16,67	,00**		1,463,199
Brand Energy Email	[0;1]	,00	,00	,00	,01**		1,463,199
Context Web (1)	[0;1354]	,22	3,02	13,91	,00**		1,463,199
Drive PM	[0;3828]	,29	8,72	29,91	,05**		1,463,199
Ebay MCUK	[0;5139]	1,14	9,23	8,09	,02**		1,463,199
Facebook4	[0;7]	,00	,01	1209,63	,00		1,463,199
Itvcom	[0;4]	,00	,01	338,69	,00		1,463,199
ITVcom MCUK	[0;132]	,01	,42	30,09	,00**		1,463,199
Context Web (2)	[0;4]	,00	,00	901,60	,000		1,463,199
MCUK Quidco	[0;30]	,02	,27	15,09	,49**		1,463,199
KYahooQ2006	[0;2912]	,25	6,67	26,28	,09**		1,463,199
Sky MCUK	[0;873]	,01	,95	131,96	,00		1,463,199
Specific Media MCUK	[0;13550]	,43	17,42	40,66	,01**		1,463,199
TheTimes1	[0;13]	,00	,02	354,47	,01**		1,463,199
Unanimis MCUK	[0;5905]	,37	8,28	22,33	,01**		1,463,199
Up My Street MCUK	[0;3]	,00	,00	802,38	,00		1,463,199
Yahoo MCUK	[0;273]	,16	1,43	9,01	,02**		1,463,199
** Correlation is significant at the 0.01 level * Correlation is significant at the 0.05 level							

Table 4.24 PCA: Purchases and Type of Banner

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	0.02	0.13	6,50	1.00	**0,62**	1,463,199
Default Click	[0;12322]	0.05	15.28	305,06	0.00		1,463,199

Dynamic Click	[0;30]	0.03	0.31	10,33	**0.62****		1,463,199
Standard Click	[0;19057]	4.07	31.45	7,73	0.09**		1,463,199
**Correlation is significant at the 0.01 level							

Table 4.25 PCA: Purchases and Rank

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	0.02	0.13	6,50	1.00	**0,20**	1,463,199
Rank1_max	[1;23156]	4,18	45,55	10,90	0,07**		1,463,199
Rank2_max	[0;479]	6,42	18,43	2,87	0,01		25,573
Rank3_max	[1;19]	1,65	,56	0,33	**0,51****		25,573
**Correlation is significant at the 0.01 level							

Table 4.26 PCA: Purchases and Type of Segment

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	,02	0.13	6,50	1.00	**0,38**	1,463,199
Segment Bronze	[0;2]	,00	,02	0,00	,17**		1,463,199
Segment Gold	[0;7]	,00	,05	0,00	**,34****		1,463,199
Segment GS	[0;7]	,00	,07	0,00	**,53****		1,463,199
Segment Silver	[0;3]	,00	,01	0,00	,11**		1,463,199
Segment SB	[0;12]	,01	,10	10,00	,75**		1,463,199
**Correlation is significant at the 0.01 level							

Table 4.27 PCA: Purchases and Cost per Click

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	,02	,13	6,50	1,00	**(-) ,07**	1,463,199
Cost per Click Mean	[0;5,45]	,64	,63	0,98	(-) ,07**		12,816
**Correlation is significant at the 0.01 level							

Table 4.28 PCA: Purchases and Click Through Rate

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	,02	,13	6,50	1,00	**,29**	1,463,199
Mean Click Through Rate	[0;2]	,00	,01	0,00	,29**		1,456,904
**Correlation is significant at the 0.01 level							

Table 4.29 PCA: Purchases and Average Position

	Min-Max Value	Mean	Standard Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	,02	,13	6,50	1,00	**(-) ,17**	1,463,199
Average Position	[1;10]	1,39	,82	0,59	(-) ,17**		12,840
**Correlation is significant at the 0.01 level							

Table 4.30 PCA: Purchases and Search Click

	Min-Max Value	Mean	Std. Deviation	Coefficient of Variation	Pearson Correlation	Mean Correlation	N
Purchases: OD	[0;19]	,02	,13	7,52	1,00	**,40**	1,463,199
Search Click	[0;475]	,01	,16	13,50	,40**		1,463,199
**Correlation is significant at the 0.01 level							

From a statistical point of view, only the variables marketed with two or one 'stars' are significant to accomplish our research goals. On the other hand, as said in the first part of this thesis, decision makers must be integrated the quantitative standpoint with a qualitative approach (opinion mining and/or expert judgment) in order to optimize the level of marketing performance. Sometimes, the experience of the staff and the personal knowledge are a very important elements to improve marketing performance. Also, the high level of the coefficients of variation indicates possible cases of multicollinearity among variables.

Finally, the following table shows the mean correlation value related to each group of the variables and some important qualitative notes directly established with the Company EVP.

Table 4.31 Mean Correlation Value per Classes of Variables

Variable Group	Mean Correlation Value	Business Opinion
Purchases	1	Target or Dependent Variable
Quantity Sold	0,97	Redundant Variable
Type of Banner	0,62	Standard and Dynamic Click
Number of Purchases	0,61	Redundant Variable
Average Position Best Five	0,39	
Brand Search	0,38	
Segment	0,38	Low Predictive Value
Head Flag	0,34	Low Predictive Value
Search Click	0,40	
Click Through Rate	0,29	
Search Engine Name	0,24	Search Engine Google and Generic Search
Rank	0,20	Rank 1 Max
Match Type	0,18	Match Type Broad and Exact
Hour of the Purchase	0,18	Could be affected by External Factors
Name of the Campaign	0,17	Low Predictive Value
Day of the Purchase	0,17	Could be affected by External Factors
Creative Height	0,16	Low Predictive Value
Creative Width	0,14	Low Predictive Value
Min Search Click	0,12	Low Predictive Value

Keywords Campaign	0,10	Strategic Key Factors (Keywords Campaign: Brand and Top Performers)
Impression or Click	0,07	Strategic Key Factors (2010-2011)
Day of Impression or Click	0,06	Low Predictive Value
Hour of Impression or Click	0,05	Strategic Key Factors (Impression or Click from 1:00pm to 4pm)
Site Name	0,05	Strategic Key Factors (Affiliate Marketing)
Creative Type	0,04	Low Predictive Value
Keywords of Advertising Groups	0,04	Low Predictive Value
Advertisement Name	-	No available
Cost per Click	(-) ,07	Reverse Relationship
Average Position	(-) ,17	Reverse Relationship

Notice that qualitative approaches is indispensable. Expert judgment and opinion mining help decision makers in the choices of the most important variables to collect in the final database. In fact, despite a good value of correlation index, some variables will be eliminated from the final dataset because of irrelevant for the analysis. Variables related to the day and the hour of purchase has been wiped out from the ending database for instance. These variables sometimes could be affected from external factors (e.g. the date on which salaries are paid, time of break, time at which work starts and finish, etc.) difficult to control. Particular attention must be paid on the variables 'Standard Clicks' and 'Dynamic Click' because they could be a key variables for the analysis. In general, each potential customer before purchasing a service should perform a single standard click at least. For this reason, the variable cannot be eliminated from the dataset despite its low correlation value with the number of purchases. On the other hand, the high correlation value of the variable 'Dynamic Click' could suggest some anomalies in the data analyzed. Also, despite the really low value of the correlations index, it could be interesting to identify when (in terms of hours) potential customers generate impression or click in order to improve the accuracy of marketing performance.

These examples underlines, once again, the strategic importance of qualitative approaches within global organizations in order to get the best strategic business solution. In fact, to estimate the correlation index is a good start point but not enough for an accurate and effectiveness data analysis in today's global business

4.5 Multicollinearity Analysis

Before explaining the results from a predictive models it is worth defining a great database. Multicollinearity analysis is a good method for discovering redundant variables. The collinearity situation among variables could be a problem in the database analyzed. Literature identify different causes of multicollinearity (among others, improper use of dummy variables, including a variable that is computed from other variables in the equation, including the same or almost the same variable twice). These causes imply some sort of errors from researchers. Nevertheless, it may just be that variables actually are highly correlated (Lattin et al. 2003). The tables below refer to the

multicollinearity problem focusing the attention on two collinearity statistics: the tolerance and the variance inflation factor. The tolerance index is equal to $1-R^2$ [15]. A small tolerance value indicates that the variable under consideration is almost a perfect linear combination of the independent variables already in the equation and that it should not be added to the regression equation. All variables involved in the linear relationship will have a small tolerance. Instead, the Variance Inflation Factor (VIF) measures the impact of collinearity among the variables in a regression model. This statistic is equal to 1/Tolerance. An important advice is that in literature there is not a general rule about the interpretation of the VIF. High values of the VIF means that the variables are identical thus one of these must be deleted from the database. Generally, from a statistical point of view, VIF less than 10 or 3 are acceptable in literature. In our case, according to the EVP, we have accepted VIF less than 10.

Table 4.32 Collinearity Situation (First Interaction)

Variables	**COLLINEARITY STATISTICS**	
	TOLERANCE	VIF
Mean Click Trough Rate	,902	1,109
Average Position	,732	1,365
Rank1 Max	,016	62,923
Dynamic Click	**,000**	**3749,830**
Standard Click	,056	17,726
Average Position Best Five	,061	16,335
Brand Search	,175	5,705
Impression or Click at 1:00pm	,106	9,425
Impression or Click at 2:00pm	,107	9,370
Impression or Click at 3:00pm	,122	8,217
Impression or Click at 4:00pm	,154	6,506
Impression or Click 2010	,027	36,563
Impression or Click 2011	,036	27,653
Match Type Broad	,463	2,158
Match Type Exact	,126	7,947
Search Engine on Google	,111	8,994
Search Click	,050	19,953
Site Name: Conion MCUK	,635	1,574
Site Name: Adjug6	,588	1,699
Site Name: Affiliate Window	,001	990,412
Site Name: Drive PM	,617	1,620

[15] The coefficient R^2 is equivalent to the square of the linear correlation coefficient thus it takes values between 0 and 1. It is equal to 0 when there is not relationship among variables, while it is equal to 1 when the fit is perfect. (Giudici, 2003 p.91)

Site Name: MCUK Quidco	**,000**	**2744,429**
Site Name: MCUK Yahoo Q2006	,516	1,936

The variables Dynamic Click and Site Name: MCUK Quidco seem to be redundant. Just 18,515 out of 1,463,199 potential customers perform a dynamic click into the marketing campaign. As consequence, decision makers supposes that the variable could represente 'fraud clicks[16]'. It means that competitors could generate dynamic click creating a directly connection with the payment page (MCUK Quidco). These 'fraud click' increase the click trough rate of the competitor. From a business point of view 'fraud click' must be continuously monitored and reported to the competent authorities in this field. Finally, Table 4.33 confirms the redundancy between the variables.

Table 4.33 Dynamic Click and Site Name: MCUK Quidco: Descriptive Analysis

	Minimum Value	Maximum Value	Mean
Dynamic Click	0,00	30,00	0,03
Site Name: MCUK Quidco	0,00	30,00	0,02

The EVP suggests to eliminate from the dataset the variable Site Name: MCUK Quidco because of to know the number of Dynamic Click generated by potential customers could be more strategic than to know the name of a specific banner to which is exposed to a potential customer.

Table 4.34 Collinearity Situation (Second Interaction)

Variables	COLLINEARITY STATISTICS	
	TOLERANCE	VIF
Mean Click Trough Rate	,902	1,109
Average Position	,732	1,365
Rank1 Max	**,016**	**62,923**
Dynamic Click	,715	1,398
Standard Click	,056	17,725
Average Position Best Five	,061	16,330
Brand Search	,175	5,705
Impression or Click at 1pm	,106	9,408
Impression or Click at 2pm	,107	9,366
Impression or Click at 3pm	,122	8,194
Impression or Click at 4pm	,154	6,478
Impression or Click 2010	**,027**	**36,563**

[16] 'Click fraud occurs when a Web Users click on a sponsored link with the malicious intent of hurting a competitor or gaining undue monetary benefits' (Asdemir, Yurtseven, and Yahya, 2008 p.61).

Impression or Click 2011	,036	27,646
Match Type Broad	,464	2,157
Match Type Exact	,126	7,947
Search Engine on Google	,111	8,993
Search Click	,050	19,952
Site Name: ConionMCUK	,643	1,556
Site Name: Adjug6	,589	1,698
Site Name: Affiliate Window	,720	1,388
Site Name: Drive PM	,617	1,620
Site Name: MCUKYahooQ2006	,516	1,936

A special attention has been paid on the variable 'Rank1 Max' because of it seems to be correlated with many marketing activities (generation of standard click, number of impression or click in 2010 and 2011, the hour of impression or click performed by a potential customers). Table 4.35 confirms the variables redundancy.

Table 4.35 Pearson Correlation Analysis

Variables	Pearson Correlation Value
Mean Cost per Click	-,00
Mean Click Trough Rate	,00
Average Position	,01
Rank1 Max	1,00
Dynamic Click	0,09
Standard Click	**0,84****
Average Position Best Five	0,06
Brand Search	0,06
Impression or Click 1pm	**0,62****
Impression or Click 2pm	**0,56****
Impression or Click 3pm	**0,53****
Impression or Click 4pm	**0,53****
Impression or Click 2010	**0,90****
Impression or Click 2011	**0,77****
Match Type: Broad	0,04
Match Type: Exact	0,05
Quantity Sold	0,07
Number of Purchases 2010	0,06
Number of Purchases 2011	0,04
Search Engine on Google	0,06

Search Click	0,07
Site Name: Conion MCUK	0,24
Site Name: Adjug6	0,30
Site Name: Affiliate Window	0,10
Site Name: Drive PM	0,26
Site Name: MCUK Quidco	0,05
Site Name: MCUK Yahoo Q2006	0,37
Purchases	0,07

In this case, for a major accuracy of the analysis, it is worthwhile to eliminate the variable 'Rank1 Max' from the final database.

Table 4.36 Collinearity Situation (Third Interaction)

Variables	**COLLINEARITY STATISTICS**	
	TOLERANCE	VIF
Mean Click Through Rate	,902	1,108
Average Position	,732	1,365
Dynamic Click	,715	1,398
Standard Click	**,056**	**17,725**
Average Position Best Five	,061	16,330
Brand Search	,175	5,705
Impression or Click at 1pm	,106	9,408
Impression or Click at 2pm	,107	9,366
Impression or Click at 3pm	,122	8,194
Impression or Click at 4pm	,154	6,478
Impression or Click 2010	,075	13,417
Impression or Click 2011	,061	16,412
Match Type: Broad	,464	2,157
Match Type: Exact	,126	7,947
Search Engine on Google	,111	8,993
Search Click	**,050**	**19,951**
Site Name: ConionMCUK	,643	1,556
Site Name: Adjug6	,589	1,698
Site Name: Affiliate Window	,720	1,388
Site Name: Drive PM	,617	1,620
Site Name: MCUK Quidco	,516	1,936

Standard Clicks is a general click that could be played every time. It represents the first activity performed by potential customers before purchasing a service and its predictive value seem to be strong. But, despite this, the variable will be removed from the dataset due to its low correlation value (0,09) compare to the variable 'Search Click' (0,12).

Table 4.37 Collinearity Situation (Fourth Interaction)

Variables	COLLINEARITY STATISTICS	
	TOLERANCE	VIF
Click Trough Rate Mean	,905	1,105
Average Position	,732	1,365
Dynamic Click	,717	1,395
Average Position Best Five	**,061**	**16,329**
Search Brand	,175	5,704
Impression or Click at 1pm	,111	9,037
Impression or Click at 2pm	,108	9,258
Impression or Click at 3pm	,122	8,194
Impression or Click at 4pm	,172	5,807
Impression or Click 2010	,144	6,934
Impression or Click 2011	,102	9,840
Match Type: Broad	,464	2,154
Match Type: Exact	,126	7,941
Search Engine on Google	,111	8,986
Search Click	**,050**	**19,936**
Site Name: ConionMCUK	,671	1,491
Site Name: Adjug6	,653	1,531
Site Name: Affiliate Window	,721	1,387
Site Name: Drive PM	,660	1,514
Site Name: MCUK Quidco	,527	1,899

Comparing the variable 'Search Click' and the variable 'Average Position Best Five' the EVP suggests to eliminate the first variable because it represents just an exposure that is a search click, while the variable 'Average Position Best Five' shows the average position of a search term in the first five positions on the search engine. In other words, this latter variable has a stronger predictive value than the first variable thus really important to improve the marketing performance level.

Table 4.38 Collinearity Situation (Fifth Interaction)

Variables	COLLINEARITY STATISTICS	
	TOLERANCE	VIF
Click Through Rate Mean	,905	1,105
Average Position	,804	1,244
Dynamic Click	,717	1,395
Average Position Best Five	,131	7,621

Brand Search	,175	5,700
Impression or Click at 1pm	,111	9,036
Impression or Click at 2pm	,108	9,251
Impression or Click at 3pm	,122	8,194
Impression or Click at 4pm	,172	5,806
Impression or Click 2010	**,144**	**6,934**
Impression or Click 2011	**,102**	**9,838**
Match Type: Broad	,483	2,069
Match Type: Exact	,134	7,436
Search Engine on Google	,125	8,018
Site Name: ConionMCUK	,671	1,491
Site Name: Adjug6	,654	1,529
Site Name: Affiliate Window	,721	1,387
Site Name: Drive PM	,661	1,514
Site Name: MCUK Quidco	,527	1,896

Variables as 'Impression or Click 2010' and 'Impression or Click 2011' must be eliminating from the model because they are a decomposition of the Purchases variable. They have been useful just for marketing reasons in order to evaluate in which period the marketing campaign is more profitable.

Table 4.39 Collinearity Situation (Sixth Interaction)

Variables	**COLLINEARITY STATISTICS**	
	TOLERANCE	VIF
Click Trough Rate Mean	,906	1,103
Average Position	,804	1,243
Dynamic Click	,721	1,388
Average Position Best Five	,132	7,599
Brand Search	,176	5,692
Impression or Click at 1pm	,126	7,928
Impression or Click at 2pm	,110	9,065
Impression or Click at 3pm	,123	8,124
Impression or Click at 4pm	,175	5,719
Match Type: Broad	,484	2,066
Match Type: Exact	,135	7,430
Search Engine on Google	,125	7,997
Site Name: ConionMCUK	,924	1,082
Site Name: Adjug6	,855	1,170
Site Name: Affiliate Window	,721	1,387
Site Name: Drive PM	,904	1,106
Site Name: MCUK Quidco	,700	1,429

Table 4.40 provides the variables that will be include in the predictive data mining models (please, see next chapter). At this stage, the VIF confirms that in database

analyzed no variables are redundant. In fact, the VIF is less than 10. In conclusion, the following table provides the final database.

Table 4.40 Final Databases

Variables	Description	Variables Measures
Dynamic Click	Number of times that a potential customer click on Banner Moving	Scale
Click Through Rate	Click Through Rate. It is a way of measuring the success of an online campaign for a particular product or service. The click through rate advertisement is defined as the number of clicks on an advertisements divided by the number of times the advertisement is shown	Scale
Average Position Best Five	Number of times that the site name appears in the Top Best Five after the research in the search engine	Scale
Brand Search	Number of times that potential customer digits one of the brand name company in the search engine	Scale
Match Type: Broad	Number of times that a potential customer digits one of the keyword on the search engine	Scale
Match Type: Exact	Number of times that a potential customer digits an exact keyword on the search engine	Scale
Purchases	Target Variable. If the potential customers purchases a service online the variable is marks with 1 otherwise 0	Scale
Purchases: OD	Target Variable. Number of service purchased by potential customer	Scale
Impression or Click at 1pm	Number of times that a potential customer performs an impression or click at 1pm	Scale
Impression or Click at 2pm	Number of times that a potential customer performs an impression or click at 2pm	Scale
Impression or Click at 3pm	Number of times that a potential customer performs an impression or click at 3pm	Scale
Impression or Click at 4pm	Number of times that a potential customer performs an impression or click at 4pm	Scale
Average Position	Average position of a search term in the search engine	Scale
Search Engine on Google	Number of times that potential customers insert a company keyword on Google	Scale
Site Name: Aconion MCUK	Number of times that potential customers are exposed to a specific banner	Scale
Site Name: Adjug6	Number of times that potential customers are exposed to a specific banner	Scale
Site Name: Affiliate Window	Number of times that potential customers are exposed to a specific banner	Scale
Site Name: Drive PM	Number of times that potential customers are exposed to a specific banner	Scale
Site Name: Yahoo Q2006	Number of times that potential customers are exposed to a specific banner	Scale

The final dataset contains 1,463,199 potential customers and 19 quantitative variables related to their purchase behavior. The database is composed by quantitative and qualitative variables. The qualitative variables have been treated as quantitative variables because they have been categorized into groups. The target variable is dichotomous so it can assume only two values: 0 (bad potential customer) and 1 (good potential customer).

4.6 Linking Research Question to Research Strategy

The aim of this section is to underline the research purpose and the approach used in the dissertation. The tables and the graphs below describe in an accurate way the research

problem developed in the next chapter. Remark that the marketing campaign analyzed concerns three months specifically from December 2010 to February 2011.

In our case, the objective for the company is to discover the main marketing activities in which to maximize their future investments in order to increase the probability of customer conversion in the web marketing sector so decrease the number of churners.

Table 4.41 shows that between 2010 and 2011 the number of customer conversion, which is equal to the number of potential customers who purchase the service offered by the marketing campaign, is drastically dropped. Indeed, in 2011 less than 20,632 potential customers have purchased the service online than in 2010. This strong result is the main reason that leads to develop this research.

Table 4.41 Comparison between the number of client in 2010 and 2011

	2010	2011	Delta 2010 - 2011
Number of Potential Customers	1,463,199	1,463,199	
Number of Impression or Click	923,317	649,729	(-) 273,588
Number of Customers	23,039	2,407	(-) 20,632

These results emphasize the importance of the research questions that drive this PhD thesis and the weakness of statistical data analysis. These findings show as traditional statistical models are definitely inadequate to forecast marketing performance in global organizations with an outside-in perspective. Due to the considerable differences in terms of clients between 2010 and 2011 the company must be analyze in deep the data available in order to uncover emerging trends and discover hidden relations within data with high added value for their business growth. More precisely, Table 4.42 shows the percentage variation on average between both the number of purchases and the number of impressions or click for the total length of the campaign.

Table 4.42 Year Deviation between Purchases and Impression Click

	2010	2011	Average Percentage Variation
Average Purchases	1,63%	0,17%	**1,46%**
Average Impressions or Click	2,35%	1,80%	**0,55%**
Valid Observations	1,463,199	1,463,199	

In 2011 only 0,17% of potential customers on average purchased the service offered by the company compare 1,63% in 2010, while the average of impression click per potential customer is equal to 1,80% in 2011 and 2,35% in 2010. Thus, 2010 is more profittable than 2011. In details, the average of impression clicks in 2011 decreased of 0,55%, while the average of purchases is reduced of 1,46%. Additionally, despite a great number of impression clicks both in 2010 and in 2011, the number of purchases in 2011 decrese of 1,46% than 2010. This reseault is really scaremongering for the firm because in 2011 the average of customer conversion decrease more than one point percentage. One of the main reason of this could be joined to the micro-seasonality effect in December (for instance: Christmas Event). For this reason, marketers should better explore and investigate the data in order to identify the best marketing drivers on

which to concentrate their future investiments for minimizing the number of churners and maximize the probability of customer conversion. In fact, decision makers should focus their attention on the following questions:

- Why between 2010 and 2011 the probability of customer conversion decreases more than one point percentage?
- What is changed in the customer purchase behavior?
- In which marketing activities the company must focus their investments?
- Traditional Forecasting Models or Predictive Data Mining Models?

Chapter 5
Empirical Findings and Managerial Implications

5.1 Introduction

The main purpose of this empirical research is to demonstrate that data mining technologies are the main tools for global organizations with an outside-in perspective to predict accurate marketing performance. Special focus is paid in managing the potential customer risk of churn and on the identification of the best marketing driver which lead potential customers in a customer state. The data analyzed concern to launch a quarterly online marketing campaign. The first part of the campaign was proposed in December 2010 while the second part in January and February 2011. The target variable is related to the number of purchases in a given period and it can be both continuous and dichotomous because of different models foresee diverse types of data. The final database is composed by 1,463,199 potential customers and 19 quantitative variables related to their purchase behavior. Finally, it is worthwhile mentioning that the statistical outputs obtained refer to SPSS software.

This chapter is organized as follows. It starts with an exploratory analysis in order to describe a brief presentation of the data. After, it shows one of the most predictive data mining models such as logistic regression and decision trees able to identify the best marketing drivers that overlook the company decision making process. Then, it illustrates the evaluation criteria about the models previously implemented. Finally, a predictive scenario simulation related to the probability of customer conversion concludes the chapter.

5.2 About the Company

The dataset analyzed was provided by a global company that offers digital data driven marketing solutions across all interactive channels: digital, direct response, relationship based media and design. The Group works and complements each other bringing together professional expertise, proven strategic insight, and an advanced digital campaign management and optimization system, which allows maximizing impact for the advertiser interactive marketing investment. Currently this company operates in most different countries across Europe, North America, South America, Asia, Africa and Australia serving over 600 clients, including the market leaders in many industries such as: Air France, Danone, Expedia, Fidelity, France Telecom, Hyundai Kia, Nike, Peugeot/Citroen, Repsol, Reckitt Benckiser and Vodafone**,** amongst others. For privacy reasons no more details about the company can be given.

5.3 Description of the Database Analyzed

Before presenting the findings of the predictive data mining analysis it is worthwhile to describe the variables collected in the database placing attention on their statistical measures. Thus, Table 5.1 provides a description of them.

Table 5.1 Final Database

Variables	Description	Variables Measures
Dynamic Click	Number of times that a potential customers click on Banner Moving	Scale
Click Through Rate	Click Through Rate. It is a way of measuring the success of an online campaign for a particular product or service. The click through rate advertisement is defined as the number of clicks on an advertisements divided by the number of times the advertisement is shown	Scale
Average Position Best Five	Number of times that the site name appears in the Top Best Five after the research in the search engine	Scale
Brand Search	Number of times that potential customer digits one of the brand name company in the search engine	Scale
Match Type: Broad	Number of times that a potential customer digits one of the keyword on the search engine	Scale
Match Type: Exact	Number of times that a potential customer digits an exact keyword on the search engine	Scale
Purchases	Target Variable. If the potential customers purchases a service online the variable is marks with 1 otherwise 0	Scale
Purchases: Original Data (OD)	Number of times that a potential customer purchase the service online	Scale
Impression or Click at 1pm	Number of times that a potential customer performs an impression or click at 1pm	Scale
Impression or Click at 2pm	Number of times that a potential customer performs an impression or click at 2pm	Scale
Impression or Click at 3pm	Number of times that a potential customer performs an impression or click at 3pm	Scale
Impression or Click at 4pm	Number of times that a potential customer performs an impression or click at 4pm	Scale
Average Position	Average position of a search term in the search engine	Scale
Search Engine on Google	Number of times that potential customers insert a company keyword on Google	Scale
Site Name: Aconion MCUK	Number of times that potential customers are exposed to a specific banner	Scale
Site Name: Adjug6	Number of times that potential customers are exposed to a specific banner	Scale
Site Name: Affiliate Window	Number of times that potential customers are exposed to a specific banner	Scale
Site Name: Drive PM	Number of times that potential customers are exposed to a specific banner	Scale
Site Name: Yahoo Q2006	Number of times that potential customers are exposed to a specific banner	Scale

5.4 Exploratory Analysis of the Target Variables

Table 5.2 shows the purchases frequency of the potential customers in a given period. Just 1.70% of potential customers purchased the service at least once. A really restricted number of potential customers have purchased the service offered more than once. One potential customer bought 19 times a service proposed by the company. This result is interesting for marketers because of it is very unusual that a potential customer purchases the same service more than one time into the same marketing campaign.

Probably the potential customer could be an headquarter of the global company which bought the service for each subsidiary within their group. Besides, the low percentage of customer conversion indicates that the web marketing strategies implemented by the company are ineffective and inefficient. Thus, a data mining analysis is the main business solution able to discover hidden relationships in huge and massive databases for increasing the percentage of customer conversion and decrease the risk of churn. Just these hidden relationships are the success for global companies with an outside-in perspective in today's global business.

Table 5.2 Frequency of the variable 'Purchases: OD'

Purchases	Absolute Frequency	Relative Frequency
.00	1,437,766	98,262
1.00	24,754	1,692
2.00	608	0,042
3.00	44	0,003
4.00	13	0,001
5.00	5	0
6.00	1	0
7.00	2	0
8.00	1	0
10.00	2	0
13.00	1	0
15.00	1	0
19.00	1	0
Total	1,463,199	100

Table 5.3 notes the frequency of the purchase related to the target variable in a dichotomous form. The level of lost information is minimum because of only the 0,046% of potential customers purchased the service more than one time in the same marketing campaign.

Table 5.3 Frequency of the variable 'Purchases'

Purchases	Absolute Frequency	Relative Frequency
.00	1,437,766	98,26
1.00	25,433	1,74
Total	1,463,199	100

More precisely, from the table emerges that 98.26% of potential customers did not purchase the service online against 1.74% of them who bought it. In other words, 25,433 potential customers are 'good' because of they purchased the service offered through a marketing campaign, while 1,437,766 are 'bad' inasmuch they did not purchase the service online. Remark that despite the dichotomization of the variable, a few information has been lost. The percentage difference in terms of customer conversion is equal to 0.004%.

Table 5.4 Descriptive Values of the variables 'Purchases' and 'Purchases: OD'

	Minimum Value	Maximum Value	Arithmetic Mean	Variance	Standard Deviation	Coefficient of Variation
Purchases: OD	,00	19,00	,018	,020	,139	7,72
Purchases	,00	1,00	,017	,017	,130	7,64

Table 5.4 provides some position and dispersion indices related to the target variable. On average 1.80% of potential customers purchase the service online. The minimum and maximum values confirm that no outliers are present in the variables distribution. Finally, the high values of the coefficient of variation address that the arithmetic mean is an inaccurate indicator in explaining the distribution of the variables.

5.5 Pearson Correlation Analysis

Owing to the huge volume of the data, in order to discover the main variables to get into the predict data mining model we have estimated the value of the Person Correlation Index between the target variables (Purchases) and each independent variables contained in the database. Notice that the target variable used is quantitative continuous.

Table 5.5 Pearson Correlation Analysis

Variables	Pearson Correlation Value
Purchases	1,00
Dynamic Click	,62**
Average Click Through Rate	,29**
Average Position Best Five	,38**
Brand Search	,39**
Match Type: Broad	,16**
Match Type: Exact	,38**
Impression or Click at 1pm	,10**
Impression or Click at 2pm	,10**
Impression or Click at 3pm	,10**
Impression or Click at 4pm	,10**
Average Position	,40**
Search Engine on Google	,39**
Site Name: Aconiom MCUK	,09**
Site Name: Adjug6	,05**
Site Name: Affiliate Window	,41**
Site Name: Drive PM	,05**
Site Name: Yahoo Q2006	,09**

From the previously table emerges that no redundant variables are collected in the database. According to the EVP, the variables related to a specific banner, which are

exposed potential customers will be include in the predictive models because of they could contain important strategic value. In addition, the EVP suggests to continuously monitor the variable Dynamic Click because it will represent the number of 'fraud click' generated by potential customers. Finally, variables such as 'Average Position Best Five' and 'Brand Search' seem to be redundant even though they contain different business information.

The next sections provide the results related to the implementation of the main data mining models able to predict accurate marketing performance within global companies with an outside-in perspective.

5.6 Data Mining and Computational Predictive Models

The aim of this section is to develop accurate predictive models for discovering the main variables that are closely related to the probability of customer conversion in order to minimize the number of churners and maximize the profitability of the company. In fact, the final goal for decision makers is to identify the best marketing drivers in which concentrate the future investments. According to the current literature logistic regression models and decision trees compare to others data mining or computational predictive models such as neural networks and cluster analysis are one of the best models to forecast marketing performance, in particular the risk of churn. For instance, models based on the neural networks are inadequate to implement within global organization because of really complicate to analyze and interpret by decision makers. This point of view turns out to be in contrast with the statistical literature because of many papers argues about the strategic importance of the neural networks to predict marketing performance. In the light of these observations we have not developed neural networks models in this research. Notice that this dissertation want to propose a contact point between corporations and academies. Thus, it should be meaningless to develop predictive models difficult to implement in the reality. Once again, it emerges the enormous distance between corporations and academies. On the other hand, cluster analysis could be a good model to identify homogeneous groups of potential customers but it is a not rife models in churn predictions. It is helpful in marketing field during the decision making process in supporting the development of global marketing strategies, if implemented in a good manner. In our case, we have tried to develop both the K-Means[17] and Two-Steps[18] Cluster Analysis but unfortunately without success. Probably, the database analyzed is too much homogeneous in fact from a statistical point of view emerges just one cluster. Remark that we cannot develop a hierarchical cluster analysis due to the nature of the data available. According to Rezanková (2009) this model can be developed only whit categorical data. Finally, an example of Cluster Analysis based on K-means algorithms is provided by the following tables, while the following graph

[17] The simplest and most popular among iterative and hill climbing clustering algorithms is the K-means algorithm (Krishna and Narasimha Murty, 1999). The main aim of the K-mean algorithms is to divide M point in N dimensions into K clusters so that the within cluster sum of squares is minimized (Hartigan and Wong, 1979).

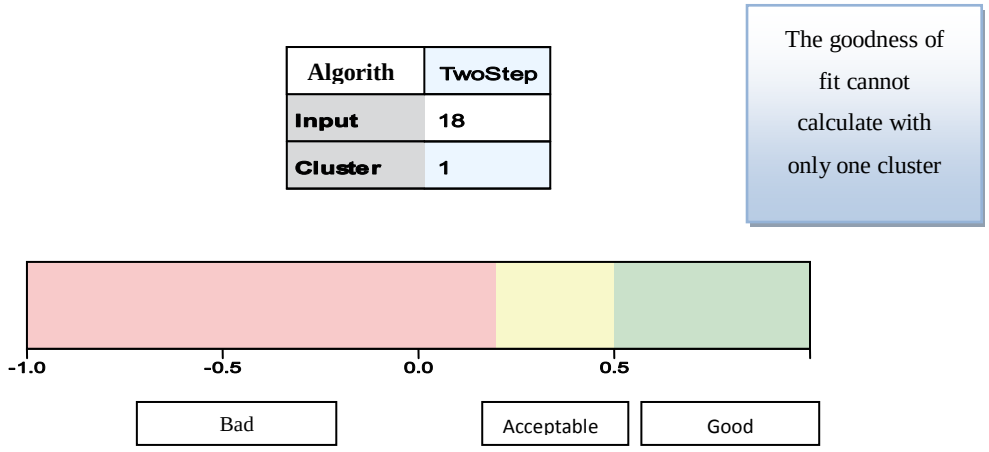
[18] The Two Step cluster method is a scalable cluster analysis algorithm designed to handle very large data sets. It can handle both continuous and categorical variables or attributes. It requires only one data pass. It has two steps: pre-cluster the cases (or records) into many small sub-clusters and cluster the sub-clusters resulting from pre-cluster step into the desired number of clusters. It can also automatically select the number of clusters (Yain, Murty, and Flynn, 1999).

shows the results of a Cluster Analysis based on Two-Step algorithm. The result related to the Two-step algorithm is very strange because of these latter is one of the main algorithm used for huge databases (Ester, Kriegel and Xu, 1995; Alexandrov, Gelboukn and Rosso, 2005).

Table 5.6 Cluster Analysis based on K-Means Algorithm

	Number of Cases in each Cluster	
Cluster	1	12,812
	2	2
	3	1
	Valid	12,815
	Missing	1,450,384

Figure 5.1 Cluster Analysis based on Two-Steps Algorithm: Summary Model



In our case the cluster analysis is not a good tool to detect the probability of churner. SPSS classify in a good way only 12,812 potential customers out of 1,463,199. Probably the database analyzed is too homogeneous. This result is coherent with the recent literature because of a really few scientific papers argue about the cluster analysis to resolve marketing problems (Morgan, 1978; Duen-Ren Liua and Ya-Yueh Shih, 2004; Nie, Chen, Zhang and Guo, 2010; Karahoca, 2011). But, despite this, a large number of companies develop cluster analysis based on various algorithms to identify the risk of churn falling in enormous traps. Often, decision makers do not know the real meaning of the statistical models or the algorithms but however implement them.

5.7 Hierarchical Logistic Regression Analysis

The target variable is dichotomous so it can assume only two values: 0 and 1. If the variable is equal to 0 it means that the potential customer did not buy the service. On the other hand, if the variable is equal to 1 it means that the potential customer purchased the service offered by the marketing campaign.

The following tables show the results of the hierarchical logistic regression. We have chosen this model because of it achieve to detect 'step by step' even of small changes of the variables. In fact, marketers through opinion mining or human judgment approaches can decide to insert or remove from the database. For instance, stepwise logistic regression based on backward or forward method are optimal from a statistical point of view but inaccurate from a business standpoint. In fact, the stepwise method automatically decide which variables to include in the model based on pre-defined statistical criteria. Instead, backward or forward methods include/exclude the best/worst predictor in every step. Notice that an opinion mining or human judgment approach is no applicable with the stepwise logistic regression. As consequence, hierarchical logistic regression should be one of the main approach to develop within global organization with an outside-in perspective able to predict the main drivers which could maximize the level of the marketing performance. Finally, the accuracy of the models is confirmed by the Criteria Based on the Loss Functions such as PPC and ROC curve.

Table 5.7 Hierarchical Logistic Regression Analysis (First Model)

Variables	Estimate Value	Standard Error	Wald Statistic	Significance Level *	Exp (β)
Average Click Through Rate	,673	,209	10,341	,001	1,959
Average Position	-,055	,029	3,568	,059	,947
Dynamic Click	2,503	,213	138,364	.000	12,222
Average Position Best Five	-,544	,075	52,390	,000	,580
Brand Search	,943	,057	272,236	,000	2,569
Impression or Click at 1pm	-,003	,006	,222	,637	,997
Impression or Click at 2pm	-,005	,007	,618	,432	,995
Impression or Click at 3pm	,000	,005	,004	,947	1,000
Impression or Click at 4pm	-,010	,005	3,738	,053	,990
Match Type: Broad	,364	,059	38,596	,000	1,439
Match Type: Exact	,947	,062	230,470	,000	2,579
Search Engine on Google	-,284	,061	21,967	,000	,753
Site Name: Conion MCUK	,047	,006	61,160	,000	1,048
Site Name: Adjug6	,002	,002	1,705	,192	1,002
Site Name: Affiliate Window	-1,120	,239	22,017	,000	,326
Site Name: Drive PM	-,003	,001	12,055	,001	,997
Site Name: MCUK Yahoo Q2006	,004	,001	7,338	,007	1,004

*Sig. < 0.05

Table 5.8 Hierarchical Logistic Regression (Second Model)

Variables	Estimate Value	Standard Error	Wald Statistic	Significance Level *	Exp (β)
Average Click Through Rate	.674	.209	10.447	.001	1.963
Average Position	-.055	.029	3.569	.059	.947
Dynamic Click	2.504	.213	138.486	.000	12.237
Average Position Best Five	-.545	.075	52.511	.000	.580
Brand Search	.944	.057	272.868	.000	2.571
Impression or Click at 4pm	-.016	.003	34.081	.000	.985
Match Type: Broad	.365	.059	38.910	.000	1.441
Match Type: Exact	.949	.062	231.410	.000	2.582
Search Engine on Google	-.285	.061	22.209	.000	.752
Site Name: Conion MCUK	.046	.006	60.970	.000	1.047
Site Name: Adjug6	.002	.001	1.818	.178	1.002
Site Name: Affiliate Window	-1.137	.238	22.882	.000	.321
Site Name: Drive PM	-.003	.001	16.679	.000	.997
Site Name: MCUK Yahoo Q2006	.003	.001	6.720	.010	1.003

*Sig. < 0.05

Table 5.9 Hierarchical Logistic Regression Analysis (Third Model)

Variables	Estimate Value	Standard Error	Wald Statistic	Significance Level *	Exp (β)
Average Click Through Rate	.655	.208	9.912	.002	1.925
Average Position	**-.055**	**.029**	**3.650**	**.056**	**.946**
Dynamic Click	2.505	.213	138.669	.000	12.250
Avgpos_best5	-.544	.075	52.346	.000	.581
Brand Search	.943	.057	272.632	.000	2.568
Impression or Click at 4pm	-.015	.002	36.135	.000	.985
Match Type: Broad	.366	.059	39.071	.000	1.442
Match Type: Exact	.950	.062	232.254	.000	2.586
Search Engine on Google	-.286	.061	22.395	.000	.751
Site Name: Conion MCUK	.047	.006	62.889	.000	1.048
Site Name: Affiliate Window	-1.142	.237	23.141	.000	.319
Site name: Drive PM	-.003	.001	16.598	.000	.997
Site Name: MCUK Yahoo Q2006	.003	.001	6.487	.011	1.003

*Sig < 0.05

Table 5.10 Logistic Regression Analysis (Fourth Model)

Variables	Estimate Value	Standard Error	Wald Statistic	Significance Level *	Exp (β)
Average Click Through Rate	4.251	.246	299.276	.000	70.144

Dynamic Click	8.032	.059	18427.986	.000	3078.246
Average Position Best Five	.969	.102	89.560	.000	2.635
Brand Search	1.889	.085	494.237	.000	6.611
Impression or Click at 4pm	.019	.003	45.070	.000	1.019
Match Type: Broad	1.752	.085	421.894	.000	5.766
Match Type: Exact	2.714	.095	810.470	.000	15.096
Search Engine on Google	.352	.096	13.488	.000	1.422
Site Name: Conion MCUK	.013	.001	144.151	.000	1.013
Site Name: Affiliate Window	-.510	.074	47.662	.000	.601
Site Name: Drive PM	-.005	.001	49.566	.000	.995
Site Name: MCUK Yahoo Q2006	.001	.001	.747	.387	1.001

*Sig. < 0.05

Table 5.11 Logistic Regression Analysis (Fifth Model)

VARIABLES IN THE EQUATION	Estimate Value	Standard Error	Wald Statistic	Significance Level *	Exp (β)
Average Click Through Rate	4.245	.246	298.819	.000	69.722
Dynamic Click	8.035	.059	18483.196	.000	3086.543
Average Position Best Five	.971	.102	90.027	.000	2.641
Brand Search	1.888	.085	494.367	.000	6.606
Impression or Click at 4pm	.019	.003	49.044	.000	1.019
Match Type: Broad	1.753	.085	422.188	.000	5.769
Match Type: Exact	2.716	.095	811.705	.000	15.113
Search Engine on Google	.351	.096	13.386	.000	1.420
Site Name: Conion MCUK	.013	.001	144.343	.000	1.013
Site Name: Affiliate Window	-.508	.074	47.389	.000	.602
Site Name: Drive PM	-.005	.001	49.640	.000	.995

*Sig. < 0.05

Table 5.12 Logistic Regression Analysis (Final Model)

VARIABLES IN THE EQUATION	Estimate Value	Standard Error	Wald Statistic	Significance Level *	Exp (β)
Average Click Through Rate	4,194	,245	294,090	,000	66,267
Dynamic Click	7,788	,041	**35851,716**	**,000**	**2411,917**
Average Position Best Five	,972	,102	90,300	,000	2,643
Brand Search	1,887	,085	494,667	,000	6,601
Impression or Click at 4pm	,018	,002	80,482	,000	1,018
Match Type: Broad	1,760	,085	427,276	,000	5,811
Match Type: Exact	2,725	,095	819,180	,000	15,254
Search Engine on Google	,346	,096	13,086	,000	1,414

*Sig. < 0.05

Being the relationship too strong from a statistical standpoint, the variable related to the number of Dynamic Click generated by potential customers must be removed from the model in order to polluting the other variables. Probably, as showed by Table 5.12, the results obtained can be erroneously due to the following motivations: fraud click, dynamic banner that links to the purchase page or to a genuine loyalty of customer to Quidco[19]. Instead, from a business standpoint, the variable 'Dynamic Click' should not be remove from the model because of it could be a strategic marketing driver for the maximization of the company profitability in today's competitive landscape.

5.8 Classification Decision Tree

A classification tree based on CART algorithm and on the Gini Impurity is provided in this paragraph. The main reason of this choice is that the CART algorithm is one of the most popular criteria used to manage business problems in global companies. Notice that this algorithm prefer to stop the growth of the tree trough a pruning mechanism rather than with a stopping criterion based on the significance of the chi-squared test such as the CHAID algorithms (Kass, 1980). Others decision trees algorithm such as C4.5 and C5.0 are presents in the current literature even if they are mainly used in engineering fields rather than in business contests (Giudici, 2003).

Before explaining the business strategic value provided by a classification tree a description of the database used in the development of this techniques is following presented.

Table 5.13 Description of the data used in the Classification Decision Tree

Variables	Description	Variables Measures
Dynamic Click	Number of times that a potential customer click on Banner Moving	Scale
Click Through Rate	Click Through Rate. It is a way of measuring the success of an online campaign for a particular product or service. The click through rate advertisement is defined as the number of clicks on an advertisements divided by the number of times the advertisement is shown	Scale
Average Position Best Five	Number of times that the site name appears in the Top Best Five after the research in the search engine	Scale
Brand Search	Number of times that potential customer digits one of the brand name company in the search engine	Scale
Match Type: Broad	Number of times that a potential customer digits one of the keyword on the search engine	Scale
Match Type: Exact	Number of times that a potential customer digits an exact keyword on the search engine	Scale
Purchases	Target Variable. If the potential customers purchases a service online the variable is marks with 1 otherwise 0	Scale
Impression or Click at 1pm	Number of times that a potential customer performs an impression or click at 1pm	Scale
Impression or Click at 2pm	Number of times that a potential customer performs an impression or click at 2pm	Scale
Impression or Click at 3pm	Number of times that a potential customer performs an impression or click at 3pm	Scale

[19] Specific banner to which it is exposed a potential customers.

Impression or Click at 4pm	Number of times that a potential customer performs an impression or click at 4pm	Scale
Average Position	Average position of a search term in the search engine	Scale
Search Engine on Google	Number of times that potential customers insert a company keyword on Google	Scale
Site Name: Conion MCUK	Number of times that potential customers are exposed to a specific banner	Scale
Site Name: Adjug6	Number of times that potential customers are exposed to a specific banner	Scale
Site Name: Affiliate Window	Number of times that potential customers are exposed to a specific banner	Scale
Site Name: Drive PM	Number of times that potential customers are exposed to a specific banner	Scale
Site Name: Yahoo Q2006	Number of times that potential customers are exposed to a specific banner	Scale

In order to avoid polluting the other variables we have directly removed from the model the variables 'Dynamic Click' in the construction of the classification decision tree. The next table provides an interpretation of the classification tree from a statistical standpoint. While, the following graph provides a classification tree helpful to identify the best marketing predictor related to the number of purchases in a given marketing campaign.

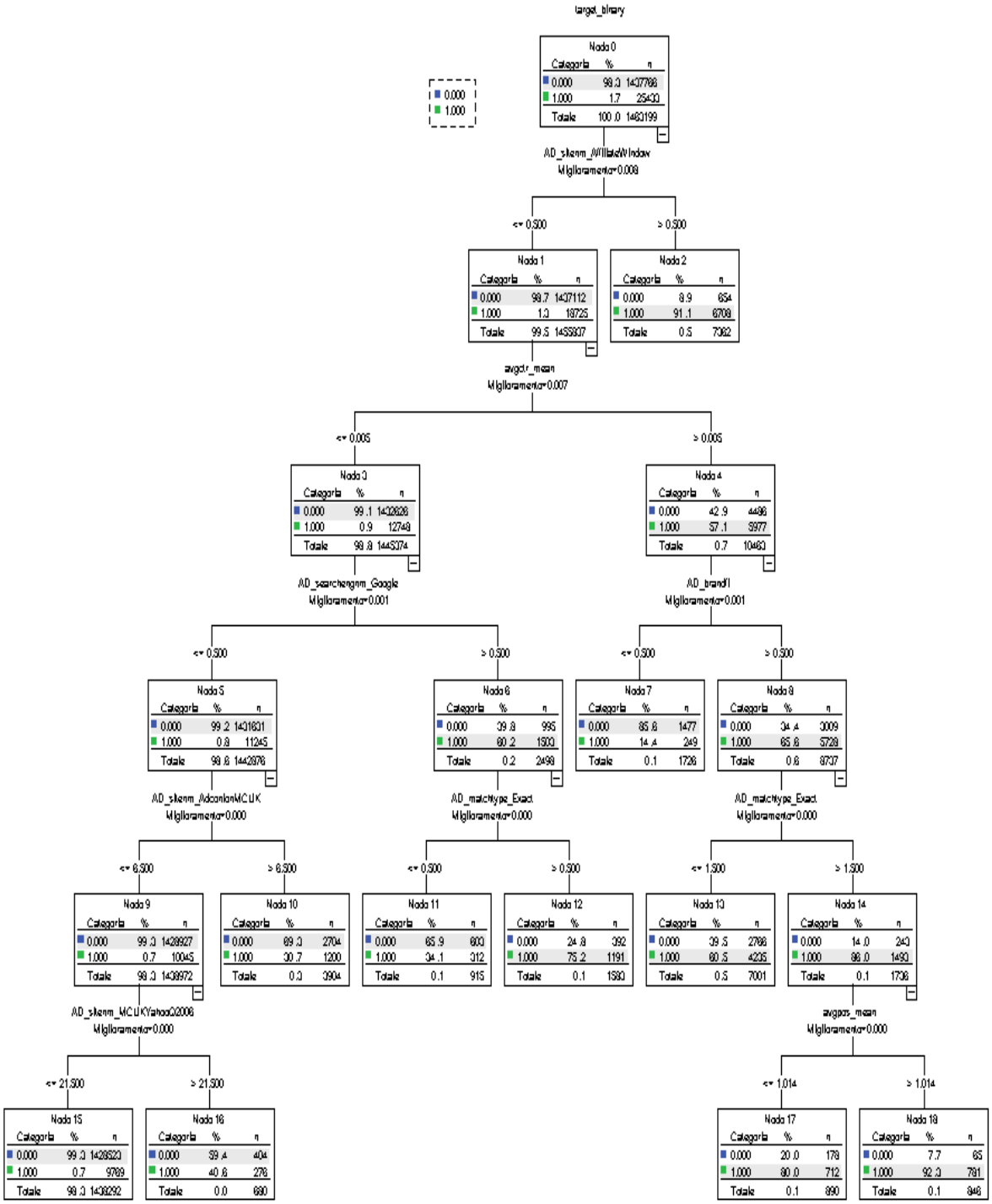
Table 5.14 Interpretation of the Classification Tree

Node	Customer category	Purchase probability	Number of potential customers
2	The potential customer has been exposed to a banner on an affiliate website at least one time.	91%	7,362
7	The potential customer has never been exposed to a banner on an affiliate website **AND** has been exposed to banners whose mean click through rate is greater than 0,5% **AND** has never digit the company's brand name on a search engine.	14%	1,726
10	The potential customer has never been exposed to a banner on an affiliate website **AND** has been exposed to banners whose mean click through rate is lower than 0,5% **AND** has never digits a campaign keyword on a search engine **AND** visited the 'Conion MCUK' website more than 6 times.	31%	3,904
11	The potential customer has never been exposed to a banner on an affiliate website **AND** has been exposed to banners whose mean click through rate is lower than 0,5% **AND** has searched on Google at least one time **AND** has never digits a specific campaign keyword on a search engine.	35,40%	915
12	The potential customer has never been exposed to a banner on an affiliate website **AND** has been exposed to banners whose mean click through rate is lower than 0,5% **AND** has searched on Google at least one time **AND** has digits a specific campaign keyword on a search engine at least once.	75%	1,583

13	The potential customer has never been exposed to a banner on an affiliate website **AND** has been exposed to banners whose mean click through rate is greater than 0,5% **AND** has digits the company's brand name on a search engine at least one time **AND** has digits a specific campaign keyword on a search engine either one or zero times.	61%	7,001
15	The potential customer has never been exposed to a banner on an affiliate website **AND** has been exposed to banners whose mean click through rate is lower than 0,5% **AND** has never digit a campaign keyword on Google **AND** has visited Conion MCUK website less than seven times **AND** has visited MCUKYahooQ2006 website less than 21 times	1%	1,438,292
16	The potential customer has never been exposed to a banner on an affiliate website **AND** has been exposed to banners whose mean click through rate is lower than 0,5% **AND** has never digit a campaign keyword on Google **AND** has visited Conion MCUK website less than seven times **AND** has visited MCUK Yahoo Q2006 website at least 22 times.	41%	680
17	The potential customer has never been exposed to a banner on an affiliate website **AND** has been exposed to banners whose mean click through rate is greater than 0,5% **AND** has digit the company's brand name on a search engine at least one time **AND** has digits exactly a campaign keyword on a search engine at least two times **AND** visited websites that have an average search position lower than 1,014.	80%	890
18	The potential customer has never been exposed to a banner on an affiliate website **AND** has been exposed to banners whose mean click through rate is greater than 0,5% **AND** has digit the company's brand name on a search engine at least one time **AND** has digits exactly a campaign keyword on a search engine at least two times **AND** visited websites that have an average search position greater than 1,014.	92%	846

The classification decision tree is coherent with the regression ones. In fact, 'Affiliate Websites' are a key driver of customer conversion. The 91% out of 7,000 potential customers that have visited an affiliate websites purchased the company's service online. Also, node 15 discriminates a customer category that is very unlikely to buy the service proposed by the company. More precisely, potential customers that neither visited affiliate websites nor search specific keyword campaign and have a probability of customer conversion lower than 1%. Instead, node 10 identifies a group of potential customer with a low probability (31%) of purchase. This group is mainly characterized by Conion MCUK visitors who do not visit affiliate websites and not search for campaign keywords on the internet. On the contrary, node 13 demonstrates that activities such as digit the company's name on a search engine and being exposed to banners with an higher click through rate increase the conversion probability. This category is composed by 7,001 potential customers and has a probability of purchase of 60%. Finally, the Company EVP confirms the coherence of the results from a managerial point of view.

Graph 5.1 Classification Decision Tree



5.9 Evaluation of the Predictive Models

This section aim to identify the best predictive data mining techniques in forecasting the main marketing drivers performed by potential customers in which global organizations should be concentrate their future investments.

5.9.1 Evaluation of the First Predictive Model

Table 5.15 shows the maximum likelihood estimates corresponding to the final model and the statistical significance of the parameters. For the entire explanatory variables we obtained a significance level (p-value) lower than 0.05. In other words, the eight explanatory variables selected through the hierarchical logistic regression are significantly associated with the purchases made by potential customers and are useful in explaining whether a variable is a good predictor for the customer conversion.

Table 5.15 Hierarchical Logistic Regression Analysis (Optimal Model)

VARIABLES IN THE EQUATION	Estimate Value	Standard Error	Wald Statistic	Significance Level *	Exp (β)
Average Click Through Rate	4,194	,245	294,090	,000	66,267
Dynamic Click	7,788	,041	**35851,716**	**,000**	**2411,917**
Average Position Best Five	,972	,102	90,300	,000	2,643
Brand Search	1,887	,085	494,667	,000	6,601
Impression or Click at 4pm	,018	,002	80,482	,000	1,018
Match Type: Broad	1,760	,085	427,276	,000	5,811
Match Type: Exact	2,725	,095	819,180	,000	15,254
Search Engine on Google	,346	,096	13,086	,000	1,414

*Sig. < 0.05

Table 5.16 Evaluation of the Hierarchical Logistic Regression Model

OBSERVED VALUES		PREDICTED VALUES		
		Purchases		Percentage Correct
		0	1	
Purchases	0	1,435,046	2,285	99,8
	1	4,363	15,210	77,7
Correctness of the Predictive Model				**99,5***

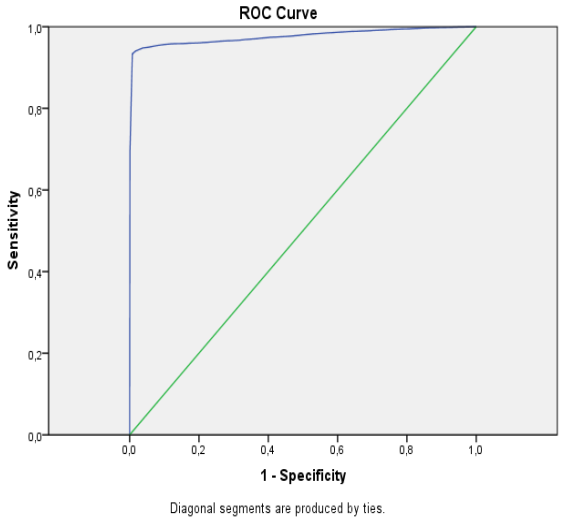
*The cut-off value is 0.05

Table 5.15 shows the strong predictive power of the number of dynamic click generated by potential customers. Probably, marketers are predicting sales using sales or there is a genuine loyalty of customer to Quidco. In contrast, the Company EVP address this unreal result to the fraud click. Remark that this variable must be checked at any time due to its strategic business meaning. Developing a marketing campaign without variables that represent moving banner would be meaningless from a business point of view. Finally, the confusion matrix confirms that the predictive accuracy of the hierarchical logistic regression model is really highly because is equal to 99.5%. In

general statistical software such as SPSS, STATA and R identify a cut-off point equal to 0,05 or 0,01.

The next graph shows the Receiver Operating Characteristic (ROC) curve. The latter criterion is stronger than the Confusion Matrix because it considers the interaction for any fixed cut-off value. In particularly this criteria considers on the x-axis 1-Specificity (false positive so the proportion of non-events predicted as events) and on y-axis the Sensitivity (the proportion of events predicted as such).

Graph 5.2 ROC curve: Purchases



According to Swets (1988) the accuracy of the model is highly accurate because the Area Under Curve (AUC) is equal to 97,6%. Qualitative approaches definitely overlook quantitative approaches because only an excellent knowledge of the market and the sector enables decision makers to draw accurate business strategies. Thereby, in global markets, data analysis exclusively based on quantitative or qualitative approaches are completely inadequate for predicting accurate business performance in particular in web marketing sector. Therefore, data mining analyses overtake the traditional statistical data analysis. Finally, from a scientific standpoint this model is due to the absurd value generate by the 'Dynamic Click' variable. As consequence, the next paragraph shows the same model without the variables 'Dynamic Click' in order to compare the results.

5.9.2 Evaluation of the Second Predictive Model

Table 5.17 point up the main marketing activities in which the company will need to invest in the future in order to maximize the probability of customer conversion and minimize the number of churners.

Table 5.17 Hierarchical Logistic Regression

Variables	Estimate Value	Standard Error	Wald Statistic	Significance Level*	Exp (β)
Average Click Through Rate	2,923	,223	171,955	,000	18,605
Average Position Best Five	,597	,086	47,973	,000	1,818
Search Brand	1,676	,071	549,417	,000	5,343

Impression or Click at 4pm	,109	,002	2909,043	,000	1,115
Match Type: Broad	1,289	,071	327,345	,000	3,631
Match Type: Exact	2,089	,079	700,440	,000	8,073
Search Engine on Google	,502	,080	39,553	,000	1,652

*Sig < 0.05

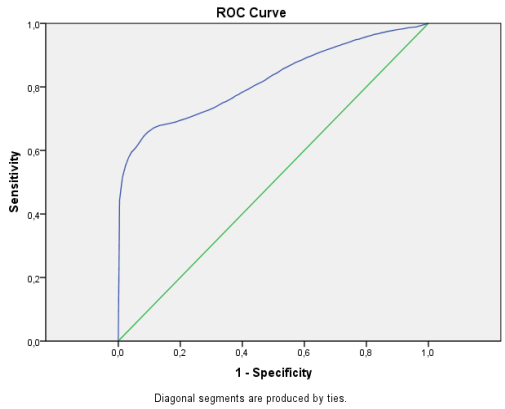
Table 5.18 Confusion Matrix

OBSERVED VALUES		PREDICTED VALUES		
		Purchases		Percentage Correct
		0	1	
Purchases	0	1,434,637	2,694	99,8
	1	13,017	6,556	33,5
Correctness of the Predictive Model				**98,9***

*The cut-off value is 0.05

Table 5.18 correctly classified the 99,8% of potential customers belonging to the class of bad potential customers, while the 33,5% of them are good potential customers. In this case, we can observe that the model previously implemented is really useful to forecast the probability of churn. However, the model is also adequate to predict the probability of customer conversion because of it predicts with an high accuracy 6,556 good potential customers out of 9,250. In fact, despite the exclusion of the variable 'Dynamic Click', the correctness of the predictive model is moderate because the AUC is equal to 82.10% (Swets, 1988). Graph 4.9 confirms as said before.

Graph 5.3 ROC curve: Purchases



5.9.3 Evaluation of the third Predictive Model

Table 5.19 notes that the classification decision tree correctly classify 13,627 out of 17,692 potential customer belonging the group of good potential customers. On the other hand, the classification tree correctly identify 1,433,711 out of 1,445,517 bad potential customer. In other words, the classification tree accurately predicts the 99,7% of potential customers who could be potential churners and the 53,6% of them as customer. In general, based on cut-off of 0,05 the accuracy of the classification tree is really high because of equal to 98,9% (Swets, 1988).

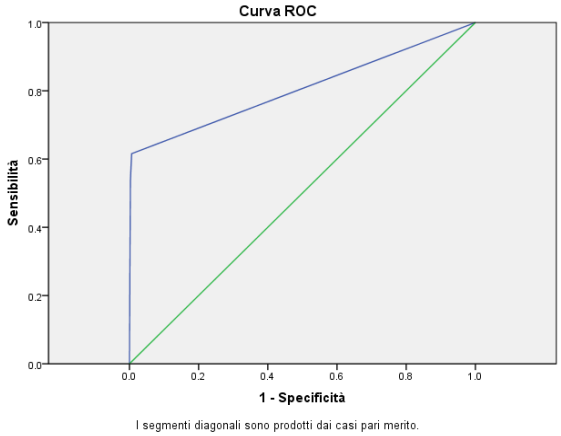
Table 5.19 Confusion Matrix

OBSERVED VALUES	PREDICTED VALUES		
	Purchase Variable		
	0	**1**	Percentage Correct
0	1,433,711	4,055	99,7
1	11,806	13,627	53,6
Total Overall Percentage			98,9*

*The cut-off value is 0.05

Due to the weakness of the Confusion Matrix the following graph provides a robust measure to evaluate the accuracy of the previously predictive model.

Graph 5.4 ROC curve: Purchases



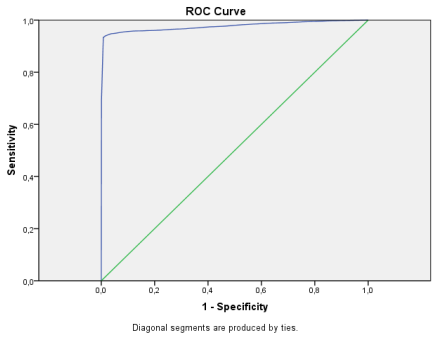
According to Swets (1988) the predictive accuracy of the classification decision tree is moderate because of is equal to 80.20%.

5.10 Hierarchical Logistic Regression and Classification Tree: Which is the best?

The following graphs show the difference among the previously predictive models implemented. Our final choice is based both quantitative and qualitative approaches. It was really difficult to select the main variables to include in the final database because of opinion mining approaches often overlooked the quantitative results. For this reason, the data mining methodology is an accurate solution in order to predict the probability of customer conversion thus to minimize the probability of churn in global organizations with an outside-in perspective in today's competitive landscape.

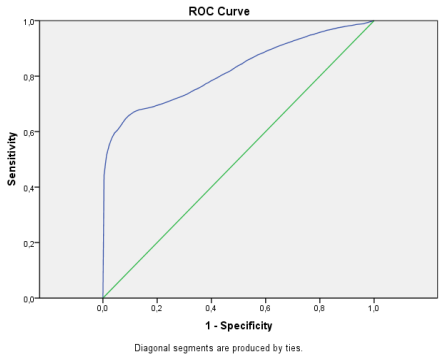
In order to establish the optimal model we have used the criteria based on the loss functions such as Confusion Matrix and the ROC curve. At this stage, decision makers must be paid the attention mainly on the ROC curve because of it is stronger than the confusion matrix from a statistical standpoint. Remark that the statistical results are often affected by many unpredictable exogenous factors that could shape the strategic business decisions in particular in today's global business.

Graph 5.5 ROC Curve Hierarchical Logistic Regression (First Model)



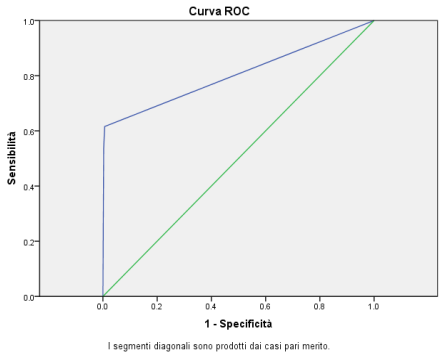
AUC = 97,60%

Graph 5.6 ROC Curve Hierarchical Logistic Regression (Second Model)



AUC = 82,10%

Graph 5.7 Decision Tree ROC Curve

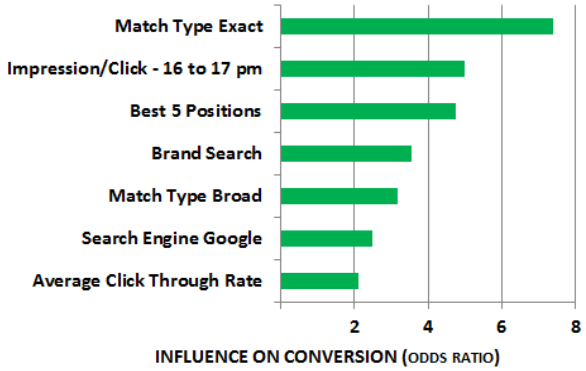


AUC = 80,20%

From the graphs emerge that both models have a strong predictive accuracy in explaining both the risk of churn and the customer conversion probability. In the first case an AUC equal to 97.6% means that the accuracy of the logistic regression model is highly accurate. Despite this astonishing accuracy this result is unreal both from a statistical and business perspective. In this case the data are soiled by the variable 'Dynamic Click'. In fact, according to the company this variable should be not considered from the strategic decision because affect too much the data and consequently the final results. First, as said before, the variable 'Dynamic Click' could represent: dynamic banner that links to the purchase page, a genuine loyalty of customers to Quidco, and fraud click. Second, the percentage of potential customers that perform a dynamic click is approximately equal to 1.7%. As a consequence of this low percentage, the company EVP address that the variable 'Dynamic Click' refers to fraud

click. Thirdly, the variable should be remove from the model because too unrealistic though strategic to get business decisions. On the other hand, the second model with an AUC equal to 82.10% has a moderate accuracy thus it is a strong model for identifying the best marketing activities in which the company will need to invest in the next year. Finally, the graph related to the classification tree guarantee that this predictive model is moderate accurate as well to estimate the main marketing drivers within global organizations. Analyzing the results emerges that an appropriate use of computational data mining models such as hierarchical logistical regression combined with a subjective human judgment help to improve the future accuracy of marketing performance in order to raise the number of customer conversion. Graph 5.2 shows the best marketing drivers that to lead a high level of customer conversion. The hierarchical logistic regression and the classification tree identify just 7 out 42 variables particularly significant in customer conversion into the web marketing sector.

Graph 5.8 Best Marketing Drivers



Indeed, Table 5.20 shows that the hierarchical logistic regression predict correctly 87% of the non-purchasers and 68% of the purchasers.

Table 5.20 Prediction Accuracy of the Model

Target Variable	Prediction		Percentage Correct
	no purchase	purchase	
no purchase	85.5%	13.2%	87%
purchase	0.4%	0.9%	68%

In other words, the hierarchical logistic regression based on an 'Enter Method' is an accurate data mining techniques to predict both the risk of churn and the probability of customer conversion.

5.11 Probability of Customer Conversion Simulation

The focus of this section is to test the effectiveness of the model previously designed by decision makers in order to get strategic marketing decisions. The following tables show how change the probability of customer conversion if it varies the marketing activities carried out by potential customer. We have attributed at each marketing activities a status level, which is equal to low if the potential customer do not make the activity and high when the potential customer performs it.

Table 5.21 What- if Scenario Simulation 1

Marketing Activity	Status
Average Click Through Rate	Low
Search Engine Google	High
Match Type Broad	High
Brand Search	High
Best 5 Positions	High
Impression/Click - 16 to 17 pm	High
Match Type Exact	High
Probability of customer conversion	
93.6%	

Table 5.22 What-if Scenario Simulation 2

Marketing Activity	Status
Average Click Through Rate	Low
Search Engine Google	Low
Match Type Broad	High
Brand Search	High
Best 5 Positions	High
Impression/Click - 16 to 17 pm	High
Match Type Exact	High
Probability of customer conversion	
74.5%	

Table 5.23 What-if Scenario Simulation 3

Marketing Activity	Status
Average Click Through Rate	Low
Search Engine Google	Low
Match Type Broad	Low
Brand Search	High
Best 5 Positions	High
Impression/Click - 16 to 17 pm	High
Match Type Exact	High
Probability of customer conversion	
38.2%	

Table 5.24 What-if Scenario Simulation 4

Marketing Activity	Status
Average Click Through Rate	Low
Search Engine Google	Low
Match Type Broad	Low
Brand Search	Low
Best 5 Positions	High
Impression/Click - 16 to 17 pm	High
Match Type Exact	High
Probability of customer conversion	
22.7%	

Table 5.25 What-if Scenario Simulation 5

Marketing Activity	Status
Average Click Through Rate	Low
Search Engine Google	Low
Match Type Broad	Low
Brand Search	Low
Best 5 Positions	Low
Impression/Click - 16 to 17 pm	High
Match Type Exact	High
Probability of customer conversion	
8.5%	

Table 5.26 What-if Scenario Simulation 6

Marketing Activity	Status
Average Click Through Rate	Low
Search Engine Google	Low
Match Type Broad	Low
Brand Search	Low
Best 5 Positions	Low
Impression/Click - 16 to 17 pm	Low
Match Type Exact	High
Probability of customer conversion	
2.5%	

Particularly interesting appears to be Scenario 2 with Scenario 3 and Scenario 4 with Scenario 5. In the first case the probability of customer conversion drastically drop down (less than 36.3%) if potential customers do not digit on the search engine certain key words bought in advance from the company. Indeed, in the second case the probability of customer conversion strongly decrease (less than 14.2%) if the marketing campaign sought by potential customer does not appear in the top five positions in the search engine.

To sum up, it is clear that the methodology used in this empirical research is effective in explaining the probability of customer conversion in global corporations. In our case, it would have been inadequate to implement traditional forecasting models based on time series because the marketing campaign was based just on three-month thus the length of time considered is too short. In addition, with the traditional statistical models is complex to analyze a huge number of variables at one and the same time. Finally, with new data, the approach proposed could be enhanced to predict futures sales using the current marketing activities.

Chapter 6
Conclusions and Discussion

Data is at the heart of many core business process in today's global business. The promise of the data mining is to find interesting patterns lurking an all these billions and trillions of bits collected in huge databases. Just finding patterns is not enough.

Businesses must respond to the patterns and act on them, ultimately turning raw data into information, information into action, and action into value. This is the virtuous cycle of data mining in a nutshell (Berry and Linoff, 2011).

Data mining needs to become an essential business process, incorporated into to other process including marketing, sales, customer support, product design, finance, and inventory control. This virtuous cycle place data mining in the larger contest of business, shifting the focus away from the discovery mechanism to the actions based on the discovery. This thesis emphasized actionable results from data mining proposing a potential contact point between academic contests and managerial worlds, and a new research field in the marketing area.

In this dissertation an extension of the current customer churn management definition has been proposed. In our case, a churner is a potential customer that did not purchase the service offering by the web marketing promotional campaign. In addition, we have investigated on the accuracy of the data mining techniques in managing potential customer churn in order to predict adequate marketing performance, especially the probability of the churn risk. Notice that no sampling techniques has been applied on the data available, for a major accuracy of the data analysis. In particular, we have tried to apply the main data mining techniques helpful for managing customer churn problem to a potential customers as to identify the best marketing drivers that lead potential customer in a customer state. Finally, a new kind of research called 'data mining research' has been proposed in literature. The data mining research is quantitative in nature but it achieves to mix both quantitative and qualitative (opinion mining) during the data process analysis. This is particularly useful for global companies with an outside-in perspective for maximizing the level of marketing performance.

Our research questions and their four assumptions have been accomplished. The research method developed is effective in explaining both the risk of churn and the probability of customer conversion. First of all, from the findings obtained emerge that marketing forecasts are often inaccurate because of the level of communication between academic and managerial reality is almost zero. In addition, there is an emerging need to build a link between marketing and statistics area and to draw up a new marketing research field both from a managerial and institutional point of view.

Traditional statistical models based on double moving average and exponential smoothing are definitely inadequate to forecast adequate marketing performance despite their continuous use by many global companies. First, with the double moving average models the predictive value of the data will be less sensitive to the actual changes and

moving average is not always a good trends indicator. Second, the predicted value always remains at the level of the past and cannot be expected to predict a higher or lower volatility of future. Instead, the exponential smoothing model requires a more complete historical data before starting the prediction and if season factors influences business sales a lot, times decomposition is more applicable than exponential smoothing. In our case, it would have been impossible to estimate both the probability of churn and the probability of customer conversion through double moving average or exponential smoothing. An important remark is that in literature a few scientific articles argue about this problem. In fact, the exponential smoothing model is the main approach used in many global organization in order for predicting marketing performance rather than their performance is totally inaccurate and without sense (Hyndman, Koehler, Ord and Snyder, 2008; Geng and Du, 2010). Once against it emerged the strategic importance of our research question argued in this dissertation. In other words, global companies develop inaccurate marketing forecast due to the implementation of inadequate predictive techniques able to manage huge amount of data. In particular, it is clear both from an analysis of the current literature and from a managerial reality an colossal cultural distance between these institutions, which could lead organizations in considerable traps in estimating both the probability of churn and the customer conversion. According to Geng and Du (2010) it is obvious that predictive data mining models compare to traditional statistical models are the key to solve business problems in presence of enormous amount of data. Finally, data mining techniques help global organizations to derive competitive knowledge from their enormous customer datasets, according to the previously literature review.

In order to accomplish our research goals a new generation of computation techniques and tools able to assist the extraction of useful competitive knowledge from the rapidly have been developed in this dissertation. Application of data mining techniques in this new field enhances the process by hastening it an improving its accuracy.

Hierarchical logistic regression based on the ‘Enter Method’ outperform the classification decision tree based on ‘CART’ algorithm. Two logistic regression models have been developed. The first model discovers an exceptionally strong performance of the ‘Web Site Quidco’ (Affiliate Marketing) in predicting customer conversion: almost any user of Quidco finally purchases the service online. The results can be erroneously due to a dynamic banner that links to the purchase page or to a genuine loyalty of customer to this specific banner. Being the relationship to strong from a quantitative stand point, the variables ‘Dynamic Click’ and ‘Web Site Quidco’ have been removed from the model so as to pollute the other variables. Instead, the second logistic regression model have identified seven marketing drivers that lead potential customers to purchase the service online. A clear connection between online marketing and customer conversion has been strongly confirmed by the data. The data analysis shows an association between purchases and almost all of the marketing trackers in the dataset. Potential customers that have searched the exact name of a keyword are seven time more likely to purchase the service than potential customer have not this.

Decision tree is coherent with regression one. In fact, affiliate web site is a key driver of customer conversion. In addition, the classification decision tree accurately discriminates a potential customer category that is very unlikely to buy the service proposed by the company. The criteria based on the loss functions confirm that the

predictive power of the hierarchical logistic regression is slightly more accurate than the classification tree in terms of AUC.
Neural networks models have been not developed for predicting both the probability of churn and the customer conversion, according to the literature review. Besides, the findings of a cluster analysis was irrelevant for our research goals probably for the high homogeneity of the database analysed.

According to Yang and Chiu (2006) is clear that neural networks analysis provides internal weights which are distributed throughout the network. As these weights do not provide insight into why the solution is valid; they are not readily understandable by the end user. Neural networking is often cited as a methodology that builds a black box. An investigation by Datta, Masand and Mani (2001) revealed that neural networks were only being used by a few companies. They state that a possible reason for this could be the lack of celerity of output. In contrast, decision trees and logistic regression are characterized by 'simple' if then rules that can sometimes be easy understand. In particular, regression models provide a straightforward relationship between the independent variables and the prediction and it could be the best data mining model able to predict the risk of churn in a manner way. It is very important to chose the best algorithm able to provide accurate and optimal estimation of the unknown parameter. In other words, if the model accurately identify the relationship sought and if these relationships make sense to then, then they more readily accept the validity if the model (Hadden, Tiwari, Roy and Ruta, 2005).

The analysis of the literature notes that the future selection phase in building potential customer churn model is fundamental in order to implement optimal strategic business decision in today competitive landscape. During the research it could lead several advantages such as a significant improvement over the performance of a classification by reducing the number of bands to a smaller set of informative features with a consequently greatly reduction of the processing time.

A certain cases lower-dimensional databases could be better, especially when the availability of the training dataset is limited. This fact is particularly true for neural networks. Despite these strength, in the literature there is a lack regarding the feature selection problem, while many authors have focused mainly developing future extraction methods.

The main trends for developing models to predict customer behaviour include regression analysis, neural networks and decision trees. Regression analysis is the most popular and accurate choice to manage the customer churn problem. Instead, neural networks have received slight more interest than decision trees. Focusing the attention on churn prediction we can observe that decision trees, neural networks, support vector machine, statistical methods, and some new techniques have already been investigated. No work has been carried out specifically for the purpose of customer churn management, using certain other powerful predictive techniques such as Bayesian neural networks. Fuzzy logic has never been investigated in either churn management or customer relationship management, for example.

From the literature review we have identified four further research. The first area concerns data semantics. Only a very limited amount of scientific articles argue about

the data semantics. As said before, a customer churn management framework cannot be developed and implemented without a clear knowledge and understanding of the data. As consequence, the authors propose the development of a framework that could be followed by researchers and developers to achieve competence at the data semantics of development. The second area related to the future selection because of the data that could be extract from enormous databases should be completely accurate. The current literature suggests that previous work done for predictive customer churn has either neglected a feature selection phase, or failed to document one. Researchers suggest to investigate about the best variables through neural networks models. This model could be helpful both companies and researchers to extract all variables with high predictive value collected in the datasets studied. The third area concerns the actual predictive model. Regression analysis and decision trees models have been used due to their usefulness in prediction and classification both for researchers and companies, independently from their size. The known lack in neural networks in churn prediction is connect to the fact that classification rules are not output in an easily understandable form. Some authors note that neural networks combined with a powerful rule could be a very power customer churn prediction model. No framework that include the causes of churn are developed in the current literature. Thus, there is an emerging need of this. Finally, the fourth area refers to the inclusion of a model able to divide loyal and non-loyal customers.

Futures research could concerns an analysis of the synergy between neural networks and genetic algorithm in order to draw an accurate churn prediction model able to discover the best variables useful to predict customer churn. Also, a framework for calculating 'loyalty' index and an analysis of the causes of churn should be provided. A text mining analysis joined with an opinion mining could support the design of this new framework. Instead, from a methodological standpoint, whit new data the model created can be enhanced to predict future purchases using current marketing activities. It could be interesting to develop a latent cluster analysis for identifying potential customer partition based on their behaviour and a survival analysis for estimating how many time customers will remain so in the company. In addition, we can test the effectiveness of the data mining techniques in small-medium enterprises (SMEs), in international and multinational companies. Data mining methodology could be a strategic tool for global and international entrepreneurship in order to discover new business opportunities and maximize the level of companies profitability.

Finally, marketing literatures makes data mining seem to easy. Just apply the automated algorithms created by the best minds such as, neural networks, decision trees, and genetic algorithms, and decision makers will be on good way to untold successes. Although algorithms are important, data mining solution is more than just a set of powerful techniques and data structures. The techniques must be applied to the right problem, on the right data. In fact, the virtuous cycle of the data mining is an interactive learning process that builds on results over time (Berry and Linoff, 2011).

In conclusion, the virtuous cycle of data mining could be defined as an interactive learning process that builds on results overtime. Success in using data will transform global organizations from reactive to proactive. This is the virtuous cycle of data mining developed in this dissertation, for extracting maximum benefit from the techniques described in this research.

Bibliography

Ahmed, S. R. (2004). Applications of data mining in retail business. *Information Technology: Coding and Computing*, 2:455-459.

Ahn, JH., Han, SP., Lee, YS. (2006). Customer churn analysis: churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications Policy*, 30 (10-11):552-568.

Ahn, WH., Kim, WJ., Par, D. (2004). Content-aware cooperative caching for cluster-based web servers. *Journal of Systems and Software*, 69(1-2):75-86.

Alexandrov, M., Gelboukn, A., and Rosso, P. (2005). *An Approach to Clustering Abstracts.* Berlin, Germany: Springer-Verlag.

Apenyo, K. (1999). Using the entity-relationship model to teach the relational model. *SIGCSE Bulletin*, 31:78-80.

Asdemir, K., Yurtseven, O., and Yahya, MA. (2008). An Economic Model of Click Fraud in Publisher Networks. *International Journal of Electronic Commerce*,13:61-90.

Athanassopoulo, AD. (2000). Customer Satisfaction Cues to support Market Segmentation and Explain Switching Behaviour. *Journal of Business Research*, 47(3):191-207.

Au, ST., Guangqin, M., and Rensheng, W. (2011). Iterative Multivariate Regression Model for Correlated Responses Prediction. *International Conference on Cyber-enabled distributed computing and knowledge discovery.*

Au W., Chan CC., and Yao X. (2003). A novel evolutionary data mining algorithm with application to churn prediction. *IEEE Transaction on Evolutionary Computation*, 7: 532-45.

Auh, S., and Johnson, MD. (2005). Compatibility effects in evaluations of satisfaction and loyalty. *Journal of Economic Psychology*, 26:35-57.

Avrachenkov, KE., Sanchez, E. (2002). Fuzzy Markov chains and decision-making. *Fuzzy optimization and Decision Making*, 1:143-59.

Baesens, B., Verstaeten, G., Van den Poel, D, Egmont-Peterson, M., Van Kenhove, P., and Vanthienen, J. (2004). Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers. *European Journal of Operational Research*,156:508-23.

Baesens, B., Viaene, S., Van den Poel, D., Venthienen, J., and Dedene, G. (2002). Bayesian neural network learning for repeat purchase modeling in direct marketing. *European Journal of Operational Research,* 138(1):191-211.

Bartlett, PL. (1998). The Minimax Distortion Redundancy in Empirical Quantizes Design. *IEEE Transaction of Information Theory,* 44(5):1802-1813.

Bassie, LJ. (1997). Harnessing the power of intellectual capital. *Training & Development,* 51(2), 25-30.

Beckman, T. (1997). A Methodology for KM. International Association of Science and Technology for Development (IASTD*). Al and Soft Computing Conference,* Banff, Canada.

Belbaly, N., Benbya, H., Meissonier, R. (2007). *An empirical investigation of the customer Knowledge creation impact on NPD Performance.* In: Proceedings of the 40th Hawaii International Conference on System Sciences.

Berne, C., Mugica, JM, and Yesus, YM. (2001). The effect of variety-seeking on customer retention in services. *Journal of Retailing and consumer services,* 8(6):335-345.

Berry, MJA., and Linoff, GS. (2004). *Data mining techniques for marketing, sales, and customer relationship management.* New York, USA: Lohn Wiley & Sons, Inc.

Berry, MJA., and Linoff, GS. (2011). *Data mining techniques for marketing, sales, and customer relationship management.* New York, USA: Lohn Wiley & Sons, Inc.

Berson, A., Smith, S., and Thearling, K. (2000). *Building data mining applications for CRM*. McGraw-Hill.

Best, RJ. (2004). Market-Based Management. Upper Saddle River, New Jersey, USA: Prentice Hall.

Bloemer, JMM.; Brijs, T., Vanhoof, K., and Swinnen, G. (2003). Comparing complete and partial classification for identifying customers at risk. *International Journal of Research in Marketing,* 20(2):117-131

Bloemer, J., Brijs, T., and Swinnen, G. (2002). Comparing complete and partial classification for identifying customers at risk. *International Journal of Research in Marketing,*20:117-131.

Boone, DS., and Roehm, M. (2002). Retail segmentation using artificial neural networks. *International Journal of Research in Marketing,* 19:287-301

Bortiz, JE., and Kennedy, DB. (1995). Effectiveness of neural network types for prediction of business failures. *Expert System with Application,* 9:503-512.

Bose, I, and Mahapatra, R. (2001). Business Data Mining-A Machine Learning Perspective. *Information & Management*, 39(3):211-225.

Brachman, R.J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G., and Simoudis, E. (1996). Mining business databases. *Communication of the ACM*, 39(11):42-48.

Bradley, AP. (1997). The use of the area under the ROC curve in the evaluation of the machine learning algorithms. *Pattern Recognition*, 7:1145-1159.

Breiman, L., Friedman, JH., Olshen, RA., and Stone, CJ. (1984). *Classification and regression trees.* California: Wadsworth.

Bryman, A. (1988*a*). Quantity and Quality in Social Research. London, UK: Routledge.

Bryman, A., Bell, E. (2007) Business Research Methods. New York, USA: Oxford University Press.

Bryman, A., Bell, E. (2011) Business Research Methods. New York, USA: Oxford University Press.

Bryson, KM. (2008). Post-pruning in regression tree induction: An integrated approach. *Expert Systems With Applications*, 34(2):1481-1490.

Brondoni, S. (2007). *Market Driven Management ed economia d'impresa globale.* In Silvio M. Brondoni (2007), Market-Driven Management e mercati globali, Torino, Italia: Giappichelli, 19-63.

Brondoni, S. (2008). Market-Driven Management, Competitive Space and Global Network. *Symphonya Emerging Issue in Management* (www.unimib.it/symphonya), 1.

Brondoni, S. (2009). Market-Driven-Management, Competitive Customer Value and Global Network. *Symphonya Emerging Issue in Management* (www.unimib.it/symphonya), 2.

Brondoni, S., and Lambin, JJ (2001). Overture de 'Market-Driven Management'. *Symphonya Emerging Issue in Management* (www.unimib.it/symphonya), 1.

Buchnowska, D. (2011). *Customer Knowledge Management Models: Assessment and Proposal.* Department of Business Informatics. University of Gdansk, Sopot, Poland.

Buckinx, V., Verstraeten, G., and Van den Poel, D. (2007). Predicting customer loyalty using the internal transactional database. *Journal of Interactive Marketing*, 32(1):125-134.

Bueren, A., Schierholz, R., Kolbe, L., Brenner, W. (2004). *Customer Knowledge management-improving performance of customer relationship management with*

knowledge management. In: Proceedings of the 37th Annual Hawaii International Conference on System Sciences.

Burez, J., and Van den Poel, D. (2007). CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications,* 32(2):277-288.

Cameron, S., and Price, D. (2009). *Business Research Methods. A Practical Approach.* UK: CIPD Enterprise Limited.

Canning, G. (1982). Do a value analysis of you customer base. *Industrial Marketing Management*, 11:89-93.

Carneiro, A. (2000). How does knowledge management influence innovation and competitiveness. *Journal of Knowledge Management,* 4(2):87-98.

Carter, S., Lee, K. (2005). Global Marketing Management. UK: Oxford University Press.

Chen, ZY, Fan, ZP, and Sun, M. (2012). A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data, *European Journal of Operational Research,* 223(2):461-472.

Chen, IJ, and Popovich, K. (2003). Understanding customer relationship management (CRM). *Business Process Management Journal*, 9:672-88.

Chen, Y., and Rosenthal, RW. (1996). Dynamic duopoly with slowly changing customer loyalties. *International Journal of Industrial Organization,* 14:269-96.

Chen, WC., Hsu, CC, and Hsu, JN. (2011). Optimal selection of potential customer range through the union sequential pattern by using a response model. *Expert Systems with Application,* 38:7451-7461.

Chen, GQ., Wei, Q., Liu, D., and Wets, G. (2002). Simple association rules (SAR) and the SAR-based rule discovery. *Computers & Industrial Engineering,* 43(4):721-733.

Chiang, DA., Wang, YH., and Chen, SP. (2010). Analysis on repeat-buying patterns. *Knowledge-Based System,* 23:757-768.

Chiang, D., Wang, Y., Lee, S., and Lin, C. (2003). Goal-oriented sequential pattern for network banking and churn analysis. *Expert Systems with Applications,* 25:293-302.

Cho, YB., Cho, YH., and Kim, SH. (2005). Mining changes in customer buying behaviour for collaborative recommendations. *Expert Systems with Applications,* 28(2):359-369.

Chueh, HE. (2011).Analysis of marketing data to extract key factors of telecom churn management. *African Journal of Business Management,* 5(20):8242-8248.

Cooper, D., and Schindler, P.S. (2008). *Business Research Methods.* Berkshire, UK: McGraw-Hill Higher Education.

Corniani, M. (2002). Demand-Bubble Management, Corporate Culture and Market Complexity. *Symphonya Emerging Issues in Management* (www.unimib.it/symphonya), 2.

Corniani. M (2005). Market, Segment and Demand Bubbles. *Symhonya Emerging Issues in Management (*www.unimib.it/symphonya), 2.

Crespo, F., Weber, R. (2004). A methodology for dynamic data mining based on fuzzy clustering. *Fuzzy Sets and Systems,*150:267-84.

Coussement, K., Benoit, D.F., and Van den Poel, D. (2010). Improved marketing decision making in a customer churn prediction context using generalized additive models. *Expert Systems with Applications,* 37(3):2132-2143.

Coussement, K., and Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications,* 34(1):313–327.

Coussement, K., and Van den Poel, D. (2009). Improving Customer Attrition Prediction by Integrating Emotions from Client/Company Interaction Emails and Evaluating Multiple Classifiers, *Expert Systems with Applications,* 36(3):6127-6134.

Daneshgar, F., and Bosanquet, L. (2010). Organizing Customer Knowledge in Academic Libraries. *Electronic Journal of Knowledge Management,* 8(1):21-32.

Day, GS. (1990). *The Market-Driven Strategy.* New York, US: The Free Press.

Day, GS. (1994). The capabilities of market-driven organizations. *The Journal of Marketing,* 58(4):37-52.

Day, GS. (1998). What Does It Mean to be Market-Driven? *Business Strategy Review,* 9(1):1-14.

Day, GS. (1999). *The Market-Driven Organization.* New York, US: The Free Press.

Day, GS. (2000). Managing market relationships. *Journal of the Academy of Marketing Science,* 28(1):24-30.

Day, GS (2001). Market Driven Winners. *Symphonya Emerging Issues in Management* (www.unimib.it/symphonya), 2.

Datta, P., Masand, B., Mani, DR., and Li, B. (2001). Automated cellular modeling and prediction on a large scale. *Issues on the Application of Data Mining,* 485-502.

De Bock, KW., and Van den Poel, D. (2011). An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. *Expert Systems with Applications*, 38(10):12293-12302.

DeLong, ER., DeLong, D.M. and Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic: a non-parametric approach. *Biometrics,* 44:837-845.

Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods*. New York, US: McGraw-Hill.

Deshpande, R., and Webster, FE. (1989). Organizational Culture and Marketing: Defining the Research Agenda. *The Journal of Marketing,* 53(1):3-15.

Dierkes, T., Bichler, M., and Krishnan, R. (2011). Estimating the effect of word of mouth on churn and cross-buying in the mobile phone market with Markov logic networks. *Decision Support Systems,* 51(3):361-371.

Duen-Ren L., and Ya-Yueh, S. (2004). Integrating AHP and data mining for product recommendation based on customer lifetime value. *Information & Management,* 42(3): 387-400.

Ester M., Kriegel HP, and Xu X. 1995. (1995). *A Database Interface for Clustering in Large Spatial Databases.* First International Conference on Knowledge Discovery and Data Mining. Montreal, Canada: AAAI Press.

Farvaresh, H., and Sepehri, MM. (2011). A data mining framework for detecting fraud in telecommunication. *Engineering Application of Artificial Intelligence*, 24(1):182-194.

Fayyad, UM. (1996). Data Mining and Knowledge Discovery: Making Sense Out of Data. *IEEE EXPERT,* 10:20-25.

Fayyad, UM., Shapiro, P., and Smyth, P. (1996). *From Data Mining to Knowledge Discovery: An Overview.* In Advances in Knowledge Discovery and Data Mining, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, MIT Press, 471-494.

Fayyad, UM., Shapiro, P., and Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM,* 39(11):27-34.

Fayyad, U., and Stolorz, P. (1997). Data Mining and KDD: Promise and challenges. *Future Generation Computer Systems,* 13:99-115.

Fayyad, UM., and Uthurusamy, R. (2002). Evolving data into mining solutions for insights. *Communications of the ACM,* 45(8):28-31.

Fleming, KN. (2003). Markov models for evaluating risk-informed in-service inspection strategies for nuclear power plant piping systems. *Reliability Engineering and System Safety*, 83:27-45.

Fletcher, D., and Goss, E. (1993). Forecasting with neural network: An application using bankruptcy data. *Information and Management*, 24(3):159-167.

Figini, S., and Giudici, P. (2009). *Applied Data Mining for Business and Industry*. New York, USA: John Wiley & Sons, Inc.

Foley, J., and Russell, JD. (1998). Mining your own Business. Retrieved on 10 July 2007 from (www.informationweek.com).

Kolko, J., and Gordon, J. (2002). Consumer Techno graphics North America Brief. *Forrester Research.*

Gebert, H., Geib, M., Kolbe, L., and Riempp, G. (2002). Towards Customer Knowledge Management: Integrating Customer Relationship Management and Knowledge Management Concepts. *In Proceedings of ICEB Conference,* Taiwan.

Geng, Y., and Du, X. (2010). The Research of Data Mining Based Sales Forecast. IEEE.

Gibbert, M., Leibold, M., and Probst, G. (2002). Five Styles of Customer Knowledge Management and How Smart Companies Put them into Actions. *European Management Journal*, 20(5):459-460.

Giudici, P. (2003). *Applied Data Mining*. Chichester, UK: Wiley.

Giudici, P. (2010). *Data Mining*. Milano, Italy: McGraw-Hill.

Goldemberg, BJ. (2003). *CRM Automation.* Upper Saddle River, USA: Prentice Hall PTR.

Golinelli, GM. (2000). L'approccio sistemico al governo d'impresa. Padova, IT: CEDAM.

Gong ZG., Muyeba, M., and Guo, JZ (2010). Business Information query expansion through semantic network. *Enterprise Information System Journal*, 4:1-22.

Gordijin, J., Akkermans, H., and Van Villet, J. (2001). Designing and Evaluating E-E Business Models. *IEE Intelligent Systems*, 16(4):11-17.

Gordini, N. (2010). Market-Driven Management: A Critical Literature Review. *Symphonya Emerging Issue in Management* (www.unimib.it/symphonya), 2.

Green, D., and Swets, JA. (1996). *Signal detection theory and psychophysics.* New York, USA: John Wiley & Sons.

Grossman, RL., Kamath, C., Kegelmeyear, P., Kumar, V., and Novnburu, R. (2001). *Data Mining for scientific and engineering applications.* London: UK: Springer-Verlag.

Guangli, N., Rowe, W., Zhang, L., Tian, Y., and Yong, S. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, 38(12):15273-15285

Hadden, J., Tiwari, A., Roy, R., and Ruta, D. (2005). Computer assisted customer churn management: State-of-the-art and future trends. *In Computers & Operations Research*, 34:2902-2917.

Hajirezaie, M., Husseini, SMM., Barfourosh, AA., et. al. (2010). Modelling and evaluating the strategic effects of improvement programs on the manufacturing performance using neural networks. *Africans Journal of Business Management*, 4(4): 414-424.

Halfpenny, P. (1979). The Analysis of Quantitative Data. *Sociological Review*, 27:799-825.

Ham, JW., and Kamber, M. (2001). *Data Mining: concepts and techniques.* San Francisco, US: Morgan Kaufman Publishers.

Hammersley, M. (1992). *Deconstructing the Qualitative-Quantitative Divide*. In What's Wrong with Ethnography? London, UK: Rutledge.

Han, Y., and Hua, Y. Mining Sequential Patterns with consideration to recency, frequency and monetary. (2011). In proceeding of: Pacific Asia Conference on Information Systems: Quality Research in Pacific Asia, Brisbane, Queensland, Australia.

Han, SH., Lu, SX., and Leung, SCH. (2012). Segmentation of telecom customers based on customer value by decision tree model. *Expert Systems with Applications*, 39(4):3964-3973.

Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining.* Cambridge, US: MIT Press.

Hanley, JA., and McNeil, BJ. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, 143:29-36.

Hartigan, J, Wong MA. (1979). Algorithm AS 136: a K-means clustering algorithm. *Journal of the Royal Statistical Society (Applied Statistics),* 28(1):100-108.

Hastie, T., Tibshirani, R., Friedman, J. (2001). *The element of statistical learning: data mining, inference, and prediction*. Berlin, Germany: Springer.

Hyndman, RJ., Koehler, AB., Ord, JK., Snyder, DS. (2008). *Forecasting with Exponential Smoothing.* Springer.

Ho-Ha, S., Min-Bae, S., and Chan-Park, S. (2002). Customer's time-variant purchase behavior and corresponding marketing strategies: an online retailer's case. *Computers and Industrial Engineering,* 43:801-20.

Hsieh, NC. (2005). Hybrid mining approach in the design of credit scoring models. *Expert Systems with Application,* 28(4):655-665.

Hsu, C., and Wallace, WA (2007). An industrial network flow information integration model for supply chain management and intelligence transportation. *Enterprise Information System,* 1(3):327-351.

Hua, BQ. (1995). Statistical prediction of the second moving average. *Statistical Education,* 2:70-73.

Huang, BQ., Kechadi, TM., Kiernan, GB., Keogh, E. and Rashid, T. (2010). Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications. *Expert Systems with Applications,* 37:3638-3646.

Huang, B., Kechadi, MT., and Buckle, B. (2012). Customer churn prediction in telecommunication. *Expert Systems with Applications,* 39(1):1414-1425.

Hui, SC., and Jha, G. (2000). Data Mining for customer service support. *Information and Management,* 38:1-13.

Hult, GTM, Ketchen DJ. (2001). Does market orientation matter? A test of relationship between positional advantage performance. *Strategic Management Journal,* 22(9):899-906.

Hung, S.Y., Yen, DC., and Wang, HY. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications,* 31:515–524.

Hwang, H., and Euiiho, ST. (2004). An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunications industry. *Expert with Applications,* 26:181-8.

Jaamour, R. (2005). Securing web services. *Information Systems Security,* 14(4):36-44.

Jaworski, BJ., Kohli, AK., and Sahay, A. (2000). Market-Driven versus Driving Markets. *Academy of Marketing Science,* 57(3):53-70.

Jenamani, M., Mohapatra, PKJ., and Ghose, S. (2003). A stochastic model of e-customer behavior. *Electronic Commerce Research and Applications,* 2:81-94.

Jick, TD. (1979). *Process and Impacts of a Merger: Individual and Organizational Perspectives.* Doctoral dissertation. New York, USA: State School of Industrial and Labour Relations, Carnell University.

Jonker, J., Piersma, N., and Van Den Poel, D. (2004). Joint optimization of customer segmentation and market policy to maximize long-term profitability. *Expert System with Applications,* 27:159-68.

K Pal, J. (2011). Usefulness Applications of data mining in extracting information from different perspectives. *Annals of Library and Information Studies*, 58:7-16.

Kahneman, D. (2011). *Thinking, Fast and Slow.* Farrar, Straus and Giroux, New York, USA.

Kanungo, T., Mount, DM., Netanyahu, NS., Piatko, CD., Silverman, R. and Wu, AY. (2002). An Efficient K-Means Clustering Algorithms: Analysis and Implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 24:881-892.

Karahoca, A., and Karahoca, D. (2011). GSM churn management by using fuzzy c-means clustering and adaptive neuro fuzzy inference system. *Expert System with Applications,* 38(3):1814-1822.

Kass, GV. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics,* 29(2):119-127.

Kavzoglu, T., and Mather, PM. (2001) The role of feature selection in artificial neural network applications. *International Journal of Remote Sensing,* 23:2919-37.

Keisler, HJ., and Robbin, JW. (1996). *Mathematical logic and computability. http://math.hosted.pl/math_2/index.html*

Keramati, A., and Ardabili, SMS. (2011). Churn analysis for an Iranian mobile operator *Telecommunications Policy,* 35(4):344-356.

Khakabi, S., and Mohammad, RG (2010). Data Mining Applications in Customer Churn Management. International Conference on Intelligent Systems, Modelling and Simulation. *In Computer Society, IEEE.*

Kim., N., Jung, KH., Kim, YS, and Lee, J. (2012). Uniformly sub sampled ensemble (USE) for churn management: *Theory and Implementation,* 39(15):11839-11845.

Kim, H., and Yoon, C. (2004). Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. *Telecommunication Policy*, 28:751-65.

Kisioglu, P, and Topcu, YI. (2011). Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey. *Expert System with Applications,* 38(6):7151-7157.

Kitayama, M., Matsubara, R., and Izui, Y. (2002). *Power engineering society winter meeting*, 1:632-34

Kohly, AK., and Jaworsky, BJ. (1990). Market Orientation: The Construct, Research Proposition and Managerial Implications. *Journal of Marketing,* 54(2):1-18.

Kotler, P., and Amstrong, G. (2010). *Principles of marketing.* Pearson Prentice Hall, New Jersey.

KPMG (2000). Knowledge Management Survey Report. London, UK: *KPMG Consulting Publications.*

Krishna, K., and Narasimha, M. (1999). Genetic K-Means Algorithm. IEEE Transactions on Systems, part b: Cybernetics, 29(3): 433-439.

Kwok, KCM., Choy, KL., Lau, HCW., and Kwok, SK., (2007). A strategic customer relationship management system: a hybrid OLAP-neural approach. *International Journal of enterprise and network management*, 1(4):350-371.

Lambin, JJ (2002). Strategic Marketing Revisited after September 11. *Symphonya Emerging Issue in Management*, 1:15.

Lambin, JJ (2007). *Market-Driven Management.* London, UK: Palgrave MacMillan.

Lambin, JJ. (2008). *Changing Market Relationship in the Internet Age.* Louvain, Belgio: Presses Universities de Louvain.

Langley, P., and Simon, HA. (1995). Application of Machine Learning and rule introduction. *Communication of the ACM*, 38(11):54-64.

Larose, D. (2010). *Data Mining Methods and Models.* New Jersey, USA: John Wiley & Sons, Ltd.

Lattin, J., Douglas, C., and Green, P. (2003). *Analyzing Multivariate Data*. Duxbury Applied Series, Hardcover.

Lau, HC., Wong, CWY., Hui, IK., and Pun, KF. (2003). Design and Implementation of an integrated knowledge system. *Knowledge-Based System*, 16:69-76.

Lee, H. (1999). Semantics of recursive relationships in entity-relationship model. *Information and Software Technology,* 41:877-86.

Lee, H., Lee, Y., Cho, H., Im, K., and Kim, YS. (2011). Mining churning behaviours and developing retention strategies based on a partial least squares (PLS) model. *Decision Support Systems,* 52(1):207-216.

Lejeune, M. (2001). Measuring the impact of data mining on churn management. In *Internet Research: Electronic Marketing Applications and Policy*, 11(5):375-387.

Lertworasirikul, S., Fang, SC., Joines, JA, and Nuttle, HLW. (2003). Fuzzy data envelopment analysis (DEA): a possibility approach. *Fuzzy Sets and Systems,* 139(2):379-394.

Lessman, S., and Vob, S. (2009). A Reference Model for Customer Centric Data Mining with Support Vector Machines. *European Journal of Operational Research*, 199(2):520-530

Liao, SH., Chu, PH., and Hsiao, PY. (2012). Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications,* 39:11303-11311.

Lin, CS., Tzeng, GH., and Chin, YC. (2011). Combined rough set theory and flow network graph to predict customer churn in credit card accounts. *Expert Systems with Applications,* 38(1):8-15.

Liu, D., and Shih, Y. (2004). Integrating AHP and data mining for product recommendation based on customer lifetime value. *Information and Management,* 42: 387-400.

Luhn, HP. (1958). A business intelligence system. *IBM Journal of Research and Development,* 2(4):314-319.

Malhatra NK. and Birks, DF. (2003). *Marketing Research. An Applied Approach.* Harlow, UK: Prentice Hall.

Manning, B., and Thorne, C. (2003). *Demand Driven.* New York, USA: McGraw Hill.

Meyer-Base, A., and Watzel, R. (1998). Transformation radial basis neural network for relevant future selection. *Pattern recognition Letters,* 19:1031-6.

Meso, P., Troutt, MD., and Rudnicka, J. (2002). A review of naturalistic decision making research with some implications for knowledge management. *Journal of Knowledge Management,* 6(1):63-73.

Migueis, VL., Van den Poel, D., Camanho, AS., and Cunha (2012). Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert systems with Applications,* 39(12):11250-11256.

Mihelis, G., Grigoroudis, E., Siskos, Y., Politis, Y., and Malandrakis, Y. (2001). Customer satisfaction measurement in the private bank sector. *European Journal of Operational Research,* 130:347-60.

Moore, J. (1996). *The Death of Competition.* New York, USA: Harper Collins.

Morgan, FW. (1978). Profitability market segmentation: identifying the heavy users of overdraft cheeking. *Journal of Business Research,* 6:99-110.

Morrisson, DG. (1969). On the interpretation of discriminant analysis. *Journal of Marketing Research,* 6:159-163

Mozer, MC., Wolniewicz, R., Grimes, DB., Johnson, E., and Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks*, 11:690-696.

Muata, K, and Bryson, O. (2004). Evaluation of decision trees: a multi criteria approach. Computers and Operational Research, 31:1933-45.

Murthy, IK. (2010). *Data Mining-Statistics Applications: A Key to Managerial Decision Making.* (www.indiastat.com).

Narver, JC., Slater, SF. and McLachlan, DL. (2004). Responsive and Proactive Market Orientation and New-Product Success. *Journal of Product Innovation Management*, 21(5):334-347.

Network World (2001) *'What the Cost of Customer Churn Means to You'*, 18(46):43.

Ng, K., and Liu, H. (2001). Customer retention via data mining. *Issue on the Application of Data Mining*, 569-90.

Nie, G., Chen, Y., Zhang, L., and Guo, Y. (2010). Credit card customer analysis based on panel data clustering. *Procedia Computer Science,* 1(1):2489-2497.

Nie, GL., Zhang, L.L., Li, X.S., and Shi, Y. (2006). *The Analysis on the Customers Churn of Charge Email based on Data mining.* In: Sixth IEEE International Conference on Data Mining - Workshops Hong Kong, China.

Ogunde, AO., Folorunso, O., Adewale, OS., Oguneye, GO, and Ajayi, AO. (2010). Towards an Agents-Based Customer Knowledge Management System. *E-Commerce Organizations International Journal on Computer Science and Engineering.*

Panagiotis, I., Soukakos, NB., and Georgoupoulos, VPE. (2007). Two interrelated framework proposed for mapping and performance measurement of customer relationship management strategies. *International Journal of Knowledge and Learning*, 3(2/3):299-315.

Parvatiyar, A., and Sheth, JN. (2001). Customer Relationship Management: Emerging practice, process, and discipline. *Journal of Economics & Social Research,* 3:1-34.

Paspallis, N., Kakousis, A., and Papadopoulus, GA. (2010). A survey of software adaptation in mobile and ubiquitous computing. *Enterprise Information System,* 4(4):355-389.

Peng, J., Lawrence, A., and Lihua, R. (2011). *Customer Knowledge Management in International Project: A Case Study.*

Petrini, M., and Pozzebon, M. (2009). Managing sustainability with the support of business intelligence: integrating socio-environmental indicators and organisational context. *The Journal of Strategic Information Systems,*18(4):178-191.

Piao, CH., Hanc, XF, and Wu H. (2010) Research on e-commerce transaction networks using multi-agent modelling and open application programming interface. *Enterprise Information System,* 4(3):329-353.

Prinzie, A., Van Den Poel, D. (2004). Investigating purchasing-sequence for financial services using Markov, MTG and MTGg models. *Journal of Operational Research,* 170:710-34.

Prinzie, A., and Van den Poel, D. (2007). Predicting home-appliance acquisition sequences: Markov/Markov for discrimination and survival analysis for modeling sequential information in NPTB models. *Decision Support Systems,* 44(1):28:45.

Quian, SL., He, JM., and Wang, CL. (2007). Telecom Customer Churn Prediction Based on Improved SVM. *Journal of Management Sciences,* 1.

Quinlan, JR. (1993). C4.5: *programs for machine learning.* San Mateo, CA: Morgan Kaufmann.

Ram, S. (1995). Intelligent database design using the unifying semantic model. *Information Systems,* 29:191-206.

Ram, S., and Khatri, V. (2003). A comprehensive framework for modelling set based business rules during conceptual database design. *Information Systems,* 30:89-118.

Rancati, E. (2010). Market-Driven-Management, Global Markets and Competitive Convergence. *In Symphonya Emerging Issue in Management*. (www.unimib.it/symphonya), 1:75-84.

Reynolds, J. (2002). *A practical guide to CRM: building more profitable customer relationships.* New York, USA: CMP Books.

Rezanková, H. (2009). Cluster analysis and categorical data. *Statistika,*3:216-232.

Risselada, H., Verhoef, PC., and Bijmolt, HA. (2010). Staying Power of Churn Prediction Models. *Journal of Interactive Marketing,* 24(3):198-208.

Rollins, M., and Halimen, A. (2005). Customer Knowledge Management Competence: Towards a Theoretical Framework. In: *Proceedings of the 38th Hawaii International Conference on System Sciences.*

Rosen, RD. (2009). Convergence Marketing. New York, USA: Wiley.

Rowley, J. (2000). Knowledge organization for a new millennium: principles and processes. *Journal of Knowledge Management,* 4(3):217-23.

Rowley, JE. (2002). Reflections on customer knowledge management. *E-business, Qualitative Market Research International Journal,* 5(4):268-280.

Rust, RT, and Zahorik, AJ. (1993). Customer satisfaction, customer retention, and market share. *Journal of Retailing,* 69:193-215.

Rygielski, C., Wang, J., and Yen, DC. (2002). Data mining techniques for customer relationship management. *Technology in Society,* 24:483-502.

Salchenberger, L.M., Cinar, EM., and Lash, NA. (1992). Neural Network: A new tool for predicting thrift failures. *Decision Sciences,* 23:899-916.

Samimi, Y., and Aghaie, A. (2011). Using logistic regression formulation to monitor heterogeneous usage rate for subscription-based services. Computers and Industrial Engineering, 60(1):89-98.

Sato, Y. (2000). Perspective on data mining from statistical viewpoints. Knowledge Discovery and Data Mining. In: *Current issues and New Applications,* 4th Pacific-Asia Conference, PAKDD, Kyoto, Japan.

Saunders, M., Lewis, P., and Thornhill, A. (2000). *Research Method for Business Students*. UK: Pearson Education Limited.

Shammari, M. (2009). *Customer Knowledge Management: People, Process and Technology.* London, UK: IGI Global.

Shapiro, BP. (1998). What the Hell Is 'Market Oriented'? *Harvard Business Review,* 66(6):119-125.

Shim, B., Choi, K, and Suh, Y. (2012). CRM strategies for a small-sized online shopping mall based on association rules and sequential patterns. *Expert System with Applications,* 39(9):7736-7742.

Shin, HW., and Sohn, SY. (2004). Segmentation of stock trading customers according to potential value. *Expert System with Application,* 27:27-33.

Shiraishi, Y., and Fukumizu, K. (2011). Statistical approaches to combining binary classifiers for multi-class classification. *Neurocomputing,* 74(5):680-688.

Skyrme, D., and Amindon, D. (1997). *Creating the Knowledge Based Business.* London, UK: Business Intelligence Ltd.

Slotnick, SA., and Sobe, MJ. (2005). Manufacturing lead-time rules: Customer retention versus tardiness costs. *European Journal of Operational Research,* 163(3):825-856.

Smith, HA., and McKeen, JD. (2005). Developments in Practice XVIII-Customer Knowledge Management: Adding Value for Our Customers. *Communications of the Association for Information System,* 16:744-755.

Smith, KA., and Gupta, JND. (2000). Neural networks in business: Techniques and applications for the operations research. *Computers and Operations Research*, 27: 1023-1044.

Stanley F., and Narver, JC. (1994). Market Orientation, Customer Value, and Superior Performance. *Business Horizons,* 22-28.

Su, CT., Hsu, HH., and Tsai, CH. (2002). Knowledge Mining from trained neural network. *Journal of Computer Information System,* 42:61-70.

Subramaniam, LV., Faruquie, TA, Ikbal, S., Godbole, S., and Mohania, MK. (2009). Business Intelligence from Voice of Customer. *IEEE International Conference on Data Engineering.*

Sun, Z., Bebis, G., and Miller, R. (2004). Object detection using feature subset selection. *Pattern Recognition*, 37(21):65-76

Swets, JA. (1979). ROC analysis applied to the evaluation of medical imaging techniques. *Investigative Radiology,*14:109-121.

Swets, JA. (1988). Measuring the accuracy of diagnostic system. *Science,* 240:1285-1293.

Tan, KY., and Kiang, MY. (1992). Managerial applications of neural networks: The case of bank failure predictions. *Management Science,* 38, 926-947.

Tan, PN., Steinbach, M., and Kumar, V. (2011). Introduction to Data Mining. Boston, USA: Pearson Education, Ltd.

Tashakkory, A., and Teddle, C. (2003). *Handbook of Mixed Methods in Social and Behavioural Research.* Thousand Oaks, USA: Sage.

Telephony Online (2002). Standing by Your Carrier. currentissue.telophonyonline.com

Todd, DJ. (1979). Mixing Qualitative and Quantitative Methods: Triangulation in Action. *Administrative Science Quarterly,* 12(24).

Thomas, BM. (2008). *Exponential Smoothing Models.* Dallas, TX: Departments of Economics. Southern Methodist University.

Trikman, P., McCormack, K., Oliveira, MPV, and Ladeira, MB (2010). The impact of business analytics on supply chain performance. *Decision Support System,* 49(3):318-327.

Triki, A., and Zouaoui, F. (2011). Customer Knowledge Management Competencies Role in the CRM Implementation Project. *Journal of Organizational Knowledge Management.*

Tsai, HH. (2012). Global Data Mining: An Empirical study of current trends, future forecasts and technology diffusions. *Expert System with Applications,* 39:8172-8181.

Tsai, CF., Chen, and MY. (2010). Variable selection by association rules for customer churn prediction of multimedia on demand. *Expert System with Applications*, 36(10):2006-2015.

Tsai, CF, and Lui, YH. (2009). Customer churn prediction by hybrid neural network. *Expert System with Applications*, 36(10):2547-12553.

Turban, E., Aronson, JE., Liang, TP., and Sharda, R. (2007). *Decision support and business intelligence systems (8th ed.).* Pearson Education.

Vallini, C., and Simoni, C. (2009). Market-Driven Management as Entrepreneurial Approach. *Symphonya. Emerging Issues in Management* (www.unimib.it/symphonya), 1.

Van den Poel, D. (2003). Predicting Mail-Order Repeat Buying. Which Variables Matter? *Review of Business and Economics,* 0(3):371-404. Katholieke Universiteit Leuven.

Van den Poel, D., and Buckinx, W. (2005). Predicting online-purchasing behavior. *European Journal of Operational Research*, 166(2):557–575.

Van den Poel, D., and Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. European Journal of Operational Research 157(1):196-217.

Van der Spek, R., and Kingma, J. (1999). Achieving successful knowledge management initiatives. In: *Reeves, J. Liberating knowledge: Business Guide*. London, UK: Caspian Publishing.

Vasu, M., and Ravi, V. (2011). A hybrid under-sampling approach for mining unbalanced datasets: application to banking and insurance. *International Journal of Data Mining, Modelling and Management,* 3(1):75-105.

Vellido, A., Lisboa, PJG., and Meehan, K. (1999). Segmentation of the online shopping market using neural network. *Expert System with Applications,*17:303-14.

Venkatadri, M., and Lokanatha, CR. (2011). A Review on Data Mining from Past to the Future. *International Journal of Computer Applications,* 15(7):19-22.

Verbeke, W., Dejaeger, K., Hur, J. and Baesens, BMD. (2012). New insights into churn prediction in the telecommunication sector: a profit driven data mining approach. *European Journal of Operational Research*, 218(1):211-229.

Verhoef, PC., and Donkers, B. (2001). Predicting customer potential via an application in the insurance industry. *Decision Support System*, 32:189-99.

Verhoef, PC., Spring, PN., Hoekstra, JC., and Leeflang PSH. (2002). The commercial use of segmentation and predictive techniques for database marketing in the Netherlands. *Decision Support Systems,* 34:471-81.

Viaene, S., Baesens, B., Van Gestel, T., Suykens, JAK., and Van den Poel, D., Vanthienen, J., De Moor, B. and Dedene, G. (2001). Knowledge discovery in a direct marketing case using least square support vector machines. *International Journal of Intelligent Systems*, 16, (9):1023-1036.

Volz, R., Handschuh, S., Staab, S., Stojanovic, L, and Stojanovic, N. (2003). Unveiling the hidden bride: deep annotation for mapping and migrating legacy data to the semantic web. *Web Semantics: Sciences, Services and Agents on the World Wide Web*, 1:187-206.

Wang, Yi-Fan, Chiang, Ding-An, Hsu, Mei-Hua, Lin, Cheng-Jung, and Lin, I-Long (2009). A recommender systems to avoid customer churn: A case study. *Expert Systems with Applications,* 36(4):8071-8075.

Wang, G., Hao, J., Ma, J., and Huang, L. (2010). A new approach to intrusion detection using Artificial Neural Networks and Fuzzy Clustering. *Expert Systems with Applications*, 37(9):6225-6232.

Wang, H-F., and Hong, W-K. (2006). Managing customer profitability in a competitive market by continuous data mining. *Industrial Marketing Management*, 35(6):715-723.

Webster, FE. (1988). The Rediscovery of the Marketing Concept. *Business Horizons,* 31:29-39.

Webster, FE. (1992). *The Changing Role of Marketing in Corporation. Journal of Marketing*, 56(4):1-17.

Webster, FE. (2002). *Market-Driven Management.* Hoboken, USA: John Wiley & Sons.

Wei, C., and Chiu, I. (2002). Turning telecommunication call details to churn prediction: a data mining approach. *Expert System with Applications,* 23:103-12.

Wei W., Wang, B., and Towsley, D. Continous-time hidden Markov Models for network performance evaluation. *Performance Evaluation,* 49:129-46.

Wilde, S. (2011). *Customer Knowledge Management. Improving Customer Relationship Through Knowledge Application.* Heidelberg, Germany: Springer.

Williams, TA, and Shoesmith, E. (2009). *Statistics for Business and Economics.* China: C&C Offset.Wireless Review (2000). *'They Love Me, They Love Me Not'*, 17(21): 8-42.

Wireless Review (2000). *'They Love Me, They Love Me Not'*, 17(21):8-42.

Wisniewski, M. (2008). *Quantitative Methods for Decision Makers.* Harlow, UK: Prentice Hall.

Wrycza, S. (2010). *Informatyka dla ekonomistow.* Podrecznik akademicki. PWE, Warszawa.

Xia, G-E., and Jin, W-D. (2008). Model of Customer Churn Prediction on Support Vector Machine. *System Engineering- Theory and Practise,* 28(1):71-77.

Xie, Y., Li, X., Ngai, E., and Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications,* 36(3):5445-5449.

Yan, L., Wolniewicz, R., and Dodier, R. (2004). Predicting customer behaviour in telecommunication. *IEE Intelligent Systems,* 19:50-8.

Yang, C. (2003). AOL: Scrambling to Halt the Exodus. *Business Week,* 3844: 62.

Yang, L., and Chiu, C. (2006). Knowledge discovery on consumer churn prediction. *Proceedings of the 10th WSEAS International Conference on Applied Mathematics*, Dallas, Texas, USA, 523-528.

Yain, AK., Murty, MN., and Flynn, PJ. (1999). Data Clustering: A Review. *ACM Computing Surveys*, 31(3):265-323.

Yin, R. (1994). *Case study research: Design and methods*. Thousand Oaks, CA: Sage Publishing.

Yu-Bao, C., Bao-sheng, L., and Xin, Q. (2011). Study on Predictive Model of Customer Churn of Mobile Telecommunication Company. *IEEE Computer Society*, 114-117.

Yu, X., Guo, S., Guo, J., and Huan, X. (2011). An extended support vector machine forecasting framework for customer churn in e-commerce. *Expert Systems with Applications,* 38(3):1425-1430.

Zabkowski, ST., and Szczesny, W. (2012). Insolvency modelling in the cellular telecommunication industry. *Expert System with Applications,* 39(8):6879-6886.

Zanjani, M.S., Rouzbehani, R., Dabbagh, H. (2008). Proposing Conceptual Model of Customer Knowledge Management: A Study of CKM Tools in British Dotcoms. *Word Academy of Science, Engineering and Technology,* 38: 303-307.

Zhang, G., Hu, MY., Patuwo, BE., and Indro, DC. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross validation analysis. *European Journal of Operational Research,* 116:16-32.

Zhang, X., Zhu, J., Xu, S., and Wan, Y. (2012) Predicting customer churn through interpersonal influence. *Knowledge-Based Systems,* 28:97-104.

Ziemba, E., and Minich, M. (2005). *Informacja I wiedzy w przedsiebiorstwie*. In: Olenski, J., Olejniczak, Z., Nowak, J. (eds.) Informatyka. Strategie I zarzadzanie wiedza, Polskie Towarzystwo Informatyczne, Katowice.