University of Milano-Bicocca

Department of Economics, Quantitative Methods and Business Strategies

Ph.D. School in Statistics and Mathematical Finance
Ph.D. in Statistics

Model Averaging using Performance and Distance measures

Supervisors:
Prof. Fulvia Mecatti
Prof. Silvia Figini

Author:
Filomena Madormo

Academic Year 2014-2015

To my daughter Sara and her guardian angel.
Thanks dad.

Abstract

In this work we introduce the problem of forecast combination using performance and distance measures for binary outcome. The thesis is focused on model averaging for parametric and non parametric approaches, with a special attention on temporal dependent and independent models. In terms of results, we combine single models using performance measures and we investigate how distance measure based on the Mahalanobis distance can lead to interesting results for model combination.
In order to assess the stability and the predictive capability of the models at hand, we employ different cross-validation techniques: Bootstrap cross-validation, 10-fold cross validation and Leave One Out cross-validation. Empirical evidence are give on a real application to predict default probabilities of Small and Medium Enterprises

Acknowledgments

I wish to thank all those who made possible the realization of my thesis. In particular I want to thank you the Prof.sa Fulvia Mecatti for making, not only the period of drafting the thesis but the entire doctoral cycle, unique experience and encouragement you gave me in these years.
I want to thank you wholeheartedly the Prof.sa Silvia Figini for being my guidance both scientific and emotional, for supporting me and for being an example of preparation and professionalism over the past four years.
Thank Mauro, Isabella and Federica for great laughs, study days and the great feasts shared.
I want to thank my partner for having been beside both in good times and in those of discomfort due to fatigue of the study and that he never lost patience and putting up with me despite everything.
I thank my mother, and particularly you dad that look at me from heaven, for the sacrifices made so that I could reach this milestone.
I thank my brother for being proud of me for the path taken.
I thank my friends Anisha, Roberta and Samantha for not abandoning me in spite of these years I'm gone for periods of study and for accepting my every outburst.
Also, thank everyone also met only briefly during this tortuous, difficult but exciting journey for all that I have both legacy culturally and humanly.

Contents

List of Figures

List of Tables

Introduction

Literature on combining forecasts is very our extensive, the common thread of most of the work is that through the combination of models is improved the accuracy of predictions. In addition, a review of the literature shows that the simplest methods of combination provide better results than more complex. In fact, in many cases it was possible to improve the performance of the models through the simple medium of forecasts obtained by single models.

The early work on the combinations of predictions are due to Reid (1968) and Bates and Granger (1969) and are considered seminal works. From one point of view this is correct because they were the first authors to develop an analytical model to combine optimally two or more predictions and to apply their models to real-world problems. Stigler (1973) describes a work of Laplace in which considers the combination of estimates of regression coefficients. Laplace studied the properties of two estimators, one being least squares and the other a kind of order statistic, and building on their joint distribution is able to obtain a formula for the combination and concluded that not being aware of the distribution of errors the combination could not be done. After Laplace, the first work that we find in the literature is to Edgerton and Kolbe (1936) in which the authors reach a combined estimate optimal minimizing the sum of squares of the differences of standard scores for the estimates. Subsequently, Horst (1938) determines the maximization of the separation in pairs between the sampling points. Halperin (1961) has developed to minimize the mean square error and Geisser (1965) presented a paper with the Bayesian equivalent of previous work Halperin through the posterior distributions. This is the true birth of the forecast combination; starting, then, since the Seventies literature has boomed following the work of Bates and Granger (1969).

This topic has experienced intense development in the field of econometrics and the combination of probability and probability distributions. A new twist to the studies was given with a series of works of the mid seventies in which it is taken into account the relative performance of the methods of combination. We refer in particular to the work of Newbold and Granger (1974)

and Makridakis et al. (1982, 1983). In particular, the work of Newbold and Granger (1974) shows how you should ignore the correlation when estimating the weights of combination by obtaining as a result that the weights that consider the correlation have poor performance. One of the most important articles in the literature combinations of forecasts is to Granger and Ramanathan (1984) that puts those techniques in the context of regression models where the variable of interest is the observed response variable and covariates are given by the prediction obtained by the single models. Further important aspect of this work is that instead consider constraints on weights claim to apply a regression without constraints to get a better fit and therefore better forecasts. Instead, the road Bayesian approach of the problem was opened by the work of Clemen and Winkler (1986) and Diebold and Pauly (1987).
Concerning applications in the course of the years the problem of the combinations of forecast was extended to the most disparate fields. Sanders (1963) dealt with the possibility to determine models in average meteorological field. A very important extension concerned the economic world through inflation, to exchange rates, to stock prices, social and technological fields and applications to football, tourism, insurance and many other sectors.
This is, in short, the historical path of the theme on the combination of forecasts oer further reading, see the work of Clemen (1989) where report a more detailed review of the early lettarutura on this issue and an annotate leterature.

We will make a presentation of the models that we have chosen to use for the problem that we have set, ie the estimated probability of default. It is models belonging to the class of pattern classifier for the problem of the classification of statistical units. As we shall see we chose two classes of models: classical models, Logistic Regression and Classification Tree, and models of survival analysis, of Cox proportional hazards regression model and Random Survival Forest.
We will clarify what it means to combine forecasts of the same response variable obtained through different models. We will see what are the methods most commonly used in the literature for the calculation of the weights of combination and what are the desirable characteristics for a good combination method.
We describe the context in which we have worked and the combination method we have developed to obtained a average model. The innovation, based on our knowledge, that will be highlighted is the particular use we have made of the performance measures most commonly used to evaluate the models. Our contribution in the field of model combination is the introduction of performance

measures as weights of forecasts obtained by individual models and not used as measures for evaluate the performance of a model. In addition, we observed the inconsistency of the performance measures, in particular the Area Under Curve, highlighting in the work of Hand (2009), we studied the introduction of distance measures such as combination weights, specifically we used the Mahalanobis distance but you can use other distance measures. The selection of these measures as a combination of weights will be justified in following the work.

Finally, we show the application of the combination method developed to a real dataset of small German SMEs and on different datasets in which we simulate the target varaibile of our interest and we present an application to churn analysis.

Chapter 1

A look at the models

1.1 Time-dependent and time-independent models with relative performance measures

In this chapter will be presented two main categories of models: time-dependent and time-independent. With regard to the first category of models, we focus our attention on the model of Cox proportional hazards model and Random Survival Forest. With regard to the time-independent model,we will discuss the Logistic Regression and Classification Trees.
We describe, also, the performance measures chosen to assess the model in terms of discriminatory power and predicted capability.

1.2 Time-dependent models: Cox proportional hazard model and Random Survival Forest

This section aims to present the survival analysis approach as we have chosen for the problem at hand.
Survival analysis studies the time required for an event of interest occurs. The most common use of this approach is medical analysis, which takes its name and in general the terminology but in recent years the scope has expanded greatly, sociology, demography, economics and so on. It is statistical methods that analyze the distribution of the time of occurrence of an event. In other words, the survival analysis allows to estimate the probability that an event will occur at a given instant in time. The following are the main differences with classical analysis:

- the data are made from a cohort of subjects;

- may be present in the sample subjects with unknown survival time, *censored subjects*;

- time takes a particular importance, the analysis is conducted on the interval of time between the entry of the subject in the study and the end of the study.

Elements that characterize the survival analysis are:

1. an event of interest, typically the death of a patient, but can be any other event such as the default of an enterprise;

2. survival time, time between the input in the study and the occurrence of the event of interest;

3. covariates, given by features of the subjects that make up the sample.

For one thing, in this context, one determines the survival time according to the difference between the instant at which the event occurs and the time of study entry. Individuals may become part of the study at different points in time and some of them may have a survival time unknown, so-called *censored subjects*. For the latter we only know that they have a longer or equal survival time than to the end time of the study but not another. They are, basically, those who have not experienced the event during the study period or subjects that emerge from the study for reasons other than the realization of the events. In particular, this subject are defined *right-censored*.
In this chapter we do a brief description of survival analysis to introduce the following models to estimate the survival function, we refer to section 3.2 for more exaustive description. Survival times are not normally distributed and it is possible to estimate the distribution with parametric methods, semi-parametric and non-parametric. Let T a random variable that represents the survival time with distribution function $P(t) = P(T \leq t)$ and the probability density function $p(t) = \frac{dP(t)}{dt}$. The survival function $S(t)$ is the complement of the cumulative distribution function $P(t)$:

$$S(t) = P(T > t) = 1 - P(t). \tag{1}$$

Equation (1) represents the probability that the random variable T is greater than a certain time t.
Another element of the analysis of survival is the hazard function that evaluates the immediate risk of the event occurs at time t conditional on survival to that

time (see Kleinbaum and Klein, 2005):

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P\left[(t \leq T < t + \Delta t) | T \geq t\right]}{\Delta(t)},$$

where $\Delta(t)$ is a small time interval.
The difference between $S(t)$ and $\lambda(t)$ is that the first is not the failure and $\lambda(t)$ refers to the failure. This two functions are linked by a very strong relationship: just to know the shape of one of the two to be able to derive the shape of the other. Specifically, we have:

$$S(t) = exp\left(\int_0^t \lambda(u)du\right),$$

$$\lambda(t) = -\left[\frac{\frac{dS(t)}{dt}}{S(t)}\right].$$

As anticipated, for the estimation of the survival function $S(t)$ is possible to use models of different nature:

- parametric models which require specific assumptions about the distribution of the survival function ;
- semi-parametric models, where are not necessary to make assumptions about the survival function and risk function.
- non-parametric models, which do not need to make any assumptions about the structure of the model.

In literature, the model most widely used to estimate the probability of survival is the Cox proportional hazards model (Cox, 1972). One of the main features of this model is that it splits the risk function into two components:

- the baseline function ($\lambda_0(t)$), which depends only on the time;
- the sum of a linear combination of the covariates in the model ($\sum_i \beta_i x_i$), which guarantees that risk estimates are non-negative.

The Cox model is defined as proportional hazards because is made the assumption that the risk ratio is constant with respect to time or, equivalently, that the risk of a subject is proportional to the risk of any other subject. In fact, the Hazard Ratio:

$$HR = \frac{\lambda(t, x^*)}{\lambda(t, x)} = \frac{\lambda_0(t) exp\left(\sum_i \beta_i x_i^*\right)}{\lambda_0(t) exp\left(\sum_i \beta_i x_i\right)}$$

does not depend on time (see Kleinbaum and Klein, 2005). Then the Cox proportional hazards defined as follows:

$$\lambda(t,x) = \lambda_0(t)exp(x'\beta), \tag{1.1}$$

tells us that the relationship between the relative risks of two different subjects is constant over time:

$$\frac{\lambda_0(t,x^*)}{\lambda_0(t,x)} = \frac{exp(x^{*\prime}\beta)}{exp(x'\beta)}.$$

The (2) in terms of the survival function becomes:

$$S(t,x) = [S_0(t)]^{exp(x'\beta)}.$$

If we consider the cumulative hazard H, i.e. the risk of the potential instantaneous occurrence of the event at time t, we can rewrite $S(t,x)$ as follows:

$$S(t,x) = exp\left[-exp(x'\beta)H_0(t)\right] = [S_0(t)]^{exp(x'\beta)}.$$

It was pointed out that the model in question assumes the proportionality of risk so you should verify compliance with this assumption (see Kleinbaum and Klein, 2005). For this purpose, is possible to use three different approaches:

1. Graphical analysis, is realized graph of the logarithm of $H(t,x)$ for each layer of the observed variable. If the curves are parallel then it is possible to say that the assumption of proportionality is respected.

2. Test of goodness of fit, there are several tests to verify the proportionality of risks. For example, you can use the test based on the Schoenfeld residuals (see Schoenfeld, 1982). The assumption is that if the proportionality is observed for a given independent variable the Schoenfeld residuals are not related with the survival time, i.e. the Schoenfeld residuals are uncorrelated in time.

3. To introduce a time dependent variable to assess the significance of the coefficient of the product between the observed variable and a function of time.

The second type of time-dependent method we have chosen to study is the Random Survival Forest which will be presented below.
The Random Forest Survival (RFS) (see Ishwaran et al., 2008) is an extension of the method, called Random Forest (RF), introduced by Breiman (2001) for *right-censored* data. The RSF is closely related to the method of Brieman.

According to Breiman all aspects of the growth of an RF must consider the outcome. In the *right-censored* data this includes the survival time and the state of censorship. Therefore, the splitting criterion used for the growth of the tree, in RSF, must include the survival time and the information of censorship. As we will see below, with the RSF creates a forest made ââup of random survival trees. Through Bootstrap independent samples, each tree is determined by randomly selecting a subset of covariates at each node and then the node is divided on the basis of a survival criterion.
Breiman has shown that the ensemble process of learning can be improved by introducing randomness in the basic learning. The randomness is introduced both through the determination of samples Bootstrap, which produce the growth of the tree, and by the fact that at each intermediate node of the tree is a randomly chosen a subset of covariates that represent the variables on which the split will occur.
The extension of the RF to survival *right-censored* data has allowed to overcome the limitations of the methods used previously. In particular, by having recourse to the RSF is no longer necessary to consider the hypothesis of proportionality of risks on which is based, for example, the Cox model. The RSF allows us to identify data structures through the analysis of the features of the sample observed. In addition, the RSF is a non-parametric method and this allows us to capture any non-linear effects of the independent variables; with parametric methods, however, such effects should be treated with transformations of various kinds or with the appropriate methods.
The RFS algorithm works as follows (see Ishwaran et al., 2008):

1. B bootstrap samples are drawn from the dataset. Bootstrap samples are independent and exclude about 40 % of the data that constitute the Out-Of-Bag (OOB) data.

2. It creates a survival tree for each Bootstrap sample. Starting from the root, each node of the tree the algorithm identifies candidate variables for the subdivision, the selected variables will be those that maximize the difference in survival between the nodes.

3. The algorithm proceeds iteratively until the tree reaches its maximum size on the basis of the constraint that the terminal nodes must have no less than $d_0 > 0$ unique deaths.

4. It calculates the Cumulative Hazard Function. It calculates the average of the Cumulative Hazard Function (CHF) of all the trees to get the

ensemble CHF.

5. Calculating the prediction error for ensemble CHF on the OOB data.

The RSF algorithm is very similar to the $CART$ algorithm which will be described in the section 1.2. The survival trees starting from the root, the highest node of the tree, determines a split of each node into two child nodes on the basis of a given survival criterion. It is necessary to evaluate the goodness of the split. A split is considered good if it is maximized the difference in survival between the child nodes. Just as in the algorithm $CART$, is evaluated the impurity of the nodes of the tree. It measures the effectiveness of the split in the separation of the data; this means, in this context, to measure the separation of the difference in survival. In essence, the algorithm selects from all covariates X and all the values of subdivision c, the variable X^* and the value c^* that maximize the difference of survival and what makes that are isolated dissimilar cases and, therefore, create sets internally homogeneous in terms of survival. The algorithm stops when no new node can be formed as it was satisfied the criterion that requires that each node contains at least $d_0 > 0$ unique deaths.
With the RSF we want to estimate the CHF. To this end, the estimator Nelson-Aalen is used:

$$\hat{H}_h(t) = \sum_{t_{l,h} \leq t} \frac{d_{l,h}}{Y_{l,h}},$$

where $d_{l,h}$ is the number of deaths at time $t_{l,h}$ and $Y_{l,h}$ is the number of people at risk at time $t_{l,h}$.
Let $H(t|x_i)$ be the CHF for subject i at time t given a set of covariates X. Because of the binary nature of the tree, x_i belong to a single terminal node. The CHF for all subjects, and then for tree survival, is:

$$H(t|x_i) = \hat{H}_h(t). \tag{1.2}$$

The identity (1.2) relates to a single tree. To obtain the ensemble CHF need to calculate the average obtained on B survival trees. It is possible to obtain both a Bootstrap and OOB estimate. The Bootstrap ensemble CHF is:

$$H_l^*(t|x_i) = \frac{1}{B} \sum_{b=1}^{B} H_b(t|x_i).$$

In this case, to obtain the estimates of CHF are used all the trees of survival and not only those on the OOB data. In the second case, on the other hand,

we have:

$$H_l^{**}(t|x_i) = \frac{\sum_{b=1}^{B} I_{i,b} H_b^*(t|x_i)}{\sum_{b=1}^{B} I_{i,b}},$$

where $I_{i,b}$ is the indicator function that takes value one if the observation is OOB ($I_{i,b} = 1$) and zero otherwise ($I_{i,b} = 0$). H_l^{**} is the average over Bootstrap samples in which the observations are OOB.

1.3 Time-independent models: Logistic regression and Classification trees

In this section we rewiev classical models for dichotomous dependent variable: Logistic Regression and Classification Tree.
First talk about the Logistic Regression model that belongs to a more general class of linear models, called Generalized Linear Models (GLM).
A generalized linear model relates a function of the expected value of the dependent variable, unpredictable nature, with the independent variables through a linear equation. In this type of models we can distinguish three components:

1. random component (ie the response variable Y). Is given by a set of random variables (Y_i), assumed independent, each with distribution that depends on a single parameter (θ_i) and that belongs to the exponential family.

2. systematic component, which describes what are the explanatory variables and their role in the model through the following linear combination: $\sum_{i=1}^{k} \beta_i x_i = \beta_1 x_1 + ... + \beta_k x_k$, where x_i are the covariates and β_i are the parameter that define the effect of a single independent variable on the dependent variable and are, typically, not know.

3. link function, describes the link between random and systematic components. This function allows that the explanatory variables (x_i) affect the expected value of the response variable (y) in the manner described by the function $\sum_{i=1}^{k} \beta_i x_i$, not necessarily linear.

Let's see, now, specifically the Logistic Regression (LR) model. This model is particularly suitable for predicting the values of a binary dependent variable, based on the independent variables that can be of any nature, qualitative or quantitative. We consider binary a variable that takes on only two characters, for example win or not win, which in general define success and failure:

$$Y = \begin{cases} 1, & \text{if outcome is success} \\ 0, & \text{otherwise} \end{cases}$$

then Y can be defined variable Bernoulli with probability of success $P(Y = 1) = \pi$ and probability of failure $P(Y = 0) = 1 - \pi$. For more details on the logistic regression see e.g. Dobson and Barnet, 2000.

Concerning single models a competitive model is the classification tree (CART). It is possible to distinguish between decision trees and classification (regression) trees according to the response variable is quantitative or qualitative. Speaking specifically of classification (regression) trees , we can say that it is a hierarchical method that produces a partition of the observations on the basis of the relationship between the dependent variable and the independent variables. On the basis of a dependent variable is found a valid partition and the algorithm iteratively chooses the classification determined by the explanatory variables that are closest to the classification made by the response variable. Each statistical unit along the shaft from the highest node, called the root, to the terminal node, which takes the name of the leaf, undergoing a classification to each intermediate node on the basis of the independent variables, generating a subdivision of the same branch. Essentially, produces a partition of the observations into groups corresponding to the leaves of the tree. Each unit is classified according to how the independent variable associated with the leaf reaches the end of the classification procedure. In general the classification algorithms seek, starting from the root of the tree, the independent variable that determines the best subdivision of the observations so that the child nodes of the root are more homogeneous with respect to the response variable.
There are various algorithms for the construction of classification trees that differ, mainly, for the splitting criterion and stop adopted. The most common algorithms are:

- Chi-square Automatic Interaction Detection (CHAID) (see Kass,1980),
- Classification and Regression Tree (CART) (see Breiman, 1984).

1.4 Performance measures

In this section we will present the measures employed to analyze and evaluate the performance of the models studied and presented in sections 1.1 and 1.2. The performance measures must meet certain criteria and it is these that

have directed our choice. First, we need measures that were able to deal the problem of binary classification in accordance with the criterion of consistency which provides a measure for the evaluation of a model must capture the look of the performance of interest (see Hand, 2009). Specifically, we are interested in measures that are related to each other and which are combinations of rate of False Positives (FP) and False Negative (FN) with the aim of reducing both. We selected measures that are associated with the ROC curve, which is described later in this section. These measures are based on a comparison between the classification of units predicted by the classifier and their classification observed. As anticipated, the objective is to reduce classification errors and thus the FP and FN. To understand what it is we introduce what is called a confusion matrix:

	1	0
1	TP	FN
0	FP	TN

Table 1.1: Confusion Matrix

so the FP are units predicted as 1 but observed as 0 FN are the units predicted as 0 but observed like 1; while the True Positive (TP) are the observations predicted and observed how 1 and True Negative (TN) are the observation predicted and observed as 0.
A grafical instrument is the Receiver Operating Characteristic curve (*ROC curve*) (see Satchell and Xia, 2006). Suppose we want to know the rating scores that tell us which debtors of a bank will survive the next year and what debtors, however, will go into default. The person who must make that decision you can use a threshold value t and on the basis of this classify debtors. If the debtors have a credit rating score lower than t will be classified as a potential defaulter while those with a rating score higher than t will be classified as those potentially not in default. This leads to four possible outcomes of the decision-making process, and then of classification, summarized in the previous Confusion Matrix (see Table 1):

1. if the rating score is lower than t and debtors are effectively in the default decision is correct (TP);

2. if the rating score is lower than t and debtors are not in the default decision is wrong, subject not in default classified as in default (FP);

3. if the rating score is higher than t and their debtors they are not really in the default decision, and then the classification, is correct (TN);

4. if, finally, the rating score is higher than t and debtors in fact result in the default decision is wrong, subjects who are in default classified as non in default (TN).

On the basis of the above, it is possible to define the following quantity named id Hit Ratio (HR) or False Positive Ratio (FPR) or, even, Sensitivity:

$$HR = \frac{H(t)}{N_D},$$

where $H(t)$ is the number of subjects in default correctly predicted in correspondence of a cut-off equal to t, and N_D is the total number of subjects in default present in the sample. In essence, HR is the proportion of correctly classified defaulter given the cut-off t. Furthermore, we define False Alarm Rate (FAR) or False Positive Rate (FPR), or, even, 1-Specificity the following quantity:

$$FAR(t) = \frac{F(t)}{N_{ND}},$$

where $F(t)$ is the number of non-defaulters misclassified as defaulters in correspondence of a cut-off equal to t and N_{ND} is the total number of non-defaulters present in the sample studied.
Now we have all elements to define the *ROC curve*. For all possible values of t in the range of the rating score can be calculated HR and FAR. The *ROC curve* is nothing but the graph between HR and FAR, such a curve is one of the most most widely used graphics tools in the literature to evaluate the ability of classification of statistical units of a model.
Measures that will be defined below are aggregate measures defined as a function of the values of the ROC curve and they are: Area Under the ROC curve (AUC or $AUROC$), H-measure, Kolmogorov-Smirnov statistic, Area Under Convex Hull of the ROC curve ($AUCH$), Gini coefficient, Minimum Error Rate (MER), Minimun Cost-Weighted Error Rate (MWL), Specificity when Sensitivity is held fixed at 95%, Sensitivity when Specificity is held fixed at 95%.
AUC is the measure most commonly used to evaluate the performance of a classification model and is given by the following amounts:

$$AUC = \int_0^1 HR(FAR)d(FAR).$$

$AUC = 0.5$ if the model has no discriminative power, which corresponds to a decision in a completely random. $AUC = 1$ if the model has perfect discriminative power; for values ââbetween 0.5 and 1, we can assume that the

analyzed model has an acceptable discriminative power, which of course will be much better as the AUC value will be close to 1. For more details see Satchell and Xia, 2006. Over the years several authors have highlighted positive and negative aspects of such a measure. An element that makes it very attractive AUC is the fact that it is not necessary to specify a threshold value t for classification. Then, from the classification rule originates a single value, which makes the AUC a simple measure that allows to easily compare the classification rules. Is an objective measure because does not necessary, for its determination, choose the parameter values and this means that different subjects can achieve the same results on the same data. On the other hand, the AUC also presents points of weakness. First, if we compare, for example, two *ROC curve* that cross is possible that the AUC of a curve result grater than the AUC relative to the other curve, but in reality it is the latter to describe the better performance. Secondly, the AUC corresponds to the average misclassification loss over a cost ratio distribution that depends on the score distributions and this implies that the AUC evaluates a classifier using a measure which depends on the same classifier. In other words, the AUC asses different classifier using different parameters. This aspect of AUC is defined *inconsistency* of this measure, and this leads us to consider more objective measures for evaluating the performance of a classification model.
AUC implicitly makes use of a weight distribution $W(c)$. Fixing a distribution $W(c)$ that captures the uncertainty for the end user on the exact values of the costs we can define:

$$L_W = \int_c L(c; T_c) W(c) dc,$$

which takes name of averaged minimum cost-weighted loss and allows us to define the AUC in terms of L_W even if with a basic choice of $W(c)$ that depends on the classifier (for more details see Hand, 2009). This leads us to the limits already analyzed of the AUC: assess different classifiers with different metrics. A measure that allows us to overcome this limit is the H-measure (Hand, 2009):

$$H = 1 - \frac{L_W}{L_W^{max}},$$

where L_W^{max} is the average MWL of the trivial classifier. The fact that L_W is normalized for L_W^{max} and subtracted to 1 leads to the following interpretation: the higher the value of H-measure, the greater will be the performance of the model.
Kolmogorov-Smirnov statistic, considers jointly the Specificity (the proportion of TN compared to the total actual negative) and Sensitivity (is making the

relation of TP compared to the total actual positive) and is given by the maximum value which takes their sum to vary the threshold t. It may also be interpreted as the maximum vertical distance between the *ROC curve* and the diagonal.
$AUCH$ is the area under the convex hull of the *ROC curve*, we notice that a *ROC curve* should not be convex. It has been shown (see Scott, 1998) that a non-convex classifier can be improved as it is possible to run the classifier on the convex hull of its ROC curve.
Gini coefficient, is closely related to the AUC and can be defined as follows:

$$Gini = 2AUC - 1.$$

The measures we are considering trying to reach a trade-off between FP and FN. To do this we must refer to the cost of missclassification; i.e. necessary to identify if it is more grave, on the basis of the analysis conducted, have FP or FN. Let c be the cost of a FP and $1-c$, the cost of a FN. The total cost is:

$$L(c;t) = 2(c\pi_0(1 - F_0(t))) + ((1-c)\pi_1 F_1(t)), where$$

$\pi_0 = \frac{TP+FP}{n}$, $\pi_1 = \frac{TP+FN}{n}$, $F_0(t) = 1 - FPR(t)$ and $F_1(t) = 1 - TPR(t)$. This quantity allows we to generalizeMER and MWL choosing a threshold that minimizes L for each value of c:

$$MWL(c) = L(c; T_c),$$

where $T_c = \underset{t}{\operatorname{argmin}} L(c;t)$. Thus MER is a special case of MWL for $c = 0.5$.
To explain Specificity when Sensitivity is held fixed at 95% and Sensitivity when Specificity is held fixed at 95% must refer to the error rate (ER):

$$ER = \frac{FP + FN}{n}.$$

The ER considers the FP as important as the FN but this may not be appropriate in some areas, such as fraud detection. A valid alternative at the ER is to establish the Specificity with respect to Sensitivity to a given level and find the Sensitivity obtained at that level. The most common choice of the level of Specificity is 95% or 99%. Similar considerations applies to Sensitivity when Specificity is held fixed at 95%.
Particularly, we are interested in comparing the AUC of the ROC curves determined for the different models considered. For this purpose it is possible

to appeal to the DeLong test (see DeLong et al., 1988). This test is used to compare two ROC curves through AUC. Specifically, we look at the p-value for the test: after fixing the dominant model, i.e. the model with AUC greater, if the p-value is greater than 0.05 then the AUC of the model is the same as the AUC of the model with which it is compared.

1.5 Time-dependent performance measures

For the time-dependent models, the Cox proportional hazards model and Random Survival Forest presented in section 1.1, we have chosen to study and apply the Harrell Harrell Concordance index, known as *C-index*. Harrell et al. introduced the concordance *C-index* in 1982. It is a measure of the separation of two survival distributions. In the last decade this index has spread as an index for evaluating the performance of prediction in the analysis of survival. In particular, we chose to use this index because it does not depend on a fixed instant of time but it is possible to determine its value at multiple points in time. The index C is connected to the *AUC* (see Heagerty and Zheng, 2005) and can be considered as a probability of missclassification. This index is a great tool to evaluate the discriminative power of the model analyzed. To calculate the *C-index* must go through the following steps (see Ishwaran et all., 2008):

1. to form all possible couples of units over the data,

2. to exclude couples whose shortest survival time is censored. It defines the total number of pairs eligible, named *"permissible"*,

3. for each admissible couples count 1 if the shortest survival time has a bad outcome expected and count 0.5 if the expected results are related. If there are eligible couples in which both subjects are dead (for which the event occurs), count 1 if the results are related, otherwise count 0.5. If there are couples eligible but the subjects are not both dead (not the event occurred for both parties) have to count 1 if the subject died have a bad result, otherwise count 0.5. The *C-index* is given by the sum of these values over all pairs permissible,

4. *C-index* is:
$$C = \frac{Concordant}{Permissible},$$
where $Concordant$ is the all couples that are can be considered concordant.

If $C = 0.5$ the model is considered to be not good in terms of discriminative power, this corresponds to making a decision with a coin toss. If $0.7 < C < 0.8$, instead, the model has a discriminative power acceptable; if $0.7 < C < 0.8$ can be assumed that the model is characterized by a good discriminative power. Finally, if $C \leq 0.9$ the model is characterized by an excellent discriminative power.

Chapter 2

Forecast combination

2.1 Introduction

Usually, the interest of the analyst is to identify the *best* forecast. When this is identified is used by the analyst while the others forecasts are discarded. However, we must consider that the forecasts discarded may contain useful information especially when the purpose is to determine the best possible forecast. In accordance with Bates and Granger (1969), we can affirm that this information may be of two types:

- forecasts based on different variables or different information,
- forecasts based on different assumption about the relationship between the variables.

We can desume that if we take into account all the forecasts and we do a combination we can come to a better and more robust forecasting performance than what we can achieve in the case of predictions generated by individual models.
To combine forecasts need to decide which forecasts to include in the combination and which weights attributed to the forecasts considered. As described by Aiolfi, Capistrán e Timmermann (2010) this decision-making process is a two-step process; the first step have to decide to be excluded because the models are characterized by poor performance. The second step, assignment of weights, is the most important and will dedicate particular attention in section 2.3.
The existing literature concerning forecasts combination suggests that even when it is possible to identify the best model may be convenient to combine forecasts since the combination could lead to an increase in the accuracy of the forecast (see Clemen, 1989).

There are many aspects that emphasize the usefulness of recourse to forecasts combination with respect to forecasts from individual models.
Suppose we want to minimize a given loss function and to have two forecasts generated by different models of the same type of response variable: $\hat{y_1}$ and $\hat{y_2}$. Suppose, also, that the forecast $\hat{y_1}$ dominates stochastically forecast $\hat{y_2}$in the sense that the prediction forecast $\hat{y_1}$ has an expected loss less than the forecast $\hat{y_2}$. This situation leads to the choice of forecast $\hat{y_1}$ over the forecast $\hat{y_2}$. However, we must consider that it is possible to determine a combination of the two forecasts that generates an expected loss less than that which is obtained only with the forecast $\hat{y_1}$.
A second reason that can lead to the forecast combination is given by the fact that the forecasts from individual models may be suffering from misspecification. Through the combination of different models can lead to a greater robustness of the forecasts against misspecification and measurement errors.
A third argument, for combination of forecasts, is that the individual forecasts may be based on different loss functions and this topic holds even if the forecasters observe the same information set. Through the inclusion of a constant in the combination equation will pick up any undesired bias, and when the latter is constant over time, there is no need to average across different forecasts.
Another aspect to use forecasts combination is that the individual forecasts may be conditioned by structural breaks. When the data window since the most recent break is short, the models may adapt more quickly and will produce the best forecasting performance. Contrariwise, the more data is available since the most recent break, slowly adapting models can be expected. The combination of forecasts from models with different degrees of adaptability will outperform forecasts from individual models, because it is difficult, in real time, to identify structural breaks.
In addition to the reasons that highlight the strength of the forecast combinations, is possible to detect several arguments against the use of them.
A problem for many combination techniques is the non-stationarities, that can lead to instabilities in the combination weights and to difficulties in deriving a set of combination weights that performs well. Another problem regards the estimation errors that contaminate the combination weights (especially when the number of forecasts is greater than the sample size).
Forecast combination problem and standard problem of constructing a single specification have common problems but they clearly show differences. Firstly, if we suppose that the individual forecasts are unbiased then also the combined forecast will be unbiased provided that the combination weights are constrained to sum to unity and an intercept is omitted and this, on condition

that the unbiasedness assumption holds for each forecast, leads to efficiency gains. It is not convenient to impose this constraint on the coefficients of a standard regression model since predictor variables can differ significantly in their units. Second, the forecasts combined need not be point forecasts but could take the form of interval. And lastly, when the individual forecasts are generated by a quantitative models whose parameters are estimated recursively there is a potential generated regressor problem which could bias estimates of the combination weights.

2.2 When to combine?

Before trying to understand when and if it is convenient to combine forecasts from different models we present a simple example that will help us understand what it means to combine forecasts. For simplicity we consider only two individual forecasts, $\hat{y_1}$ and $\hat{y_2}$, with the respectively following forecast errors:

$$e_1 = y - \hat{y_1}$$

$$e_2 = y - \hat{y_2}$$

Is assumed that the forecast errors are unbiased:

$$E\left[e_i\right] = 0, i = 1, 2$$

Variance and Covariance of forecast errors are σ_i^2, $i = 1, 2$, and $\sigma_{1,2}$. Combined forecast will be unbiased if the weights satisfy the constrain of su o unity, so we have:

$$\tilde{y} = w\hat{y_1} + (1 - w)\hat{y_2}.$$

In the same way combined forecast error is a weighted average of single forecast errors:

$$e(w) = we_1 + (1 - w)e_2.$$

So shown by Timmermann (2006), solving for the Mean Square Error (MSE):

$$w^* = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}$$

$$1 - w^* = \frac{\sigma_1^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}$$

are the optimal combination weights. The optimality of these weights is due to the fact that comply with one of the properties most desirable for the weights of combination. In fact, these weights are greater in correspondence of more accurate forecasts, this means that the variance of the prediction errors is lower, and vice versa in the event of inaccurate forecasts.

In presence of two or more forecasts of the same variable the first question to answer is whether to combine or not to combine.
It is important to distinguish between the situation where the information sets underlying the individual forecasts is observed and the situation where they are unobserved

- *information sets are unobserved*: often is justified to combine forecasts provided that the private parts of the information sets are sufficiently important. When this is satisfied it is difficult to evaluate, but it is possible to analyze and consider the correlation between forecasts or forecast errors.

- *information sets are completely observed*: the combination may be less justified in the sense that successful combination indicates misspecification of the individual models and so a better individual model should be found.

The tests on the dominance of forecasts are neither necessary nor sufficient to decide whether or not to combine, and this because in examples where forecast $\hat{y}_1$ dominates forecast $\hat{y}_2$ (lower expected loss), yet it remains optimal to combine the two forecasts.
In general, it requires a test to see if a forecast includes all information contained in another forecast. In the case of MSE loss functions there are these tests (Chong and Hanry (1986)). Point forecasts are sufficient statistics under MSE loss and a test of pair-wise encompassing can be based on the regression

$$y_{t+h} = \beta_0 + \beta_1 \hat{y}_{t+h,t,1} + \beta_2 \hat{y}_{t+h,t,2} + e_{t+h,t}, \quad t = 1, 2, \ldots, T-h.$$

The rule of the test provides that, when the parameter restriction $(\beta_0\ \beta_1\ \beta_2) = (0\ 1\ 0)$ holds, forecast 1 encompasses forecast 2 and when, instead, we have $(\beta_0\ \beta_1\ \beta_2) = (0\ 0\ 1)$ then forecast 2 encompasses forecast 1.
It is possible that omitting one forecast, in small samples, can obtain to better out-of-sample forecasts, even if the coefficient on the ignored forecast differs from zero.
More generally, there is a test based on the hypothesis that $\beta_2 = \ldots = \beta_N$,

where forecast 1 encompasses all other models

$$y_{t+h} - \hat{y}_{t+h,t,1} = \beta_0 + \sum_{i=2}^{N} \beta_i \hat{y}_{t+h,t,i} + e_{t+h,t}.$$

When the data is not much informative and it is not possible to detect a dominant model, it makes sense to combine forecasts. The risk of not choosing the best model diminishes when more methods are considered and forecasts are averaged and this is proven by showing that the forecasting performance of a combination strategy enhances as a function of the number of models involved in the combination.
A model selection criteria, such as the SIC, to choose which forecasts to combine has been proposed by Swanson and Zeng (2001). The combination chosen by the criteria SIC seems to provide the best overall performance and rarely is dominated by other methods.
After establishing whether to combine or not, there are several ways to estimate the combination weights, $\hat{\boldsymbol{\omega}}_{t+h,t}$.

2.3 How to determine the combination weights

Existing results on the properties and performance of forecast combinations have been derived in the context of a least squares loss function.
To simplify matters we assume that the loss function only depends on the forecast error from the combination, $e^c_{t+h,t} = y_{t+h} - g(\hat{\mathbf{y}}_{t+h,t}; \boldsymbol{\omega}_{t+h,t})$, i.e. $L = L(e_{t+h})$.
The parameters of the optimal combination, $\boldsymbol{\omega}^*_{t+h,t} \in \mathcal{W}_t$, solve the problem

$$\boldsymbol{\omega}^*{}_{t+h,t} = \arg \min_{\boldsymbol{\omega}_{t+h,t} \in \mathcal{W}_t} E\left[L\big(e^c_{t+h,t}(\boldsymbol{\omega}_{t+h,t})\big)|\hat{\mathbf{y}}_{t+h,t}\right].$$

The combination weight can be found, as shown by Elliott and Timmermann (2004), as the solution to the Taylor series expansion of the loss function around the bias of the forecast error $\mu_{e_{t+h,t}} = E\left[e_{t+h,t}|\mathcal{F}_t\right]$

$$\begin{aligned}\boldsymbol{\omega}^*{}_{t+h,t} = \arg \min_{\boldsymbol{\omega}_{t+h,t} \in \mathcal{W}_t} \Bigg\{ & L(\mu_{e_{t+h,t}}) + \frac{1}{2} L''_{\mu_e} E\left[(e_{t+h,t} - \mu_{e_{t+h,t}})^2 | \mathcal{F}_t\right] \\ & + \sum_{m=3}^{\infty} L^m_{\mu_e} \sum_{i=0}^{m} \frac{1}{i!(m-i)!} E\left[e^{m-i}_{t+h,t} \mu^i_{e_{t+h,t}} | \mathcal{F}_t\right] \Bigg\},\end{aligned}$$

where $L^k_{\mu_e} \equiv \partial^k L(e_{t+h,t}) / \partial^k \omega|_{e_{t+h,t} = \mu_{e_{t+h,t}}}$.
This expansion suggests that the collection of individual forecast $\hat{\mathbf{y}}_{t+h,t}$ is useful

in as far as it can predict any of the conditional moments of the forecast error distribution.
The objective function underlying a combination problem, oftentimes, is mean squared error (MSE) loss

$$L(y_{t+h}, \hat{y}_{t+h,t}) = \theta(y_{t+h} - \hat{y}_{t+h,t})^2, \quad \theta > 0.$$

In the combination problem are involved two levels of aggregation:

- First step: summarizes individual forecasters' private information to produce point forecasts. The difference to the standard forecasting problem is that the input variable are forecasts from other models, this can leads to the bias the estimated combination weights and it could explain, in part, why combinations based on estimated weights often do not perform well.

- Second step: aggregates the vector of point forecasts to the consensus measure. This step is likely to lead to more parsimonious forecasting models when compared to a forecast based on full set of individual forecasts.

  Information is lost in both steps.

In general, we can expect that the information aggregation to increase the bias in the forecast but also reduce the variance of the forecast error.

In the literature they are found in the majority of cases, patterns of combination of the linear type. In essence, if we assume as a loss function the MSE proceed to a combination of linear forecast models. However, you can also encounter cases where build combinations of non-linear type or time-varying combination methods (see Timmermann, 2006).
We can also observe optimal combinations in case of asymmetric loss functions. However, it should be noted that the work in the literature is the result, also in this context, use theme as a loss function to minimize. In fact, it was shown that the standard properties of an optimal prediction in the case of MSE as a loss function (unbiased, absence of serial correlation between the forecast errors and increase of the error variance of the time horizon to grow) fall in the case of asymmetric loss functions.
Bates and Granger (1969) claim that it would be desirable to have more weight to predictions with less error. Also, they tell us that a good method to determine the weights of combination should meet the following properties:

1. the average weight must tend to the optimum increases forecasts,

2. the weights should be able to quickly adapt to new values in case they should occur changes in the success of one of the predictions,

3. weights should vary slightly compared to the optimal value.

One of the most important observations, in our view, is one that says that the methods for the determination of the weights should be moderately simple. One the basis of this consideration in their work, Bates and Granger (1969) examine five methods to calculate the combination weights. They consider two forecast and determine a forecast combination at time T, C_T and compute a weights k_T the satisfy the condition to sum to one. So, they have weight k_T for forecast one and weight $1-k_T$ for forecast two. This weights are determine on the basis of the forecasts error until to previous time instant T: $e_{1,1},,e_{1,T-1}$; $e_{2,1},,e_{2,T-1}$. We notice that for the first weight k_1 the authors choose the value 0.5 for all methods studied. For example the first method presented establishes the following weight:

$$w_T = \frac{E_2}{E_1 + E_2},$$

where $E_1 = \sum_{t=T_\tau}^{T-1}$ and the the same is true for E_2. For other methods we please refer Bates and Granger (1969).

In the context of forecast combination play an important role those who are called equal weights. This weights are considered by Timmerman (2006) a "natural benchmark" for the problem of forecast combination. This weights are optimal in the case in which the variances of individual forecast errors are identical and the pair-waist correlation are exactly alike. They are defined as follows:

$$\tilde{y}^{ew} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i,$$

where n is the number of forecasts want to combine.This represent the common baseline in the combination problem. In essence, it translates in determining an average of the individual predictions. The great strength of this combination scheme is due to the fact that simple methods like this are hard to beat in practice.

2.3.1 An interesting type of weights: relative performance weights

A major problem with which have to deal when you have to determine the weights of combination is the estimation of the covariance matrix of errors Σ_e. To overcome this aspect Bates and Granger (1969) and Newbold and Granger

(1974) have been proposed combination weights that disregard the correlation between the individual forecasts. Winkler and Makridakis (1983) assert that this means to treat the covariance matrix Σ_e as a diagonal matrix.
In literature we find the work of Stock and Watson (2001) that present the combination of weights that ignore the correlation between errors but are based on the relative performance of the MSE of the individual models. Let $MSE_{t+h,t,i} = \frac{1}{\tau}\sum_{t-\tau}^{t} e_{t-\tau,(t-\tau)-h,i}$ the MSE of ith models at time t. Then

$$\hat{y}^{c}_{t+h,t} = \sum_{i=1}^{N} \hat{w}_{t+h,t,i}\hat{y}_{t+h,t,i}$$

where

$$\hat{w}_{t+h,t,i} = \frac{(1/MSE^{k}_{t+h,t,i})}{\sum_{j=1}^{N}(1/MSE^{k}_{t+h,t,i})} \tag{2.1}$$

If it is assumed $k = 0$ this means giving equal weights to individual forecasts. The authors say that this weights effectively not consider the correlation between the forecast errors and perform well in case of large samples.
In letterature, many authors agree with Newbold and Granger (1974) that say the combination strategy that ignore the correlation between the individual forecast errors perform better than methods that consider this relationship. However, even today, after many years of study on the combination methods, this seems to be an open question.

2.4 The estimation problem

The combination weights are estimated on the basis of past data. Over the years, several methods have been defined to estimate the combination weights and this is due primarily to the absence of a single optimal method for their determination. Moreover, the role that is attributed to the forecast errors constitutes an element of difference. The only element that seem to have in common all methods of combination is that the starting point is the equal-weight combination described in section 1.3. One of the major disadvantages of the combined forecasts is that they introduce estimation errors related to the parameters when the weights are, in fact, it is estimated (see Yang, 2004). In most cases the problem of the estimate of the weights is addressed through last square estimation. Generally, assumes a linear model in the weights and the combination weights are estimated through Ordinary Last Square (OLS). Granger and Ramanathan (1984) consider three regression:

1. $y_{t+h} = w_{oh} + w_{h}^{'}\hat{y}_{t+h,t} + \epsilon_{t+h}$

2. $y_{t+h} = w_h^{'}\hat{y}_{t+h,t} + \epsilon_{t+h}$

3. $y_{t+h} = w_h^{'}\hat{y}_{t+h,t} + \epsilon_{t+h}$, s.t. $w_h^{'}\iota = 1$.

The regression 1. and 2. can be estimated by ordinary last square, the difference the difference between both is that the second regression omits the intercept, while the regression 3. requires the constrained last square. In general the regression does not impose that the forecast by single models are unbiased because the bias can be adjusted by intercept terms. The regression three, instead, require the unbiased forecasts. This can overcome by imposing that the weights sum to one but this constraint can leads to efficiency losses. It is necessary to signal that there are in the literature different versions of basic last square regression.
There are also works that rely estimating combination weights directly on the loss function but excluding quadratic loss function. This is the case of moment estimator. The combination weights can be obtained as an moment estimator based on sample. For more details see Elliot and Timmermann (2004).
Those just described are the methods most commonly used for the estimation of the weights of combination, also suggest the existence of non-parametric combination schemes and other methods for which reference should be made to the work of Timmermann (2006).

In our opinion the starting point, when thinking whether to build or not predictions combined, is the statement of Clemen (1989): "Combining forecasts has been shown to be practical, economical and useful". In fact, the forecast combination allows us not only to use a broader range of information than what we allow the individual models but makes possible to overcome, for example, the misspecification of single models.
Appears evident from the work in the literature that more simple are methods through which combine forecasts from individual models best are their performance. For this reason the combination methods with equal weights is, in most cases, the best or at least equal to the combination schemes much more complex. Then when you want to combine forecasts from single models it is always convenient to compare the methods selected with the simple and economical technique of equal weights.
Moreover, in this chapter we have seen what are the methods most commonly used to combine the forecasts of the same dependent variable generated based on different single models. The principal method to determine the combination weights is to assume a linear model in the weights and the combination weights are estimated through OLS.

From our point of view are of particular interest those who are called relative performance weights because, as we saw in the previous section, let us overcome the problem of forecast errors due to the estimated weight of combination. We notice that in the cases in which the combination weights are imposed instead be estimated we can overcome the problem of estimation error.
The following chapter shows our methodological suggestion to combine forecasts from single models, particularly from models belonging to two different class of models, time-dependent and time-independent, for the estimation of binary outcome.

Chapter 3

The proposal for combination forecasts

3.1 The context

The goal we are placed in this work is to identify the best model for predicting the binary response variable, in the specific case we consider the Probability of Default (PD). Before describing in detail the method that we have developed to improve forecast, we describe the framework in which we operate. The general context in which it occurs our empirical work is to credit risk.
For credit risk refers to the risk that an unexpected change in the creditworthiness of a person, towards whom there is credit exposure, causes a corresponding unexpected change in the value of the same credit position. It is not the only possibility that the counterparty becomes insolvent, in fact even the deterioration of the creditworthiness is considered credit risk. An aspect of great relevance is given by the fact that the variation of the credit position must be of unexpected nature. There are several types of credit risk, we distinguish five major categories:

1. risk of insolvency, is the possibility that an entrusted counterparty becomes insolvent. The loss made by the creditor is given, of course, by the difference between the value of the loan and how much is actually recovered;

2. risk of migration, is the risk of deterioration of the creditworthiness of a counterparty;

3. risk of recovery, is the possibility that the rate of recovery relative to exposures to counterparties defaulting is lower than originally estimated;

4. risk of exposure, is the risk that the size of the exposure against a counterpart increases unexpectedly;

5. spread risk, is the risk that the same credit rating increases the spread.

Among the different types just presented in short we focus our attention on the risk of insolvency. Specifically, as declared earlier, it is our interest to identify the best possible model to determine the probability of default (PD). PD is the counterparty risk of the recipient of credit exposure. In other words, it is the measurement of the creditworthiness by estimating the probability of insolvency. As regards the choice of the technique to be used for the estimates of the probability of a subject to be insolvent, is prossibile say that there is not a better technique in absolute but certainly we must choose the one which enables to best use the information that we have to available. To estimate the probability of default is necessary to understand when a credit position should be considered in default. Generally, a loan is considered in default when it is no longer possible or the will of the borrower to meet its financial obligations. To be more precise, it is necessary to say that there is no single definition of default but the banks agree that a credit can be defined in default when they go on suffering (no payment is made by the debtor, the creditor advances a proposal for restructuring the debt, etc.). According to the rating agency Standard Poor's default occurs when the debtor is facing an irreversible crisis which leads to exclude guarantees related to the credit position and as to suggest that it will generate the loss of part, or the whole, the capital loaned.

After clarifying which is the final object of our study, it is necessary to emphasize that we do not operate in the classic context of forecasting of PD but we place ourselves in the framework of the cohort studies. In other words, as is clear from the choice of models declared in Chapter 1, we are in the typical context of the survival analysis.
To determine the PD we have been developed systems of crew scoringn with the aim to calculate the probability that a subject that require a loan may go into default. Over the years the number of statistical models to address this need has grown more and more but the model, even today, is the most widely used Logistic Regression (LR) for binary responce variable (see Stepanova and Thomas, 2002). In recent years as part of the credit scoring was introduced also the survival analysis, branch of statistics that deals with the analysis of lifetime data. The latter, in general, is used in the medical field for studies on the administration of drugs or therapies on patients affected by a given pathology but in recent years is applied in an increasing number of sectors

(see e.g. Hosmer et al, 2008). In the context of the credit risk of the event of interest is the default.The application of the survival analysis to credit risk is due to Narain (1992) developed, then, in the future by Thomas et al. (1999). Compared to classical models used for predicting the PD, the models in the survival analysis consent to work on samples containing censored data. Censorship most common finding is called censorship right and tells us that a censored subject has not experienced the event of interest, in this case the default, during the study period; we clarify the concept better later. In the case of credit risk it is very easy to find in a sample many subjects censored because the majority of these is not in default.
Another advantage of the survival analysis is that not only allows us to study the occurrence or not of an event of interest but to study the time necessary to occur the same. In fact, the target variable is the time to the occurrence of an event, called survival time. In other words, the survival analysis refers to those statistical methods that study the distribution of time of an event occurring. The survival time is given by the difference between the time in which an event occurs and the time of entry of the subject in the study. Credit risk is closely related to the survival analysis since the latter by definition is the analysis of time-to-failure data, where failure to mean, in this case, the default. The advantages in applying models of survival analysis to the problem of credit risk are manifold. In particular, ability to insert different types of covariates in the model, flexibility in parametrizing the default intensity and expand the class of models applicable in this context
As anticipated, with survival analysis is studied a cohort of subjects for a certain period of time, this can result that its numerosity various time precisely the occurrence or nonoccurrence of the event of interest. In fact, one of the assumptions of this analysis is that the observations entering and leaving the study randomly. The main features of this type of study are described below:

1. starting point, given by the occurrence of an event well identified over time;
2. diagnostic and admission criteria: they must be clearly made explicit;
3. methods of subject recruitment\statistical units, all subjects examined by observers should be included in the study and the number and characteristics of the excluded should be carefully considered and specified;
4. control of all subjects for a certain minimum period;
5. end point, relevant event which has to be defined in advance in an objective and reliable.

The time interval between the initial and the terminal event is represented by a random variable T ($T \geq 0$) defined as *survival time*. Survival times can not be normally distributed and their distribution is described by survival function described in section 2.2 by equation (2.1). A very important aspect of the censored data is that censorship is not considered informative, this assumption means that the censored subjects have the same risk of non-censored. For this reason the variable T is not fully observable. All this means that when the data are censored very numerous than those in suffering statistical methods used may not be reliable. A widespread problem in the survival analysis is the presence of censored data, as specified in Section 2.2 we refer to right censored data. The Figure 3.1 helps to understand the characteristics of a cohort of subjects typical of survival analysis.

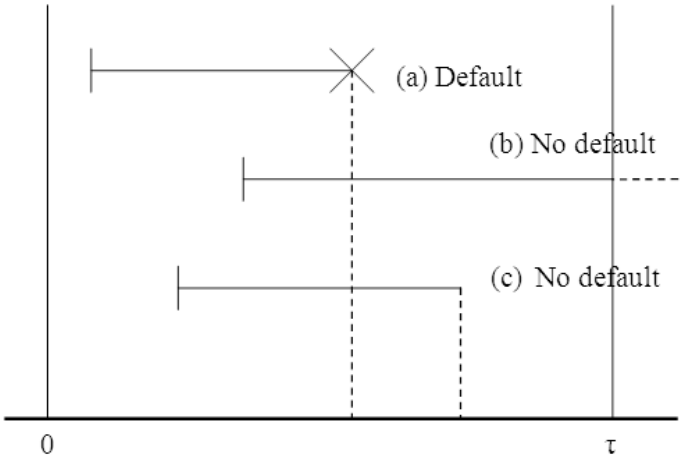

Figure 3.1: Example dataset with right censored data

On the x axis shows the time of study which goes from 0 to τ. The fact of considering a limited time interval can detrminare the presence of censored observations right. Censorship can be due to two main reasons: did not occur the event of interest during the period of the study or the subject comes from the study for reasons other than the occurrence of the event of interest. for the subject a event occurs, the default, the studio while subjects b and c represent two censored data. For the subject b is not the event occurs by the end of the study and to him we can only say that his survival time is greater than "tau", end time of the study. The third subject, c, is a subject that comes out from the study, during the study time, for different reasons after the event, for example, lost the requirements to be part of the study (see e.g. Cao et al.

2009).
In this section we have explained why we chose to use for predicting the PD also models in the survival analysis and as we shall see in Chapter 5 the latter are competitive or better than the classical models, in particular the LR. The next section, however, will present the method developed by us to combine two classic models in the analysis for the prediction of PD, LR and Classification Tree (CT), and two models of survival analysis, Cox proportional hazards regression model (COX) and Random Forest Survival (RSF), in order to improve prediction of PD. We will see not only how we built the average modelbut also as how we evaluated the performance of individual models than average models.

3.2 Combination scheme

This section will present the method studied by us to combine models of different nature in order to improve the prediction of the PD. Specifically, we will see that we determined averaged models using weights as performance measures and distance measures, to our knowledge, is an innovation in the literature of forecast combination.
The goal of our work is to be able to correctly classify statistic units available to us with the classification models presented in Chapter 1. Through combiation of these models we want to get a more accurate classification. Accuracy is the proportion of subjects correctly classified and the sum of True Positives (TP) and True Negative (TN) on total of statistical units (N) (see Table 1.1) :

$$Accuracy = \frac{TP + TN}{N}$$

This is the reason why we have chosen to use performance measures related to the Receiver Operating Characteristic (ROC) curve, as explained in Chapter 1. In fact, the area that is below the ROC curve, called Area Under the Curve (AUC), is the performance measure most widely used for the evaluation of classification models and is a measure of diagnostic accuracy.

Our problem of combination is divided into two parts:

1. combination of only time-dependent models, that allow us to have estimates of the Probability of Default (PD) for each time step of our study time,

2. combination of all four models, regardless of the fact that we are time-

dependent models or less, during this case we obtain the PD estimate of the entire time period considered.

In both cases the procedure and the combination pattern adopted are the same, the difference lies only in the fact that in the case of time-dependent models we obtain estimates for each instant of time of our interest.
The initial idea was to compare the time-dependent models with those time-independent in terms of AUC to establish the single model that performs better and then use the same measure as weight for the combination. The choice of using a performance measure as weight arises from the consideration that the literature suggests that use weights without any constraint leads to combination methods more accurate than those that pose conditions on the characteristics of the weights such as, for example, the constraint to sum to one. During studies there has arisen doubt about using AUC, largely as a measure for the evaluation of the performance of a model, considering the fact that it is a measure defined "incoherent". Observation on incoherence of the AUC as a measure of comparison of classification models was made first by Hand (2009). What is Hand says that the assessment made by AUC in comparing models does not match the one made by other measures based on misclassification costs. What we want to understand Hand is that given two models m_1 and m_2 say $AUC_{m_1} > AUC_{m_2}$ does not allow us to say anything about the misclassification costs of the same models. It seems to be that the measuring instrument, the AUC, is different depending on what you want to measure. Citing the same author we can say that "evaluates different classifiers using different metrics. It is as if one measured person A's height using a ruler calibrated in inches and person B's using one calibrated in centimetres". To demonstrate this incoherence, Hand determine a linear relationship between the expected minimum loss and the AUC. This means that if two models have the same AUC does not mean that they have the same type of expected minimum loss if you have used different distributions over proportions costs were the same for both models. For more details we refer the reader to the work of Hand (2009). After the work of Hand we can find some papers in the literature showing that the AUC is to be a coherent measure, unlike what it says Hand, if, for example, not using an optimum threshold or if you do not take into account the classification errors. In our case we say that for us the AUC is a incoherent measure because we are interested in their classification errors because, as we shall see shortly, the evaluation of combined models will be made right on the basis of statistical units correctly classified.
Inspired by the work of Hand, to overcome the problem of incoherence of the

AUC we decided to use performance measures identified by him and which are a "coherent alternative", as defined by same author, the AUC for the evaluation of performance of classification models. So we proceed with determining all the measures presented in Chapter 1 and summarized below: H-measure, Gini index, AUC, AUCH, Kolmogorov-Smirnov statistic, MER, MWL, Specificity when Sensitivity is fixed to 95%, Sensitivity when Specificity is fide to 95%. In this case, our main goal is not to use these measures to assess and compare the performance of different models considered rather use them as weights in the combination scheme ensuring, in this way, that the best model has greater weight and model worst going to weight less at the time of the combination thus eliminating the specter of incoherence. Since the measures mentioned above do not guarantee compliance with this condition, in particular the AUC, for their inconsistency we decided to calculate and use as weight distance measure. Specifically, we have focused our attention on the Mahalanobis distance for the reasons which will be indicated below without this represents a limit to the use of other distance measures.
The Mahalanobis distance is defined as the Euclidean distance between the statistical units after pulling the main components and have them standardized. Given two statistical units x_i and x_j is given by

$$D_{i,j}^{M} = \left[(x_i - x_j)' S^{-1} (x_i - x_j)\right]^{\frac{1}{2}}, \qquad i, j = 1, \cdots, n$$

where S^{-1} is the inverse of the covariance matrix and n is the total number of units. It is also known as generalized Mahalanobis distance. If the variables are uncorrelated, the matrix of variances and covariances is diagonal and it is reduced to the distance between the units after standardized variables. The Mahalanobis distance is not affected by correlations between variables, but tends to mitigate the differences between groups, if any. It is a statistical measure of the distance between the units, which is calculated net of the correlation between the variables, therefore it allows to eliminate the correlation between the variables. The choice of this distance measure is due to the following reflections. First, in multivariate statistics this measure is used to capture the similarity between two objects and we are interested to measure the similarity that exists between the response variable observed and forecast the same made by the different individual models considered. Obviously, we wish that the expected value is as close as possible to the value actually observed, so we can only assume that the model is reliable. Second, we wanted a measure that could be calculated in the same way for all the models under study view of the different nature of these models. Last but not least, we wanted a simple

measure to calculate to meet one of the main characteristics of the methods of combination.
Clarified the weights that we chose we move now to the description of the combination method adopted. First we specify that the estimates used were made on the sample for testing following the application of three different methods of cross-validation. The goal we wanted to achieve through various methods of cross-validation is the assessment of the stability of the models and then determine whether a given method of cross-validation postesse affect results. Selzionate techniques for this purpose are: Bootstrap cross-validation (Bootcv), K-fold cross-validation (K-foldcv) and Leave-one-out cross-validation (LOOcv). Considering the Bootstrap approach, we take the entire original data set; we draw a sufficiently large number B (B=200) of bootstrap samples and for each boot- strap sample we run our model. Then, we compute the Bootstrap error. We notice that it is possible to realize this validation technique by performing either a with replacement or without replacement resampling, according to the nature of the original sample data. The second method we used to assess the models is the K-fold cross vali- dation that divides the data into K subsets of equal size that are called folds. For each repetition, k1 subsamples are used to training the models and ki, with i = 1, ..., K , is the validation sample. For instance, having K=10 folds, we first use data from folds 1 to 9 as training sets and leave the 10th as the validation set. Secondly folds 1 to 8 plus 10 are used as training sets and fold 9 is left as validation set; and so on iteratively for each of the available K folds. The numbers more frequent of folds adopted are 5, 10 and 20; we choose to adopt a number of fold equal to 10. The Leave- One- Out cross validation (LOOcv) is a particular type of K-fold cross validation. In fact, it is a K- fold cross validation with $K = n$. For more details on these techniques of cross-validation we refer the reader to the work of Mogensen, Ishwaran and Gerds (2012).
First, we only worked on models in order to have time dependent PD estimates at each instant of time of the study, in order to know not only if the event occurs of interest but also when.

The scheme is organized according to the following four main steps:

1. estimation of the response variable, Y, on the basis of the test samples as a result of the techniques of cross-validation briefly described,

2. calculation of performance measures and the Mahalanobis distance for individual models at each instant of the time period of study duration (6 years),

3. use of measures calculated in step 2 as weights to create averaging models,

4. evaluation of the performance of individual models and averaging models and compare them to determine the best model, the assessment is made on the basis of the units correctly classified.

We specify that in the context of the combination of classification models we can distinguish two major strategies of combination calls fusion and selection. The difference of the two strategies is the fact that in the context of the fusion each model that is part of the ensemble have knowledge of the whole feature space, while in selection every model knows a part of feature space. In the strategies of fusion fall combinations determined as the medium while in the branch of the selection fall the strategies in which select one classifier to label the input x. This speech is to point out that our combination scheme falls within the context of the fusion strategies because we determine weighted average models.
In the case of time-dependent models we calculated all performance measures mentioned, but not only. In fact, for these models it is possible to calculate an additional measure of performance which by its nature is time-dependent and that takes the name of Harrell Concordance index, in a short C-index, described in section 1.4. While this index is inherently time-dependent, so we can easily determine the value for each point in time, other performance measures are not time-dependent, thus providing a unique value to the whole horizon. One new feature of this work is the fact that we were able to determine all the measures, including the Mahalanobis distance, for each point in time and not only on the entire time horizon.
The weights used, therefore, to create models average for time-dependent approach are: C-index, AUC, H-measure, Gini index, AUCH, Kolmogorov-Smirnov statistic, MER, MWL, Specificity when Sensitivity is fixed to 95%, Sensitivity when Specificity is fide to 95%, Mahalanobis, $\frac{1}{C}$, $\frac{1}{AUC}$, $\frac{1}{Mahalanobis}$. Moreover, we have created the following weights from vector Mahalanobis distance:

$$w_{RP} = \frac{1}{Mahalanobis}\left(\sum_{i=1}^{n}(Mahalanobis_i)^{-1}\right)^{-1} \tag{3.1}$$

Weight described by equations 3.1 is particularly important as they fall into the category of so-called relative performance weigths described in section 2.4.1 and meet one of the most important properties of the optimal combination weights, that sum to one. Our approach prefer to use weights on which are not subject to restrictions of any kind because, as the literature suggests and as described in Chapter 2, weights without constraints provide better perfor-

mance. However, for completeness of the work we have also identified weights, starting from the Mahalanobis distance, they were able to meet this important constraint. In Chapter 4 we will observe and compare the performance of combinations obtained using these particular weights.
As regards, instead, the combination scheme adopted for combining both time-dependent models is time-independent models is summarized in Figure 3.3. In substance, as already specified, the only difference that one has in this case compared to the previous is due to the fact that the performance measures are calculated on the entire time horizon and not for every single moment, since the nature of the models time-independent prevents us from obtaining such a result.

The weights used for the combinations are the same as those described for the approach time-dependent.
At this point we can see how we got combinations. First it should be noted that in addition to having calculated weighted average models, we included in our study even the simple average model without weights. This is because the literature suggests, as discussed in Chapter 2, the simple average of the forecasts obtained from the individual models is the starting point for the combination models and often should adopt it because performs better or on a par with much more complex combinations. The average models were determined in the following way:

$$C_{w_{AUC}} = \frac{f_{Cox}AUC_{Cox} + f_{RSF}AUC_{RSF} + f_{LR}AUC_{LR} + f_{CT}AUC_{CT}}{AUC_{Cox} + AUC_{RSF} + AUC_{LR} + AUC_{CT}} \quad (3.2)$$

where C is the combination model, f is the forecast by single model and AUC is the performance measure used as combinato weights. For example, we showed the average model weighed to the AUC but likewise were determined other combinations where instead AUC insert other measures of performance or the Mahalanobis distance or, even, weights indicated in equations 3.1 and 3.2. We do not report the equations of all combinations in this chapter for economies of space but all combianzioni will be present in the appendix. From the equation 3.3 we get a new forecast of the variable response of our interest, the PD, as a weighted average of the predictions obtained from single models. At this point we are going to evaluate the performance of all models considered in our study, single and combained, based on the units correctly classified. Therefore we determine the confusion matrix presented in Chapter 1 Table 1.1. Observing the confusion matrix we can see the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), all quantities defined in Chapter 1. In this case the best model is the one that will have a higher percentage of units

are correctly classified and therefore true positives and true negatives. As we will see in chapter 4 matrices of confusion we were determiante at different cut-off to determine whether the choice of a threshold particoalre could affect the result.

In this chapter we have made clear, first of all, the context in which we work, and therefore the concept of credit risk, and what was the object of our interest, namely Probability of Default (PD). Second, we observed that the typical patterns of the survival analysis have a high utility also in the context of credit risk as they allow us not only to estimate the PD but also to determine when you will experience the event of interest, namely the default, through the temporal approach. The most interesting part of this chapter is the section 3.3 which describes the scheme of combination studied. In particular, we see how we used measures of performance and the Mahalanobis distance as weights to determine weighted average models. Finally we saw that after the combinations we evaluated and compared the models on the basis of observations correctly classified through, therefore, the matrix of confusion.

Chapter 4

The empirical results on real and simulated datasets

4.1 Data description

Our empirical exercise is based on annual 1996–2004 data from a major rating agency for Small and Medium Enterprises (SMEs) in Germany, for about 750 firms (Creditform).

When handling bankruptcy data it is natural to label one of the categories as success (healthy) or failure (default) and to assign them value 0 and 1 respectively. Our data set consists of a binary response variable Solvency (Y_{it}), a Duration variable and a set of explanatory variables: $X_1, ..., X_k$ displayed in Table 1.

Given our available dataset, we computed this set of 9 financial ratios:

- Capital tied up: this ratio evaluates the turnover of short term debts with respect to sales.

- Supplier target days: it is a temporal measure of financial sustainability expressed in days that considers all short and medium term debts as well as other payables.

- Liabilities ratio: it is a measure of a companyâs financial leverage calcu lated by dividing a gross measure of longterm debt by the firmâs assets; also it highlights what debt proportion the company is using to finance its assets.

- Outside capital structure: this ratio evaluates a firmâs capability to re-ceive forms of financing other than banksâ loans.

- Cost income ratio: the cost income ratio is an efficiency measure similar to the operating margin one which is useful to measure how costs are changing compared to income.

- Trade payable ratio: this ratio reveals how often the firm payables turn over during the year; a high ratio means a relatively short time between purchase of goods and services and their payment; a low ratio may be a sign that the company has chronic cash shortages.

- Liquidity ratio: this ratio measures the extent to which a firm can quickly liquidate assets and cover short-term liabilities. It is therefore of interest to short-term creditors.

- Cash ratio: this ratio indicates the cash a company can generate in relation to its size.

- Equity ratio: it measures a companyâs financial leverage calculated by dividing a particular measure of equity by the firmâs total assets.

suggested by Creditreform based on its experience. More details about the data can be found in Figini and Giudici 2011.

4.2 Empirical evidence

The first step of our work was to select the independent variables to be used for the construction of the models. As stated in Chapter 1, one of the assumptions underlying the model of Cox regression (Cox) is the proportion of the risks, which is why we conducted the tests for verifying the proportionality of the risks is a prerequisite for applying the Cox model . To test this assumption, we used a test based on Schoenfeld residuals after estimating model parameters. Before observing the results of the tests we give an interpretation of the parameters: $\beta > 1$ means that the covariate increases the risk, $\beta < 1$ means that the covariate decreases the risk and $\beta = 1$ suggests, finally, that covariate risk and are independent. As regards the tests conducted it is the test proposed by Schienfeld (1982) and based on the residues defined by the same author. The test tells us that the assumption of proportionality is respected by a given covariate if the Schoenfeld residuals for this particular covariate are not in relation with the survival time, that is, the residues of Schienfeld are uncorrelated with time. Table 4.1 shows the result of the test on all covariates after the fitina of the Cox model on the whole set of covariates:

Variable	**p-value**
supplier target days	< 0.05
outside capital structure	> 0.05
cash ratio	> 0.05
capital tied up	< 0.05
equity ratio	> 0.05
cost income ratio	> 0.05
trade payable ratio	< 0.05
liabilities ratio	> 0.05
liquidity ratio	> 0.05
Global	< 0.05

Table 4.1: Tests for checking the proportion of the risks for the Cox model

If the p-value of the test is greater than 0.05 for all covariates and for the overall pattern then the proportional hazards assumption is verified. In this case, as is possible to see from the second column of Table 4.1, not all the explanatory variables fulfills this assumption for which we take into account the only variables that pass the test.
Secondly we fit the Logistic Regression (LR) model and let's see what covariates are significant for this model:

1. outside capital structure,
2. capital tied up,
3. cost income ratio,
4. liabilities ratio.

the last selection of the independent variables for the construction of all the models occurred crossing the results of the test of proportionality of risks in the Cox model and the significance of variables for the model of LR. This means that the variables used to build the models are the following:

1. outside capital structure,
2. cost income ratio,
3. liabilities ratio.

At this point we could build the models and begin the process of our combianzione. We will present first the results obtained on only models that are

Variable	**coef**	**exp(coef)**	**se(coef)**	**z**	**Pr(>|z|)**	
outside capital structure pc	2.5924	13.3617	0.6550	3.9580	7.57e-05	***
cost income ratio pc	3.4582	31.7597	0.5797	5.9650	2.44e-09	***
liabilities ratio pc	4.4521	85.8039	0.8349	5.3330	9.68e-08	***

Table 4.2: Fitting semi-parametric Cox model that takes into account the effect of covariates

time-dependent, the semi-parametric model of Cox and the technique non-parametric Survial Random Forest (RSF). Fitting the Cox model has provided the results reported in Table 4.2.

As we can see covariates considered have a great level of significance; mettimo out that in the initial model, with all covariates, no variable proved significant for the Cox model.
The second time-dependent model considered is the RSF. The characteristics of this model are given in Table 4.3.

Sample size	742
Number of deaths	96
Number of trees	100
Minimum terminal node size	3
Average no. of terminal nodes	77.01
No. of variables tried at each split	2
Total no. of variables	3
Analysis	RSF
Family	surv
Splitting rule	logrank *random*
Number of random split points	10
Error rate	23.54%

Table 4.3: Random Survival Forest

As is possible to see from Table 4.3 the number of trees generated is 100. Table 4.4 summarizes the model RSF obtained.

Variable	**Importance**	**Relative Imp**
outside capital structure	0.0670	1.0000
liabilities ratio	0.0470	0.7009
cost income ratio	0.0235	0.3512

Table 4.4: Importance and relative importance of variables for RSF

According to the output described in Table 4.4 of the variables that are more important in explaining our dependent variable (Solvency) are: outside

capital structure pc and liabilities ratio.

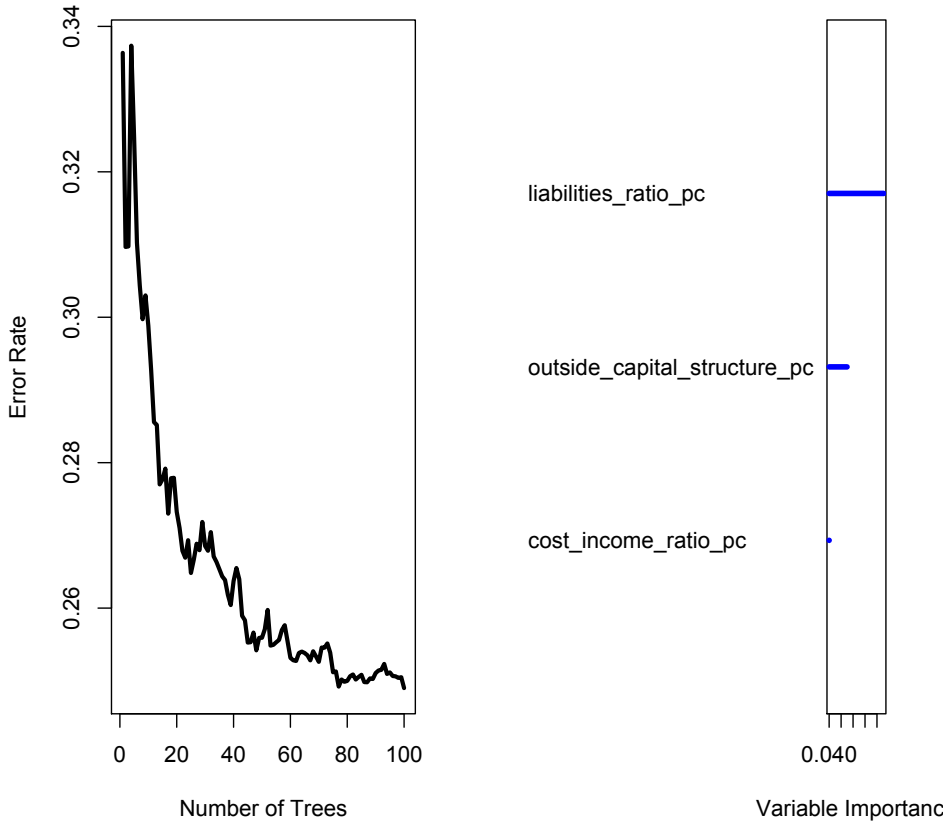

Figure 4.1: Random Survival Forest

The graph on the left in Figure 4.1 represents the fraction of the error rate for the model RSF as a function of the number of trees. As the graph on the right shows the out-of-bag value of the importance of covariates. The graphical analysis confirms the results reported in Table 4.4.
For time-dependent models we calculated the Harrell's concordance index (C-index), described in chapter 1 section 1.4. We calculated this index both insample and through the three cross-validation (cv) techniques presented in chapter 3. Please note that due to space constraints, in this thesis we will present only the results obtained following the application of 10-fold cross-validation, in However the full results can be obtained from the authors. The value of the C-index based on the sample is shown in Table 4.3 with its confidence interval.

Table 4.6 shows the values of the C-index for every moment of the time horizon of our study of the court. The time range is from 0 to 6. We inform

Model	C-index	se	lower	upper	p-value
Cox	0.1945	0.0436	0.1090	0.2800	2.504296e-12
RSF	0.9789	0.0094	0.9545	1.0000	0

Table 4.5: C-Index with CI

Model	time 1	time 2	time 3	time 4	time 5	time 6
Cox	0.8500	0.8040	0.7950	0.7950	0.7950	0.7950
RSF	0.6330	0.6330	0.7640	0.7640	0.7640	0.7640

Table 4.6: C-index 10-fold cv

you that from now on, given the stability of the results from the third point in time, we report only risulati for the first three years of study. This stability is supposed to be due to the absence of the subject censored after the third year. As is possible to see, the results from cross-validation showed a situation upside down from those shown in Table 4.5. In the latter the best model in terms of C-index is the RSF that has a high index value (0.9789), while the Cox model seems to provide worse results of those that we can have by entrusting the choice to the case (value of C -index <0.5). The results collected in Table 4.6, however, tell us that the best model, for each instant of time, is the Cox model. We notice the difference between the two models is that the Cox has a performance that worsens with increasing time while the RSF appears to provide results Sharpening time increases. It must be said that the results in sample are obtained over the entire time horizon while the results from cross-validation are determined at each point in time. Insample results, see Table 4.5, of C-index suggest that the Cox has a poor discriminative power, what we get from the Cox model is worst of example results that would realize through the toss of a coin. Instead, the RSF presents excellent discriminative power ($C - index \geq 0.9$). The results reported in Table 4.6, however, tell us that the Cox has a good discriminative power ($0.8 < C - index < 0.9$) for instant T = 1 while the power becomes acceptable ($0.7 < C - index < 0.8$) increasing time spent. The RSF has a reasonable discriminating power in the first two time instants (T = 1 and T = 2), T = 3 forwards increases the power that can be considered good.

Table 4.7 shows the performance measuresfor the two models, Cox and RSF, at different points in time. Point out that the measures set out in Table 4.7 are those described in chapter 1, SpSe99 is the specificity when the sensitivity is set at a level of 99% while SeSp99 is nothing more than the Sensitivity when

specificity is set at a level of 99%. Consider, as an example the column that shows the AUC value, we consider this measure since it is the most used to evaluate the performance of a model. What is more evident is that the Cox model presents a strong stability of the results over the period while the RSF is characterized by an improvement, in the case of AUC, the performance time increases.
We remember that in our study the main interest is not to use the measures listed in Table 4.7 as a tool for assessing the performance of the models, but to use them as weights to the forecasts obtained from these models in order to create weighed averaged models.

A very important measure for our work is the Mahalanobis distance. In fact this distance, along with the measures just observed, will be used also as a weight in the combination scheme of forecasts. We chose this distance because it allows us to assess the similarity, and so the proximity between two objects, in our case the observed Solvency and Solvency expected. Obviously what we want is that the estimate of our dependent variable, Solvency, is as close as possible to the observed value of the same variable. To interpret this measure, it is clear that the more the two objects is considered to resemble the more will be the small value of the Mahalanobis distance which is to measure the similarity.
In Table 4.8 we report the main summary statistics that describe the vector of Mahalanobis distances. From the first column we have, in order, the minimum (Min), the first quartile (Q1), the second quartile (Q2), the average (Mean), the third quartile (Q3), the maximum (Max), these indexes position used to steal information about the Mahalanobis distance. The last two columns of the same table, instead, show the measures of variability. The penultimate column shows the reciprocal of the variance while the last column we find the coefficient of variation.

At this point we have calculated all the measures we need to determine which average models as indicated by the equation (3.3). Once you get your models again we calculate average performance measures and the Mahalanobis distance so you can compare models combined with those individuals. The Table 4.9 shows the summarizing statistics for all average models, we remember that we are still working in the approach time-dependent models for which we average at different points in time and then performance measures and Mahalanobis distance to the corresponding instants, in the specific Tables 4.9 and 4.10 summarize the risults obtained in T=3, results for the moments T=1 and T=2 are shown in Appendix B. In row we have average models. Specifically,

Model	H	Gini	AUC	AUCH	KS	MER	MWL	SpSe99	SeSp99
Cox_{T1}	0.3859	0.6807	0.8403	0.8539	0.5668	0.1173	0.0976	0.1827	0.0938
RSF_{T1}	0.0274	0.0786	0.5393	0.5403	0.0797	0.1280	0.2073	0.9334	0.0313
Cox_{T2}	0.3859	0.6807	0.8403	0.8539	0.5668	0.1173	0.0976	0.1827	0.0938
RSF_{T2}	0.2350	0.4751	0.7375	0.7437	0.4082	0.1159	0.1333	0.6455	0.1146
Cox_{T3}	0.3859	0.6807	0.8403	0.8539	0.5668	0.1173	0.0976	0.1827	0.0938
RSF_{T3}	0.88755	0.9712	0.9856	0.9878	0.9301	0.0377	0.0158	0.9102	0.5729

Table 4.7: Performance measure for Cox and RSF at different time points

Model	Min	Q_1	Q_2	Mean	Q_3	Max	$\frac{1}{\sigma^2}$	CV
Cox_{T1}	0.1484	0.1856	0.2439	1.9970	0.3047	238.8000	0.0097	5.0879
RSF_{T1}	0.1527	0.1527	0.1527	1.9970	0.1527	608.8000	0.0020	11.2616
Cox_{T2}	0.1484	0.1906	0.2574	1.9970	0.3291	205.6000	0.0121	4.5591
RSF_{T2}	0.1484	0.1725	0.1725	1.9970	0.1944	110.9000	0.0155	4.0151
Cox_{T3}	0.1484	0.2572	0.4910	1.9970	0.8581	24.5800	0.0765	1.8105
RSF_{T3}	0.1484	0.3147	0.6145	1.9970	1.2590	23.8600	0.0946	1.6276

Table 4.8: Summary statistics of Mahalanobis distance for Cox and RSF for different time points

with *mean* we denote the simple average of the forecasts obtained from the individual models, $mean_C$ indicates the average of individual forecasts weighing with the C-index associated with the corresponding model and so on. While in the column, in Table 4.9 are the summary statistics of Mahalanobis distance and Table 4.10, however, we find the performance measures.

Model	**Min**	Q_1	**Median**	**Mean**	Q_3	**Max**	$\frac{1}{Variance}$	**CV**
$mean$	0.1484	0.3041	0.6204	1.9970	1.2480	17.6800	0.1006	1.5787
$mean_C$	0.1484	0.3058	0.6199	1.9970	1.2480	17.6900	0.1007	1.5781
$mean_{\frac{1}{C}}$	0.1484	0.3043	0.6218	1.9970	1.2160	17.6800	0.1005	1.5794
$mean_H$	0.1485	0.3073	0.6115	1.9970	1.2840	19.9300	0.0971	1.6071
$mean_{Gini}$	0.1484	0.3054	0.6168	1.9970	1.2540	17.5900	0.0994	1.5878
$mean_{AUC}$	0.1484	0.3023	0.6246	1.9970	1.2160	17.66000	0.1002	1.5820
$mean_{AUCH}$	0.1484	0.3027	0.6252	1.9970	1.2130	17.6700	0.1002	1.5817
$mean_{KS}$	0.1484	0.3059	0.6151	1.9970	1.2670	17.9500	0.0988	1.5929
$mean_{MER}$	0.1484	0.2833	0.6094	1.9970	1.0900	17.1200	0.0993	1.5887
$mean_{MWL}$	0.1484	0.2924	0.6077	1.9970	1.0870	19.3700	0.0973	1.6055
$mean_{SpSe99}$	0.1484	0.3034	0.6089	1.9970	1.1290	23.3500	0.0932	1.6397
$mean_{SeSp99}$	0.1484	0.3024	0.6090	1.9970	1.1240	23.9800	0.0925	1.6466
$mean_{\frac{1}{AUC}}$	0.1484	0.3057	0.6187	1.9970	1.2730	17.6800	0.1008	1.5767
$mean_M$	0.1484	0.1815	0.2104	1.9970	0.2613	118.6000	0.0169	3.8464
$mean_{\frac{1}{M}}$	0.1484	0.2178	0.5866	1.9970	1.0160	22.3700	0.0877	1.6905
$mean_{RP}$	0.1484	0.2174	0.5864	1.9970	1.0160	22.3700	0.0877	1.6906

Table 4.9: Summary statistics of Mahalanobis distance for all time-dependent average models (T=3)

We look in particular the AUC shown in Table 4.10. The average model that performs better in terms of AUC is the one obtained with the weight described in equation (3.1) followed by the model $mean_{99}$ following by $mean_{99}$. We notice that average models $mean_{\frac{1}{M}}$ and $mean_{RP}$ have the same perfor-

mance in terms of AUC (AUC=0.9671). This model performs better even the Cox that in T=3 has an AUC equal to 0.8403 while the RSF has an AUC equal to 0.9856 and this means that it has a level of performance superior to all models both individual and average, as you can see in Table 4.7.
Table 4.11 summarizes the results of the performance of individual and average models in T=3 in terms of AUC with its confidence interval (CI). We note that the average patterns obtained using as weights the performance measures are all characterized by a level of accuracy, measured in terms of AUC, lower than the RSF while all average models have a performance much better than the Cox.

The final evaluation of our models is done in terms of matrices of confusion.For every single model and we calculate the average matrix of confusion and try to understand if the estimates made on the basis of the model is correct or not. Let's look in particular:

1. True Positive (TP) when the positive forecast corresponds to a positive obsereved value,

2. True Negative (TN), when the negative forecast corresponds to a negative obsereved value,

3. False Positive (FP), when the positive forecast corresponds to a negative obsereved value,

4. False Negative(FN) when the negative forecast corresponds to a positive obsereved value.

Look, for example, the scores for the Cox model at time T=3 given in Table 4.12. The results were obtained, in this case, with a cut-off equal to 0.5. Over 742 statistical units the data involving the real situation of Solvency are 13.34% (TP+TN), versus the 86.65% that produce a bad results. We observe a big value of FN for all models and for all point in time. Based on the confusion matrix we can calculate Sensitivity and Specificity. The first is the ratio of TP and the total units planned as positive (TP+FN) and is defined accuracy rate tells us how many observations are estimated correctly positive compared to the total of positive comments. Specificity is the ratio between TN and the total of the observations provided negative (FP+TN). Continuing with our look at the Cox model in T=3 we see:

1. Sensitivity: $\frac{TP}{TP+FN} = \frac{72}{72+619} = 10.42\%$,

Model	H	Gini	AUC	AUCH	KS	MER	MWL	SpeSe99	SeSp99
$mean$	0.7070	0.9039	0.9530	0.9577	0.8290	0.0849	0.0385	0.7075	0.2188
$mean_C$	0.7028	0.6501	0.9519	0.9569	0.8275	0.0863		0.6997	0.2188
$mean_{\frac{1}{C}}$	0.7106	0.9086	0.9543	0.9585	0.8306	0.0849	0.0381	0.7121	0.2396
$mean_H$	0.8103	0.9443	0.9721	0.9750	0.8940	0.0606	0.0239	0.7864	0.3333
$mean_{Gini}$	0.7556	0.9260	0.9630	0.9671	0.8553	0.0714	0.0326	0.7523	0.3125
$mean_{AUC}$	0.7291	0.9164	0.9582	0.9621	0.8414	0.0782	0.0357	0.7260	0.2604
$mean_{AUCH}$	0.7267	0.9156	0.9578	0.9618	0.8398	0.0795	0.0361	0.7229	0.2604
$mean_{KS}$	0.7707	0.9322	0.9661	0.9693	0.8677	0.0687	0.0298	0.7663	0.3125
$mean_{MER}$	0.5414	0.8156	0.9078	0.9159	0.7139	0.1132	0.0645	0.5682	0.1354
$mean_{MWL}$	0.2299	0.7541	0.8771	0.8854	0.6189	0.1213	0.1218	0.5728	0.3333
$mean_{SpSe99}$	0.8546	0.9573	**0.9786**	0.9824	0.9153	0.0431	0.0191	0.8189	0.3958
$mean_{SeSp99}$	0.8576	0.9593	**0.9796**	0.9826	0.9153	0.0431	0.1085	0.8282	0.3958
$mean_{\frac{1}{AUC}}$	0.6823	0.8959	0.9480	0.9522	0.8151	0.0889	0.0417	0.6780	0.2140
$mean_M$	0.2877	0.5739	0.7870	0.8017	0.4970	0.1159	0.1133	0.1430	0.1146
$mean_{\frac{1}{M}}$	0.7747	0.9342	**0.9671**	0.9707	0.8693	0.0701	0.0295	0.7121	0.2604
$mean_{RP}$	0.7747	0.9342	**0.9671**	0.9707	0.8693	0.0701	0.0295	0.7121	0.2604

Table 4.10: Performance measure for all time-dependent average models (T=3)

2. Specificity: $\frac{TN}{TN+FP} = \frac{27}{27+24} = 52,94\%$.

We include Accuracy and Total Error (TE) in order to draw conclusions on perormance model. The Accuracy is the ratio between the correct units of total forecast ($\frac{TP+TN}{N}$) and measure the model's ability to predict correctly. While, TE is the complementary index of the accuracy and gives additional information about the error committed by the forecast model.

Model	AUC	CI
Cox	0.8199	0.7784-0.8614
RSF	0.9842	0.9770-0.9914
$mean$	0.9530	0.9383-0.9677
$mean_C$	0.9519	0.9370-0.9669
$mean_{\frac{1}{C}}$	0.9543	0.9399-0.9687
$mean_H$	0.9721	0.9616-0.9827
$mean_{Gini}$	0.9630	0.9504-0.9756
$mean_{AUC}$	0.9582	0.9445-0.9718
$mean_{AUCH}$	0.9578	0.9440-0.9715
$mean_{KS}$	0.9661	0.9542-0.9780
$mean_{MER}$	0.9078	0.8843-0.9314
$mean_{MWL}$	0.8771	0.8474-0.9067
$mean_{SpSe99}$	0.9786	0.9696-0.9877
$mean_{SeSp99}$	0.9796	0.9709-0.9884
$mean_{\frac{1}{AUC}}$	0.9480	0.9322-0.9637
$mean_M$	0.9335	0.9156-0.9514
$mean_{\frac{1}{M}}$	0.9671	0.9550-0.9791
$mean_{RP}$	0.9671	0.9551-0.9792

Table 4.11: AUC and relative CI for all time-dependent models (T=3)

- Accuracy: $\frac{72+27}{742} = 13.34\%$,
- TE: $1 - Accuracy = 1 - 0.1334 = 86.66\%$.

The model examined correctly predicts 13.34% of cases compared to 86.66% error. When it will positively have a chance of being correct 10.42% if it includes the negative will have a percentage of correctness highest at 52.94%. In the same way we can evaluate the performance of all other models.

Consider, now, the time-independent approach. First we calculated the performance measures and the vector of the Mahalanobis distances for predictions obtained through the single models. The table 4:16 describes the measures of performance for all four models (Cox, RSF, CT, LR), while Table 4:17 shows the measures of synthesis of Mahalanobis distance for the same models. We observe, in Table 16.4, the single best model in terms of AUC is the RSF with AUC = 0.9833, followed by LR (AUC = 0.8203), then CT with AUC = 0.7994 and finally the Cox model (0.7522).
We use the information contained in tables 4.16 and 4.17 for build average models. After obtaining the weighted average models with different performance measures and the Mahalanobis distance, again we calculate such measures to compare the results obtained on the basis of individual models with those pro-

Model	TP	TN	FP	FN
Cox	0.0970	0.0364	0.0323	0.8342
RSF	0.0283	0.0256	0.1011	0.8450
$mean$	0.0687	0.0283	0.0606	0.8423
$mean_C$	0.0687	0.0283	0.0606	0.8423
$mean_{\frac{1}{C}}$	0.0674	0.0283	0.0620	0.8423
$mean_H$	0.0526	0.0256	0.0768	0.8450
$mean_{Gini}$	0.0606	0.0256	0.0687	0.8450
$mean_{AUC}$	0.0633	0.0283	0.0660	0.8423
$mean_{AUCH}$	0.0647	0.0283	0.0647	0.8423
$mean_{KS}$	0.0566	0.0256	0.0728	0.8450
$mean_{MER}$	0.0889	0.0296	0.0404	0.8410
$mean_{MWL}$	0.0957	0.0337	0.0337	0.8369
$mean_{SpSe99}$	0.0364	0.0256	0.0930	0.8450
$mean_{SeSp99}$	0.0350	0.0270	0.0943	0.8437
$mean_{\frac{1}{AUC}}$	0.0350	0.0270	0.0943	0.8437
$mean_M$	0.0674	0.0445	0.0620	0.8261
$mean_{\frac{1}{M}}$	0.0687	0.0175	0.0606	0.8531
$mean_{RP}$	0.0687	0.0175	0.0606	0.8531

Table 4.12: Confusion matrix for all time-dependent models at cut-off=0.5 (T=3)

vided by the average patterns. Tables 4.18 and 4.19 show, respectively, the summary statistics of Mahalanobis distance and performance measures for all average models.

Looking at the table 4.19 we can see that the average model that performs best in terms of AUC $mean_{RP}$ followed by $mean_{\frac{1}{M}}$ with AUC, respectively, 0.9681 and 0.9659. We note that all the average models have a performance, in terms of AUC, better than the individual models to the exclusion of the RSF with AUC equal to 0.9833, as reported in the Table 4.16. The same conclusions can be drawn by looking at the Table 4.20 that summarizing the accuracy of the individual and average models in terms of AUC. In the third column of the same table we find the CI of AUC.

As in the case of time-dependent models, also in the time-independent context we go to observe the confusion matrices. We calculate the Specificity, the Sensitivity, Accuracy and Total Error for RSF:

1. Sensitivity: $\frac{TP}{TP+FN} = \frac{80}{80+26} = 75.47\%$,
2. Specificity: $\frac{TN}{TN+FP} = \frac{620}{620+16} = 97.48\%$,
3. Accuracy: $\frac{80+620}{742} = 94.34\%$,

Model	TP	TN	FP	FN
Cox	0.0795	0.0647	0.0499	0.8059
RSF	0.0040	0.0472	0.1253	0.8235
$mean$	0.0391	0.0512	0.0903	0.8194
$mean_C$	0.0391	0.0512	0.0903	0.8194
$mean_{\frac{1}{C}}$	0.0377	0.0512	0.0916	0.8194
$mean_H$	0.0189	0.0431	0.1105	0.8275
$mean_{Gini}$	0.0296	0.0485	0.0997	0.8221
$mean_{AUC}$	0.0377	0.0485	0.0916	0.8221
$mean_{AUCH}$	0.0377	0.0499	0.0916	0.8208
$mean_{KS}$	0.0243	0.0458	0.1051	0.8248
$mean_{MER}$	0.0593	0.0606	0.0701	0.8010
$mean_{MWL}$	0.0647	0.0606	0.0647	0.8010
$mean_{SpSe99}$	0.0108	0.0404	0.1186	0.8302
$mean_{SeSp99}$	0.0081	0.0404	0.1213	0.8302
$mean_{\frac{1}{AUC}}$	0.0418	0.0526	0.0876	0.8181
$mean_M$	0.0404	0.0687	0.0890	0.8019
$mean_{\frac{1}{M}}$	0.0350	0.0404	0.0943	0.8302
$mean_{RP}$	0.0350	0.0404	0.0943	0.8302

Table 4.13: Confusion matrix for all time-dependent models at cut-off=0.6 (T=3)

4. TE: $1 - Accuracy = 1 - 0.9434 = 5.66\%$.

The RSF correctly predicts,with a cut-off=0.5, 94.34% of cases compared to 5.66% error. When it will positively have a chance of being correct 75.47% if it includes the negative will have a percentage of correctness highest at 97.48%. In the same way we can evaluate the performance of all other models.

In conclusion we can say that we get excellent performance in terms of accuracy and correctly classified observations, when we consider the time-independent approach, while the only time-dependent models commit a high percentage of error in the classification of statistical units, however, must consider that they have a good ability to predict the units classified negatively (about 52%) compared with 10% probability of correctly classify positive units. TM

4.2.1 Churn analysis example

In many sectors of the customers become churners when they cease their subscription and decide to sign a new contract with a competitor. This event is a major concern for businesses with customers who can easily become customers of competitors, for example, credit card issuers, insurance companies, pay per

Model	TP	TN	FP	FN
Cox	0.0539	0.1213	0.0755	0.7493
RSF	0.0013	0.0849	0.1280	0.7857
$mean$	0.0121	0.1065	0.1173	0.7642
$mean_C$	0.0121	0.1065	0.1173	0.7642
$mean_{\frac{1}{C}}$	0.0121	0.1065	0.1173	0.7642
$mean_H$	0.0013	0.0916	0.1280	0.7790
$mean_{Gini}$	0.0054	0.1011	0.1240	0.7695
$mean_{AUC}$	0.0081	0.1051	0.1213	0.7655
$mean_{AUCH}$	0.0081	0.1051	0.1213	0.7655
$mean_{KS}$	0.0027	0.0984	0.1267	0.7722
$mean_{MER}$	0.0323	0.1132	0.0970	0.7574
$mean_{MWL}$	0.0418	0.1173	0.0876	0.7534
$mean_{SpSe99}$	0.0013	0.0863	0.1280	0.7844
$mean_{SeSp99}$	0.0013	0.0863	0.1280	0.7844
$mean_{\frac{1}{AUC}}$	0.0162	0.1105	0.1132	0.7601
$mean_M$	0.0135	0.1267	0.1159	0.7439
$mean_{\frac{1}{M}}$	0.0094	0.0795	0.1199	0.7911
$mean_{RP}$	0.0094	0.0795	0.1199	0.7911

Table 4.14: Confusion matrix for all time-dependent models at cut-off=0.7 (T=3)

view companies and telecommunications companies. In recent years the major changes of the markets in which these companies have made strategic management churn to compete on the market and to create a series of customer-focused marketing operations. Through good management churn the company is able to determine which customers are more likely to churn and which are the most loyal customer. Churn management also allows companies to give a value to the customers, that is to figure out when it is convenient to keep a customer and when the customer agrees that migrate to another company for little or no profitable. With all this information available to the company can implement marketing strategies that allow you to attract customers more profitable and have a lower likelihood of churn. One of the most used techniques to cope with the abandonment is the data mining that can be used for two main purposes:

1. predict whether a particular customer will churn and when it will happen,
2. understand why particular customers churn.

Through the prevision of clients who are at risk of churn, the company is able to reduce the dropout rate by offering, for example, services that convince the customer to stay. Understanding the reason for abandonment, however, it is

Model	**TP**	**TN**	**FP**	**FN**
Cox	0.0296	0.2304	0.0997	0.6402
RSF	0	0.1509	0.1294	0.7197
$mean$	0.0013	0.1819	0.1280	0.6887
$mean_C$	0.0013	0.1860	0.1280	0.6846
$mean_{\frac{1}{C}}$	0.0013	0.1819	0.1280	0.6887
$mean_H$	0.0013	0.1792	0.1280	0.6914
$mean_{Gini}$	0.0013	0.1806	0.1280	0.6900
$mean_{AUC}$	0.0013	0.1806	0.1280	0.6900
$mean_{AUCH}$	0.0013	0.1806	0.1280	0.6900
$mean_{KS}$	0.0013	0.1792	0.1280	0.6914
$mean_{MER}$	0.0108	0.1968	0.1186	0.6739
$mean_{MWL}$	0.0189	0.2170	0.1105	0.6536
$mean_{SpSe99}$	0	0.1739	0.1294	0.6968
$mean_{SeSp99}$	0	0.1725	0.1294	0.6981
$mean_{\frac{1}{AUC}}$	0.0013	0.1819	0.1280	0.6887
$mean_M$	0.0013	0.2116	0.1280	0.6590
$mean_{\frac{1}{M}}$	0.0013	0.1456	0.1280	0.7251
$mean_{RP}$	0.0013	0.1456	0.1280	0.7251

Table 4.15: Confusion matrix for all time-dependent models at cut-off=0.8 (T=3)

Model	H	Gini	AUC	AUCH	KS	MER	MWL	SpSe99	SeSp99
Cox	0.2143	0.5045	0.7522	0.7692	0.4300	0.1294	0.3235	0.1250	0
RSF	0.3111	0.5414	0.9833	0.7891	0.4470	0.1213	0.1246	0.0668	0.0833
CT	0.3701	0.5988	0.7994	0.7994	0.4926	0.0997	0.1143	0.0346	0.1636
LR	0.3456	0.6406	0.8203	0.8329	0.5306	0.1240	0.1057	0.1811	0.0521

Table 4.16: Performance measure for single models

Model	**Min**	Q_1	**Median**	**Mean**	Q_3	**Max**	$\frac{1}{Variance}$	**CV**
Cox	0.1484	0.2697	0.4122	1.9970	0.5658	136.9000	0.0249	3.1726
RSF	0.1484	0.2907	0.5992	1.9970	0.9650	15.9600	0.0945	1.6286
CT	0.2185	0.2460	0.2460	1.9970	0.3017	20.8700	0.0566	2.1038
LR	0.1484	0.3057	0.5895	1.9970	1.0070	24.8100	0.0909	1.6607

Table 4.17: Summary statistics of Mahalanobis distance for single models

useful to society to implement policies that enable the customer satisfaction before it abandons the company in favor of a competitor.

After clarifying what does the churn analysis are going to present our method of combining forecasts to a problem of churn analysis of data from a company's pay-per-view. Below in the datail riesults obtained. As regards the variables studied it comes to variables that relate to the type of subscription as, for

Model	**Min**	Q_1	**Median**	**Mean**	Q_3	**Max**	$\frac{1}{Variance}$	**CV**
$mean$	0.1484	0.3066	0.5378	1.9970	0.9830	22.7700	0.0932	1.6403
$mean_H$	0.1484	0.3131	0.5209	1.9970	0.9331	20.8400	0.0914	1.6563
$mean_{Gini}$	0.1484	0.3044	0.5298	1.9970	0.9439	22.0700	0.0930	1.6417
$mean_{AUC}$	0.1484	0.3089	0.5360	1.9970	0.9691	22.5200	0.0931	1.6405
$mean_{AUCH}$	0.1484	0.3087	0.5377	1.9970	0.9717	22.4800	0.0934	1.6385
$mean_{KS}$	0.1484	0.3058	0.5306	1.9970	0.9484	22.2300	0.0930	1.6414
$mean_{MER}$	0.1484	0.3093	0.5525	1.9970	1.0400	22.6300	0.0950	1.6243
$mean_{MWL}$	0.1484	0.3081	0.5382	1.9970	0.9976	23.2300	0.0932	1.6399
$mean_{SpSe99}$	0.1484	0.3064	0.5723	1.9970	0.9675	22.4100	0.0948	1.6262
$mean_{SeSp99}$	0.1484	0.2909	0.4130	1.9970	0.5496	22.8600	0.0732	1.8507
$mean_{\frac{1}{AUC}}$	0.1484	0.3061	0.5387	1.9970	0.9877	23.0200	0.0932	1.6401
$mean_M$	0.1484	0.3346	0.5141	1.9970	1.1250	16.5700	0.0960	1.6161
$mean_{\frac{1}{M}}$	0.1484	0.2466	0.5052	1.9970	0.8101	34.6900	0.0668	1.9372
$mean_{RP}$	0.1484	0.2502	0.5236	1.9970	0.8658	31.6300	0.0698	1.8949

Table 4.18: Summary statistics of Mahalanobis distance for average models

example, the rental of the decoder, the packages included in the offer (sports, movies, etc.), the income range, type of decoder , etc. Then we have a variable duration and the dependent variable of our interest that represents the state of the client (abandoned or not).

Table 4.25 shows the values of in sample C-index for the Cox and RSF. How can we make out of the id that index in the first column of the tabellla RSF has an excellent discriminating power suggested by a C-index equal to 0.9320 while the Cox model has an acceptable discriminative power (C-index = 0.7142).

In Table 4.26 we have the performance measures for the single models. The first thing you notice is that the Cox model and LR are characterized by the

Model	H	Gini	AUC	AUCH	KS	MER	MWL	SpSe99	SeSp99
$mean$	0.5993	0.8621	0.9311	0.9373	0.7448	0.0916	0.0575	0.7214	0.3438
$mean_H$	0.6141	0.8688	0.9344	0.94018	0.7668	0.0903	0.0525	0.7229	0.3542
$mean_{Gini}$	0.5951	0.8586	0.9293	0.9371	0.7413	0.0903	0.0583	0.7074	0.3438
$mean_{AUC}$	0.5985	0.8609	0.9305	0.9373	0.7448	0.0903	0.0575	0.7167	0.3646
$mean_{AUCH}$	0.5987	0.8612	0.9306	0.9373	0.7448	0.0903	0.0575	0.7183	0.3646
$mean_{KS}$	0.5923	0.8575	0.9288	0.9361	0.7386	0.0916	0.0589	0.7059	0.3438
$mean_{MER}$	.5988	0.8601	0.9301	0.9369	0.7429	0.0957	0.0579	0.7214	0.3125
$mean_{MWL}$	0.6064	0.8657	0.9329	0.9388	0.7522	0.0916	0.0558	0.7260	0.3438
$mean_{SpSe99}$	0.5360	0.8191	0.9095	0.9167	0.6960	0.1132	0.0685	0.6610	0.1563
$mean_{SeSp99}$	0.6755	0.8959	0.9479	0.9532	0.8028	0.0903	0.0444	0.7586	0.3021
$mean_{\frac{1}{AUC}}$	0.6011	0.86291	0.9314	0.9378	0.7475	0.0916	0.0569	0.7260	0.3438
$mean_M$	0.5286	0.8010	0.9005	0.9092	0.7183	0.1145	0.0635	0.7183	0.0521
$mean_{\frac{1}{M}}$	0.7481	0.9319	0.9659	0.9705	0.8191	0.0593	0.0407	0.7647	0.3750
$mean_{RP}$	0.7604	0.9362	0.9681	0.9726	0.8287	0.0553	0.0386	0.7771	0.3750

Table 4.19: Performance measure for all model combinations

same level of performance. All models have a good level of accuracy but the models that perform best are the Cox model and LR.

Table 4.27 gathers summary statistics of the Mahalanobis distance. Recall that the indicators considered are, starting from the leftmost column, the minimum, the first quartile, median, average, the third quartile, maximum, the reciprocal of the variance and the variation coefficient. Table 4.28 reports summary statistics of Mahalanobis distance for average models.

Model	AUC	CI
Cox	0.7522	0.7064-0.7980
RSF	0.9833	0.9761-0.9905
CT	0.7994	0.7465-0.8522
LR	0.8203	0.7788-0.8618
$mean$	0.9311	0.9119-0.9503
$mean_H$	0.9344	0.9158-0.9530
$mean_{Gini}$	0.9293	0.9097-0.9490
$mean_{AUC}$	0.9305	0.9111-0.9498
$mean_{AUCH}$	0.9306	0.9113-0.9499
$mean_{KS}$	0.9288	0.9090-0.9485
$mean_{MER}$	0.9301	0.9108-0.9494
$mean_{MWL}$	0.9329	0.9141-0.9516
$mean_{SpSe99}$	0.9095	0.8867-0.9323
$mean_{SeSp99}$	0.9479	0.9323-0.9636
$mean_{\frac{1}{AUC}}$	0.9314	0.9124-0.9505
$mean_M$	0.9005	0.8776-0.9234
$mean_{\frac{1}{M}}$	0.9659	0.9528-0.9791
$mean_{RP}$	0.9681	0.9557-0.9806

Table 4.20: AUC and relative CI for all models

Model	TP	TN	FP	FN
Cox	0.0162	0.83560	0.1132	0.0350
RSF	0.1078	0.8356	0.0216	0.0350
CT	0.0121	0.8585	0.0876	0.0418
LR	0.0135	0.8571	0.1105	0.0189
$mean$	0.0391	0.8639	0.0903	0.0067
$mean_H$	0.0458	0.8639	0.08362	0.0067
$mean_{Gini}$	0.0391	0.8625	0.0903	0.0081
$mean_{AUC}$	0.0391	0.8639	0.0903	0.0067
$mean_{AUCH}$	0.0391	0.8639	0.0903	0.0067
$mean_{KS}$	0.0391	0.8639	0.0903	0.0067
$mean_{MER}$	0.0364	0.8652	0.0930	0.0054
$mean_{MWL}$	0.0391	0.8639	0.0903	0.0067
$mean_{SpSe99}$	0.0229	0.8598	0.1065	0.0108
$mean_{SeSp99}$	0.0189	0.8518	0.0741	0.0553
$mean_{\frac{1}{AUC}}$	0.0391	0.8639	0.0903	0.0067
$mean_M$	0.0526	0.8329	0.0768	0.0377
$mean_{\frac{1}{M}}$	0.0283	0.8679	0.1011	0.0027
$mean_{RP}$	0.0283	0.8679	0.1011	0.0027

Table 4.21: Confusion matrix for all models at cut-off=0.5

Model	TP	TN	FP	FN
COX	0.0094	0.8437	0.1199	0.0270
RSF	0.0863	0.8491	0.0431	0.0216
CT	0.0418	0.8585	0.0876	0.0121
LR	0.0040	0.8652	0.1253	0.0054
$mean$	0.0148	0.8679	0.1146	0.0027
$mean_H$	0.0148	0.8693	0.1146	0.0013
$mean_{Gini}$	0.0121	0.8693	0.1173	0.0013
$mean_{AUC}$	0.0135	0.8679	0.1159	0.0027
$mean_{AUCH}$	0.0135	0.8679	0.1159	0.0027
$mean_{KS}$	0.0121	0.8679	0.1173	0.0027
$mean_{MER}$	0.0162	0.8679	0.1132	0.0027
$mean_{MWL}$	0.0162	0.8679	0.1132	0.0027
$mean_{SpSe99}$	0.0135	0.8679	0.1159	0.0027
$mean_{SeSp99}$	0.0391	0.8625	0.0903	0.0081
$mean_{\frac{1}{AUC}}$	0.0148	0.8679	0.1146	0.0027
$mean_M$	0.0243	0.8477	0.1051	0.0229
$mean_{\frac{1}{M}}$	0.0108	0.8693	0.1186	0.0013
$mean_{RP}$	0.0121	0.8693	0.1173	0.0013

Table 4.22: Confusion matrix for all models at cut-off=0.6

Model	TP	TN	FP	FN
Cox	0.0040	0.8504	0.1253	0.0202
RSF	0.0647	0.8612	0.0647	0.0094
CT	0.0418	0.8585	0.0876	0.0121
LR	0	0.8679	0.1294	0.0027
$mean$	0.0067	0.8706	0.1226	0
$mean_H$	0.0054	0.8706	0.1240	0
$mean_{Gini}$	0.0067	0.8706	0.1226	0
$mean_{AUC}$	0.0067	0.8706	0.1226	0
$mean_{AUCH}$	0.0067	0.8706	0.1226	0
$mean_{KS}$	0.0067	0.8706	0.1226	0
$mean_{MER}$	0.0067	0.87066	0.1226	0
$mean_{MWL}$	0.0067	0.8706	0.1226	0
$mean_{SpSe99}$	0.0027	0.8706	0.1267	0
$mean_{SeSp99}$	0.0202	0.8679	0.1092	0.0027
$mean_{\frac{1}{AUC}}$	0.0067	0.8706	0.1226	0
$mean_M$	0.0067	0.8625	0.1226	0.0081
$mean_{\frac{1}{M}}$	0.0040	0.8706	0.1253	0
$mean_{RP}$	0.0054	0.8706	0.1240	0

Table 4.23: Confusion matrix for all models at cut-off=0.7

Model	TP	TN	FP	FN
Cox	0.0013	0.8544	0.1280	0.0162
RSF	0.4580	0.8693	0.0836	0.0013
CT	0	0.8706	0.1294	0
LR	0	0.8706	0.1294	0
$mean$	0	0.8706	0.1294	0
$mean_H$	0	0.8706	0.1294	0
$mean_{Gini}$	0	0.8706	0.1294	0
$mean_{AUC}$	0	0.8706	0.1294	0
$mean_{AUCH}$	0	0.8706	0.1294	0
$mean_{KS}$	0	0.8706	0.1294	0
$mean_{MER}$	0	0.8706	0.1294	0
$mean_{MWL}$	0	0.8706	0.1294	0
$mean_{SpSe99}$	0	0.8706	0.1294	0
$mean_{SeSp99}$	0.0027	0.8706	0.1267	0
$mean_{\frac{1}{AUC}}$	0	0.8706	0.1294	0
$mean_M$	0	0.8706	0.1294	0
$mean_{\frac{1}{M}}$	0	0.8706	0.1294	0
$mean_{RP}$	0	0.8706	0.1294	0

Table 4.24: Confusion matrix for all models at cut-off=0.8

In Table 29.4 are the performance measures for all average models. Looking at the column corresponding to the AUC we may note that the model that presents the best performance is the average model weighed with the relative performance weight (AUC=0.8721) followed by the average model weighed with reciproca of Mahalanobis distance (AUC=0.8710). If we compared with the results for individual models we see that the Cox model (AUC = 0.8915) and LR (AUC = 0.8915) have a performance in terms of accuracy better than $mean_{RP}$ and $mean_{\frac{1}{M}}$. While the CT (AUC = 0.8714) has a performance almost equal to the last two. The model achieves a performance worse is the RSF (AUC = 0.8328) compared to all single and average models.
Table 4.30 summarizes the results just commented with regard to the AUC and reports the respective confidence intervals.
Tables 31.4, 4.32, 4.33 and 4.34 show the results of the confusion matrices. We focus our attention on the table 4.34. We observe, for example, the results for the model $mean_{RP}$. We calculate the Specificity, the Sensitivity, Accuracy and Total Error:

1. Sensitivity: $\frac{TP}{TP+FN} = \frac{3549}{3549+13132} = 21.28\%$,

2. Specificity: $\frac{TN}{TN+FP} = \frac{17546}{17546+774} = 95.78\%$,

P	Cox	RSF	CT	LR	$mean$	$mean_H$	$mean_{Gini}$	$mean_{AUC}$	$mean_{AUCH}$	$mean_{KS}$	$mean_{MER}$	$mean_{MWL}$
real data	0.7522	0.9833	0.7994	0.8203	0.9311	0.9344	0.9293	0.9305	0.9306	0.9288	0.9301	0.9329
0.01	0.5540	1	0.9251	0.7056	0.5436	0.5150	0.5127	0.5181	0.5272	0.5021	0.6345	0.6275
0.02	0.5230	1	0.8827	0.6700	0.9965	0.9995	0.9989	0.9978	0.9976	0.9991	0.8528	0.8341
0.03	0.6647	0.9787	0.5908	0.6088	0.9099	0.9773	0.9769	0.9420	0.9415	0.9760	0.8935	0.6829
0.04	0.6521	0.9028	0.6557	0.6141	0.8872	0.8963	0.8963	0.8928	0.8923	0.8964	0.8872	0.8413
0.05	0.7214	0.9993	0.6232	0.6081	9988	0.9993	0.9992	0.9992	0.9992	0.9992	0.7653	0.5649
0.06	0.5668	1	0.8103	0.6489	0.9999	1	1	1	1	1	0.7451	0.7348
0.07	0.6120	0.9999	0.8835	0.6045	0.9985	0.9999	0.9998	0.9994	0.9992	0.9998	0.8322	0.7617
0.08	0.7468	0.9982	0.7984	0.6084	0.9902	0.9983	0.9969	0.9942	0.9941	0.9976	0.8397	0.7413
0.09	0.6109	0.8252	0.5622	0.6046	0.7779	0.8090	0.8115	0.7938	0.7914	0.8074	0.7744	0.7399
0.10	0.6858	0.9977	0.7165	0.5799	0.9846	0.9972	0.9962	0.9914	0.9912	0.9967	0.9077	0.6446
0.15	0.6575	0.9952	0.6414	0.5991	0.9982	0.9970	0.9981	0.9986	0.9987	0.9983	0.8183	0.7896
0.20	0.7235	0.7119	0.5599	0.5627	0.6700	0.6345	0.6682	0.6689	0.6668	0.6495	0.6743	0.6782
0.25	0.7065	0.9971	0.5000	0.5772	0.9976	0.9979	0.9979	0.9980	0.9979	0.9982	0.6089	0.6035
0.30	0.7558	0.9962	0.5000	0.5517	0.9983	0.9981	0.9982	0.9986	0.9987	0.9985	0.5647	0.5267
0.35	0.7639	0.6027	0.5000	0.5330	0.5571	0.7422	0.6681	0.5928	0.5988	0.6991	0.5192	0.5186
0.40	0.7668	0.9915	0.5457	0.5540	0.9959	0.9952	0.9953	0.9958	0.9962	0.9960	0.6864	0.7196
0.45	0.7441	0.8323	0.5000	0.5451	0.7928	0.7803	0.8062	0.7986	0.7922	0.7849	0.7991	0.8014
0.50	0.7557	0.8702	0.5681	0.5605	0.8277	0.8343	0.8473	0.8367	0.8321	0.8268	0.8228	0.8219
0.60	0.7350	0.9835	0.500	0.5212	0.9908	0.9920	0.9912	0.9925	0.9920	0.9924	0.7905	0.6697
0.70	0.7483	0.9829	0.6123	0.5448	0.9905	0.9945	0.9932	0.9938	0.9936	0.9949	0.7594	0.6349
0.80	0.7185	0.6314	0.5000	0.5027	0.5276	0.6862	0.6077	0.5459	0.5674	0.6596	0.5588	0.5502
0.90	0.8046	0.9733	0.7796	0.6257	0.8701	0.9498	0.9771	0.9384	0.9113	0.9342	0.5836	0.5162

Table 4.25: AUC for all models for real and simulated data (1)

P	Cox	RSF	CT	LR	$mean_{SpSe99}$	$mean_{SeSp99}$	$mean_{\frac{1}{AUC}}$	$mean_M$	$mean_{\frac{1}{M}}$	$mean_{RP}$
real data	0.7522	0.9833	0.7994	0.8203	0.9095	0.9479	0.9314	0.9005	0.9659	0.9681
0.01	0.5540	1	0.9251	0.7056	0.5122	0.5462	0.5646	0.5258	0.5324	0.5413
0.02	0.5230	1	0.8827	0.6700	0.9997	0.9999	0.9951	0.9578	0.9988	0.9989
0.03	0.6647	0.9787	0.5908	0.6088	0.9132	0.9787	0.8860	1	0.9912	0.9986
0.04	0.6521	0.9028	0.6557	0.6141	0.8972	0.8413	0.8418	0.8549	0.9561	0.9348
0.05	0.7214	0.9993	0.6232	0.6081	0.9992	0.9993	0.9969	0.9697	1	1
0.06	0.5668	1	0.8103	0.6489	1	1	0.9994	0.9798	0.9999	0.9999
0.07	0.6120	0.9999	0.8835	0.6045	0.9999	1	0.9959	0.9693	0.9979	0.9980
0.08	0.7468	0.9982	0.7984	0.6084	0.9985	0.9983	0.9845	0.9505	0.9986	0.9987
0.09	0.6109	0.8252	0.5622	0.6046	0.8155	0.8066	0.7518	0.7260	0.9999	0.8613
0.10	0.6858	0.9977	0.7165	0.5799	0.9979	0.9975	0.9723	0.9445	1	0.5536
0.15	0.6575	0.9952	0.6414	0.5991	0.9963	0.9961	0.9958	0.9479	0.9944	0.9949
0.20	0.7235	0.7119	0.5599	0.5627	0.6998	0.5853	0.6715	0.7304	0.6748	0.6685
0.25	0.7065	0.9971	0.5000	0.5772	0.9971	0.9972	0.9957	0.9523	0.9631	0.9476
0.30	0.7558	0.9962	0.5000	0.5517	0.9962	0.9964	0.9954	0.9605	0.9677	0.9531
0.35	0.7639	0.6027	0.5000	0.5330	0.5472	0.5391	0.5299	0.5233	0.6766	0.7226
0.40	0.7668	0.9915	0.5457	0.5540	0.9910	0.9917	0.9943	0.9388	0.9456	0.9428
0.45	0.7441	0.8323	0.5000	0.5451	0.8309	0.8180	0.7865	0.7982	0.7547	0.7510
0.50	0.7557	0.8702	0.5681	0.5605	0.8680	0.8640	0.5106	0.7785	0.7650	07684
0.60	0.7350	0.9835	0.500	0.5212	0.9843	0.9847	0.9824	0.9048	0.7588	0.7718
0.70	0.7483	0.9829	0.6123	0.5448	0.9832	0.9843	0.9731	0.8570	0.6979	0.7149
0.80	0.7185	0.6314	0.5000	0.5027	0.5571	0.5087	0.5042	0.5085	0.5824	0.5764
0.90	0.8046	0.9733	0.7796	0.6257	0.9797	0.9835	0.7423	0.7288	0.6622	0.6406

Table 4.26: AUC for all models for real and simulated data (2)

3. Accuracy: $\frac{3549+13132}{35000} = 47.66\%$,

4. TE: $1 - Accuracy = 1 - 0.4766 = 52.34\%$.

$mean_{RP}$ correctly predicts,with a cut-off=0.8, 47.66% of cases compared to 52.34% error. When it will positively have a chance of being correct 21.28% if it includes the negative will have a percentage of correctness highest at 95.78%. In the same way we can evaluate the performance of all other models at all cut-off.

Table 4.35 reports the results of the AUC in different points in time for the Cox model and the RSF, we emphasize that it is cross-validation results obtained by 10-fold cross-validation method. for the first four time instants the AUC of RSF is smaller in AUC of the Cox model, starting from T = 15 the tendency is reversed and the RSF shows higher AUC than those for the Cox model. It is specified that, for the analysis of churn the time instants are considered on a monthly basis unlike what happened in the case of the analysis of PD incui time is expressed in years. We note also that the AUC from the table 4:35 to RSF increases all'aumetare time while the Cox follows this same trend up to T = 51, then he begins to address decreasing with increasing time.

Table 4.36, instead, shows the results in terms of C-index, always obtained with the method 10-fold cross-validation. Also in this case the RSF shows values of the C-index almost always greater than the Cox model, so it has a greater discriminating power in many time instants. We note that the decrease in the C-inex for the Cox model is smaller than what happens to the RSF. Infact, in T=3 the C-index of Cox is 0.9570 while the RSF is 0.9810, T=72 in the C-index for the Cox model is 0.9003 while the RSF is equal to 0.8790.

Table 4.37 shows the summary statistics of Mahalanobis distance for models single time instant T=3. While the Table 4.38 gathers perfrmance measures always for the single models in T=3. From Table 4.87 let 's see now that the model with a best performance at T=3 is the Cox model compared to RSF.

Table 4.39 shows summury statistics for all average models inT=3. we note that these indicators are all very close to each other for each combination of patterns.

Table 4.40 contains the results of the measures of performance for all models average. Focusing on the column of the AUC, of our interest, we see that the performance of all models combianti is almost similar, with respect to the individual models, we can say that, in terms of AUC, the Cox has a perfromance slightly higher than weighted average models as compared to RSF these models are best. In this case, therefore, the single model is improved, even if only slightly, the combined models. As usual, we summarize the results in terms of

AUC in T=3, both for the individual models for both patterns average, in the Table 4.41 where we report the value of the AUC and its confidence interval. Tables 4.42, 4.43, 4.44 and 4.45 summarize the confusion matrices in correspondence, respectively, of the cut-off 0.5, 0.6, 0.7 and 0.8 for all models analyzed. We look average models $mean_{RP}$ in Table 4.45 and determine:

1. Sensitivity: $\frac{TP}{TP+FN} = \frac{3717}{3717+30531} = 10.85\%$,
2. Specificity: $\frac{TN}{TN+FP} = \frac{130}{130+606} = 17.66\%$,
3. Accuracy: $\frac{3717+130}{35000} = 10.99\%$,
4. TE: $1 - Accuracy = 1 - 0.1099 = 89.01\%$.

The $mean_{RP}$ correctly predicts,with a cut-off=0.8, 10.99% of cases compared to 89.01% error. When it will positively have a chance of being correct 10.85% if it includes the negative will have a percentage of correctness highest at 17.66%. The results obtained for the other models can be commented on in the same way. We deduce that in the churn analysis time-dependent approach commit a high percentage of error in the classification of units, while the time-independent approach has better performance in terms of units are correctly classified.
The results for the time-dependent models were calculated for a time period that ranges from 3 to 72 months. We stress that their analysis of the PD, conducted in section 4.3, for churn analysis time is expressed in months and not in years. For reasons of economy of space it is not possible to report the results for all time instants but we considered that the results obtained are similar to those presented to the instant T=3.

Model	**C-index**	**se**	**lower**	**upper**	**p-value**
Cox	0.7142	0.0051	0.7043	0.7241	0
RSF	0.9320	0.0015	0.9282	0.9357	0

Table 4.27: C-Index for Cox and RSF with CI for churn analysis

Model	**H**	**Gini**	**AUC**	**AUCH**	**KS**	**MER**	**MWL**	**SpSe99**	**SeSp99**
Cox	0.5059	0.7830	0.8915	0.8927	0.6221	0.0848	0.0818	0.1946	0.3546
RSF	0.4465	0.6656	0.8328	0.8555	0.5626	0.0921	0.0946	0.2831	0.2697
CT	0.5207	0.7428	0.8714	0.8714	0.6481	0.0909	0.0762	0.5861	0.3437
LR	0.5059	0.7830	0.8915	0.8927	0.6221	0.0848	0.0818	0.1946	0.3547

Table 4.28: Performance measure for single models for churn analysis

Model	**Min**	Q_1	**Median**	**Mean**	Q_3	**Max**	$\frac{1}{Variance}$	**CV**
Cox	0.1409	0.1869	0.2809	2.0000	0.3532	33.0200	0.0547	2.1375
RSF	0.1409	0.1901	0.2395	2.0000	0.2423	60.9700	0.0405	2.4839
CT	0.1904	0.2374	0.2374	2.0000	0.4209	32.8000	0.0592	2.0545
LR	0.1409	0.1869	0.2809	2.0000	0.3532	33.0200	0.0547	2.1375

Table 4.29: Summary statistics of Mahalanobis distance for single models for churn analysis

Model	**Min**	Q_1	**Median**	**Mean**	Q_3	**Max**	$\frac{1}{Variance}$	**CV**
$mean$	0.1409	0.1907	0.2403	2.0000	0.2451	60.7700	0.0407	2.4778
$mean_H$	0.1409	0.1907	0.2405	2.0000	0.2455	60.7400	0.0408	2.4769
$mean_{Gini}$	0.1409	0.1907	0.2406	2.0000	0.2456	60.7400	0.0408	2.4768
$mean_{AUC}$	0.1409	0.1907	0.2404	2.0000	0.2453	60.7600	0.0407	2.4774
$mean_{AUCH}$	0.14094	0.1907	0.2404	2.0000	0.2452	60.7600	0.0407	2.4776
$mean_{KS}$	0.1409	0.1907	0.2405	2.0000	0.2455	60.7500	0.0407	2.4771
$mean_{MER}$	0.1409	0.1907	0.2403	2.0000	0.2450	60.7900	0.0407	2.4782
$mean_{MWL}$	0.14094	0.1906	0.24022	2.0000	0.2447	60.8000	0.0407	2.4787
$mean_{SpSe99}$	0.1409	0.1905	0.2406	2.0000	0.2453	60.7600	0.0407	2.4774
$mean_{SeSp99}$	0.1409	0.1905	0.2407	2.0000	0.2460	60.7100	0.0408	2.4760
$mean_{\frac{1}{AUC}}$	0.1409	0.1906	0.2403	2.0000	0.2450	60.7800	0.0407	2.4782
$mean_M$	0.1409	0.1698	0.1887	2.0000	0.1913	118.4000	0.0213	3.4228
$mean_{\frac{1}{M}}$	0.1409	0.2015	0.2592	2.0000	0.2660	68.8600	0.0420	2.4384
$mean_{RP}$	0.1409	0.2006	0.2590	2.0000	0.2661	72.0600	0.0413	2.4619

Table 4.30: Summary statistics of Mahalanobis distance for average models for churn analysis

Model	**H**	**Gini**	**AUC**	**AUCH**	**KS**	**MER**	**MWL**	**SpSe99**	**SeSp99**
$mean$	0.4562	0.7269	0.8635	0.8731	0.5670	0.0919	0.0937	0.3877	0.2726
$mean_H$	0.4570	0.7303	0.8651	0.8744	0.5670	0.0919	0.0937	0.3906	0.2726
$mean_{Gini}$	0.4570	0.7304	0.8652	0.8742	0.5670	0.0919	0.0937	0.3912	0.2726
$mean_{AUC}$	0.4567	0.7285	0.8643	0.8738	0.5669	0.0919	0.0937	0.3892	0.2726
$mean_{AUCH}$	0.4568	0.7279	0.8640	0.8737	0.5670	0.0919	0.0937	0.3886	0.2726
$mean_{KS}$	0.4569	0.7298	0.8649	0.8741	0.5670	0.0919	0.0937	0.3898	0.2726
$mean_{MER}$	0.4559	0.7255	0.8627	0.8726	0.5670	0.0919	0.0937	0.3876	0.2721
$mean_{MWL}$	0.4554	0.7227	0.8613	0.8716	0.5667	0.0919	0.0938	0.3844	0.2716
$mean_{SpSe99}$	0.4594	0.7320	0.8660	0.8771	0.5680	0.0919	0.0935	0.3803	0.2721
$mean_{SeSp99}$	0.4579	0.7348	0.8674	0.8755	0.5670	0.0918	0.0938	0.3938	0.2726
$mean\ \frac{1}{AUC}$	0.4558	0.7253	0.8627	0.8724	0.5670	0.0919	0.0937	0.3873	0.27219
$mean_M$	0.4499	0.7268	0.8634	0.8714	0.5917	0.1121	0.0884	0.3744	0932
$mean\ \frac{1}{M}$	0.5180	0.7419	0.8710	0.8856	0.6129	0.0797	0.0838	0.3981	0.3343
$mean_{RP}$	0.5182	0.7442	0.8721	0.8865	0.6136	0.0780	0.0836	0.3983	0.3223

Table 4.31: Performance measure for all model combinations for churn analysis

Model	AUC	CI
Cox	0.8915	0.8681-0.8969
RSF	0.8328	0.8254-0.8402
CT	0.8714	0.8649-0.8779
LR	0.8915	0.8861-0.8969
$mean$	0.8635	0.8577-0.8692
$mean_H$	0.8651	0.8595-0.8708
$mean_{Gini}$	0.8652	0.8595-0.8708
$mean_{AUC}$	0.8643	0.8586-0.8700
$mean_{AUCH}$	0.8640	0.8583-0.8697
$mean_{KS}$	0.8649	0.8593-0.8706
$mean_{MER}$	0.8627	0.8570-0.8685
$mean_{MWL}$	0.8613	0.8555-0.8672
$mean_{SpSe99}$	0.8660	0.8604-0.8716
$mean_{SeSp99}$	0.8674	0.8618-0.8729
$mean_{\frac{1}{AUC}}$	0.8627	0.8569-0.8684
$mean_M$	0.8634	0.8581-0.8688
$mean_{\frac{1}{M}}$	**0.8710**	0.8651-0.8768
$mean_{RP}$	**0.8721**	0.8663-0.8779

Table 4.32: AUC and relative CI for all models for churn analysis

Model	TP	TN	FP	FN
Cox	0.0540	0.8601	0.0695	0.0164
RSF	0.1050	0.3652	0.0185	0.5113
CT	0.0476	0.8615	0.0759	0.0151
LR	0.0540	0.8601	0.0695	0.0164
$mean$	0.1141	0.4581	0.0093	0.4184
$mean_H$	0.1140	0.4669	0.0095	0.4097
$mean_{Gini}$	0.1139	0.4671	0.0096	0.4095
$mean_{AUC}$	0.1141	0.4603	0.0094	0.4162
$mean_{AUCH}$	0.1141	0.4593	0.0094	0.4172
$mean_{KS}$	0.1140	0.4661	0.0095	0.4105
$mean_{MER}$	0.1142	0.4569	0.0093	0.4196
$mean_{MWL}$	0.1143	0.4494	0.0092	0.4271
$mean_{SpSe99}$	0.1169	0.4660	0.0066	0.4105
$mean_{SeSp99}$	0.1137	0.4838	0.0098	0.3927
$mean_{\frac{1}{AUC}}$	0.1141	0.4573	0.0093	0.4192
$mean_M$	0.1147	0.4763	0.0088	0.4003
$mean_{\frac{1}{M}}$	0.1137	0.4499	0.0098	0.4266
$mean_{RP}$	0.1135	0.4552	0.0100	0.4213

Table 4.33: Confusion matrix for all models at cut-off=0.5 for churn analysis

Model	TP	TN	FP	FN
Cox	0.0450	0.8673	0.0785	0.0092
RSF	0.1048	0.3736	0.0187	0.5031
CT	0.0339	0.8715	0.0896	0.0050
LR	0.0450	0.8673	0.0785	0.0092
$mean$	0.1123	0.4833	0.0112	0.3932
$mean_H$	0.1121	0.4895	0.0113	0.3871
$mean_{Gini}$	0.1121	0.4898	0.0113	0.3867
$mean_{AUC}$	0.1122	0.4869	0.0113	0.3896
$mean_{AUCH}$	0.1122	0.4865	0.0113	0.3901
$mean_{KS}$	0.1121	0.4891	0.0113	0.3875
$mean_{MER}$	0.1123	0.4704	0.0111	0.4061
$mean_{MWL}$	0.1121	0.4666	0.0114	0.4099
$mean_{SpSe99}$	0.1134	0.4892	0.0101	0.3873
$mean_{SeSp99}$	0.1121	0.4982	0.0114	0.3783
$mean_{\frac{1}{AUC}}$	0.1121	0.4754	0.0113	0.4011
$mean_M$	0.1125	0.5121	0.0109	0.3644
$mean_{\frac{1}{M}}$	0.1124	0.4677	0.0111	0.4088
$mean_{RP}$	0.1123	0.4774	0.0111	0.3991

Table 4.34: Confusion matrix for all models at cut-off=0.6 for churn analysis

Model	TP	TN	FP	FN
Cox	0.0361	0.8717	0.0873	0.0048
RSF	0.1042	0.3894	0.0193	0.4871
CT	0.0334	0.8715	0.0896	0.0050
LR	0.0361	0.8717	0.0873	0.0048
$mean$	0.1092	0.4952	0.0143	0.3813
$mean_H$	0.1011	0.5195	0.0224	0.3571
$mean_{Gini}$	0.1101	0.5049	0.0133	0.3716
$mean_{AUC}$	0.1101	0.4971	0.0134	0.3794
$mean_{AUCH}$	0.1097	0.4967	0.0138	0.3799
$mean_{KS}$	0.1100	0.5039	0.0135	0.3726
$mean_{MER}$	0.1088	0.4909	0.0147	0.3856
$mean_{MWL}$	0.1076	0.4881	0.0159	0.3885
$mean_{SpSe99}$	0.1104	0.5044	0.0131	0.3721
$mean_{SeSp99}$	0.1110	0.5099	0.0125	0.3666
$mean_{\frac{1}{AUC}}$	0.1081	0.4917	0.0154	0.3848
$mean_M$	0.1099	0.5365	0.0135	0.3400
$mean_{\frac{1}{M}}$	0.1058	0.4857	0.0177	0.3908
$mean_{RP}$	0.1089	0.4939	0.0146	0.3827

Table 4.35: Confusion matrix for all models at cut-off=0.7 for churn analysis

Model	**TP**	**TN**	**FP**	**FN**
Cox	0.0306	0.8738	0.0929	0.0027
RSF	0.1031	0.3987	0.0204	0.4778
CT	0.0339	0.8715	0.0896	0.0050
LR	0.0306	0.8738	0.0929	0.0027
$mean$	0.1014	0.5071	0.0221	0.3694
$mean_H$	0.1011	0.5195	0.0224	0.3571
$mean_{Gini}$	0.1010	0.5202	0.0225	0.3563
$mean_{AUC}$	0.1014	0.5090	0.0221	0.3675
$mean_{AUCH}$	0.1014	0.5087	0.0221	0.3678
$mean_{KS}$	0.1013	0.5140	0.0222	0.3625
$mean_{MER}$	0.1014	0.5052	0.0221	0.3713
$mean_{MWL}$	0.1015	0.4669	0.0220	0.3796
$mean_{SpSe99}$	0.1011	0.5196	0.0223	0.3569
$mean_{SeSp99}$	0.1006	0.5381	0.0229	0.3384
$mean_{\frac{1}{AUC}}$	0.1014	0.5058	0.0221	0.3707
$mean_M$	0.1011	0.5652	0.0224	0.3113
$mean_{\frac{1}{M}}$	0.1015	0.4976	0.0219	0.3789
$mean_{RP}$	0.1014	0.5013	0.0221	0.3752

Table 4.36: Confusion matrix for all models at cut-off=0.8 for churn analysis

Time	**AUC Cox**	**AUC RSF**
T_3	0.7112	0.6205
T_6	0.7375	0.6776
T_9	0.7375	0.7086
T_{12}	0.7392	0.7260
T_{15}	0.7372	0.7461
T_{18}	0.7282	0.7567
T_{21}	0.7305	0.7636
T_{24}	0.7241	0.7894
T_{27}	0.7207	0.8038
T_{30}	0.7249	0.8205
T_{33}	0.7282	0.8291
T_{36}	0.7299	0.8346
T_{39}	0.7310	0.8422
T_{42}	0.7319	0.8481
T_{45}	0.7324	0.8530
T_{48}	0.7327	0.8564
T_{51}	0.7317	0.8613
T_{54}	0.7316	0.8654
T_{57}	0.7312	0.8678
T_{60}	0.7261	0.8701
T_{63}	0.7237	0.8726
T_{66}	0.7189	0.8742
T_{69}	0.7147	0.8756
T_{72}	0.7120	0.8763

Table 4.37: AUC for Cox and RSF at different points in time for churn analysis

Time	C-index Cox	C-index RSF
T_3	0.9570	0.9810
T_6	0.9640	0.9760
T_9	0.9670	0.9720
T_{12}	0.9470	0.9460
T_{15}	0.9460	0.9430
T_{18}	0.9480	0.9410
T_{21}	0.9490	0.9360
T_{24}	0.9470	0.9310
T_{27}	0.9450	0.9280
T_{30}	0.9410	0.9220
T_{33}	0.9380	0.9190
T_{36}	0.9360	0.9160
T_{39}	0.9320	0.9110
T_{42}	0.9310	0.9100
T_{45}	0.9300	0.9080
T_{48}	0.9270	0.9050
T_{51}	0.9270	0.9020
T_{54}	0.9210	0.8980
T_{57}	0.9180	0.8940
T_{60}	0.9150	0.8910
T_{63}	0.9100	0.8860
T_{66}	0.9080	0.8840
T_{69}	0.9040	0.8810
T_{72}	0.9003	0.8790

Table 4.38: C-index of Cox and RSF at different points in time

Model	Min	Q_1	Median	Mean	Q_3	Max	$\frac{1}{Variance}$	CV
Cox	0.1409	0.1590	0.1607	2.0000	0.1607	145.6000	0.0162	3.9272
RSF	0.1409	0.1622	0.1630	2.0000	0.1630	108.7000	0.0224	3.3405

Table 4.39: Summary statistics of Mahalanobis distance for single models for churn analysis (T=3)

Model	H	Gini	AUC	AUCH	KS	MER	MWL	SpSe99	SeSp99
Cox	0.2069	0.4223	0.7112	0.7289	0.3575	0.1086	0.1391	0.4665	0.1789
RSF	0.1449	0.2409	0.6205	0.6295	0.2090	0.1073	0.1712	0.5241	0.2210

Table 4.40: Performance measure for single models for churn analysis (T=3)

Model	**Min**	Q_1	**Median**	**Mean**	Q_3	**Max**	$\frac{1}{Variance}$	**CV**
$mean$	0.1409	0.1620	0.1645	2.0000	0.1645	101.7000	0.0199	3.5435
$mean_C$	0.1409	0.1620	0.1645	2.0000	0.1645	100.9000	0.0200	3.5378
$mean_{\frac{1}{C}}$	0.1409	0.1620	0.1645	2.0000	0.1645	102.4000	0.0198	3.5492
$mean_H$	0.1409	0.1616	0.1641	2.0000	0.1641	111.3000	0.0190	3.6256
$mean_{Gini}$	0.1409	0.1614	0.1638	2.0000	0.1638	116.4000	0.0186	3.6701
$mean_{AUC}$	0.1409	0.1619	0.1643	2.0000	0.1643	105.5000	0.0196	3.5752
$mean_{AUCH}$	0.1409	0.1619	0.1643	2.0000	0.1643	105.8000	0.0195	3.5775
$mean_{KS}$	0.1409	0.1614	0.1639	2.0000	0.1639	115.8000	0.0186	3.6648
$mean_{MER}$	0.1409	0.1620	0.1645	2.0000	0.1645	102.0000	0.0199	3.5463
$mean_{MWL}$	0.1409	0.1625	0.1646	2.0000	0.1646	95.7000	0.0205	3.4965
$mean_{SpSe99}$	0.1409	0.1625	0.1646	2.0000	0.1646	98.3400	0.0202	3.5168
$mean_{SeSp99}$	0.1409	0.1620	0.1645	2.0000	0.1645	101.2000	0.0200	3.5400
$mean_{\frac{1}{AUC}}$	0.1409	0.1625	0.1646	2.0000	0.1646	97.7300	0.0203	3.5122
$mean_M$	0.1409	0.1585	0.1610	2.0000	0.1610	93.3200	0.0177	3.7588
$mean_{\frac{1}{M}}$	0.1409	0.1636	0.1659	2.0000	0.1659	112.3000	0.0215	3.4100
$mean_{RP}$	0.1409	0.1585	0.1610	2.0000	0.1610	93.5300	0.01774	3.7611

Table 4.41: Summary statistics of Mahalanobis distance for all time-dependent average models for churn analysis (T=3)

Model	H	Gini	AUC	AUCH	KS	MER	MWL	SpSe99	SeSp99
$mean$	0.2006	0.4048	0.7024	0.7150	0.3166	0.1055	0.1479	0.4605	0.1975
$mean_C$	0.2008	0.4045	0.7022	0.7149	0.3162	0.1056	0.1480	0.4605	0.1953
$mean_{\frac{1}{C}}$	0.2002	0.4052	0.7026	0.7152	0.3167	0.1056	0.1479	0.4604	0.1976
$mean_H$	0.2017	0.4121	0.7060	0.7190	0.3252	0.1056	0.1460	0.4608	0.1923
$mean_{Gini}$	0.2045	0.4170	0.7085	0.7220	0.3333	0.1055	0.1443	0.4606	0.1895
$mean_{AUC}$	0.2013	0.4085	0.7043	0.7172	0.3199	0.1056	0.1472	0.4603	0.1952
$mean_{AUCH}$	0.2018	0.4088	0.7044	0.7175	0.3218	0.1056	0.1468	0.4603	0.1952
$mean_{KS}$	0.2049	0.4167	0.7083	0.7222	0.3336	0.1055	0.1443	0.4606	0.1895
$mean_{MER}$	0.2005	0.4050	0.7025	0.7150	0.3168	0.1055	0.1479	0.4605	0.1976
$mean_{MWL}$	0.1990	0.3990	0.6995	0.7121	0.3119	0.1055	0.1490	0.4611	0.1985
$mean_{SpSe99}$	0.2002	0.4021	0.7010	0.7135	0.3154	0.1055	0.1482	0.4607	0.1976
$mean_{SeSp99}$	0.2007	0.4046	0.7023	0.7149	0.3164	0.1056	0.1480	0.4605	0.1953
$mean_{\frac{1}{AUC}}$	0.1998	0.4010	0.7005	0.7131	0.3154	0.1056	0.1482	0.4610	0.1976
$mean_M$	0.2195	0.4082	0.7041	0.7191	0.3160	0.1089	0.1481	0.4605	0.1728
$mean_{\frac{1}{M}}$	0.1910	0.4018	0.7009	0.7119	0.3172	0.1059	0.1478	0.4605	0.1930
$mean_{RP}$	0.2195	0.4086	0.7043	0.7192	0.3164	0.1089	0.1480	0.4604	0.1726

Table 4.42: Performance measure for all model combinations for churn analysis (T=3)

Model	AUC	CI
Cox	**0.7112**	0.7029-0.7194
RSF	0.6205	0.6112-0.6298
$mean$	0.7024	0.6939-0.7109
$mean_C$	0.7022	0.6937-0.7107
$mean_{\frac{1}{C}}$	0.7026	0.6941-0.7111
$mean_H$	0.7060	0.6976-0.7145
$mean_{Gini}$	0.7085	0.7001-0.7169
$mean_{AUC}$	0.7043	0.6958-0.7127
$mean_{AUCH}$	0.7044	0.6959-0.7129
$mean_{KS}$	0.7083	0.6999-0.7168
$mean_{MER}$	0.7025	0.6940-0.7110
$mean_{MWL}$	0.6995	0.6910-0.7080
$mean_{SpSe99}$	0.7010	0.6925-0.7095
$mean_{SeSp99}$	0.7023	0.6938-0.7108
$mean_{\frac{1}{AUC}}$	0.7005	0.6920-0.7090
$mean_M$	0.7041	0.6956-0.7127
$mean_{\frac{1}{M}}$	0.7009	0.6924-0.7093
$mean_{RP}$	0.7043	0.6958-0.7129

Table 4.43: AUC and relative CI for all models for churn analysis (T=3)

Model	TP	TN	FP	FN
Cox	0.1153	0.0005	0.0065	0.8761
RSF	0.1136	0.0003	0.0099	0.8763
$mean$	0.1161	0.0002	0.0074	0.8763
$mean_C$	0.1161	0.0002	0.0074	0.8763
$mean_{\frac{1}{C}}$	0.1162	0.0002	0.0729	0.8763
$mean_H$	0.1161	0.0003	0.0074	0.8762
$mean_{Gini}$	0.1161	0.0003	0.0074	0.8763
$mean_{AUC}$	0.1162	0.0002	0.0073	0.8763
$mean_{AUCH}$	0.1162	0.0002	0.0073	0.8763
$mean_{KS}$	0.1161	0.0003	0.0074	0.8763
$mean_{MER}$	0.1162	0.0002	0.0073	0.8763
$mean_{MWL}$	0.1158	0.0002	0.0077	0.8763
$mean_{SpSe99}$	0.1160	0.0002	0.0075	0.8763
$mean_{SeSp99}$	0.1161	0.0002	0.0074	0.8763
$mean_{\frac{1}{AUC}}$	0.1157	0.0002	0.0077	0.8763
$mean_M$	0.1168	0.0002	0.0067	0.8763
$mean_{\frac{1}{M}}$	0.1131	0.0005	0.0104	0.8760
$mean_{RP}$	0.1168	0.0002	0.0067	0.8763

Table 4.44: Confusion matrix for all models at cut-off=0.5 for churn analysis (T=3)

4.3 Summary

In this chapter we have presented several examples of application of our method combianzione models. We have shown that it is possible to apply the method to all the problems of analysis in which you have a binary response variable through the use of two different datasets, the first concerning the prediction of probability of default (section 4.3) and the second, which takes into account the churn analysis (section 4.3.1). In conclusion for PD example we can say that we get excellent performance in terms of accuracy and correctly classified observations, when we consider the time-independent approach, while the only time-dependent models commit a high percentage of error in the classification of statistical units, however, must consider that they have a good ability to predict the units classified negatively compared with probability of correctly classify positive units.
If we look at the approach to the time-dependent case of churn analysis we can say that the probability of having the observations correctly classisìficate is almost the same as which ones to make mistakes in the classification.It must, however, be observed that the models, in this case, have a high probability to correctly classify the negative units. Regarding the time-dependent approach we can say that the probability of making errors in classification is very high, then even in this case works best time-independet approaches.
If we want to evaluate the performance of the models in terms of AUC, we can say that the models with average weights constructed with the Mahalanobis distance have a level of performance higher than all the others. Therefore, the method identified, that is to use measures of distance as weights combination of forecasts obtained from the individual models and overcomes the problem of the inconsistency of the measures of performance, it seems to provide added value compared to other methods.

Model	**TP**	**TN**	**FP**	**FN**
Cox	0.1017	0.0013	0.0132	0.8752
RSF	0.1109	0.0004	0.0126	0.8761
$mean$	0.1113	0.0012	0.0122	0.8753
$mean_C$	0.1113	0.0012	0.0122	0.8753
$mean_{\frac{1}{C}}$	0.1113	0.0012	0.0122	0.8753
$mean_H$	0.1119	0.0012	0.0116	0.8753
$mean_{Gini}$	0.1121	0.0012	0.0114	0.8753
$mean_{AUC}$	0.1115	0.0012	0.0120	0.8753
$mean_{AUCH}$	0.1115	0.0012	0.0120	0.8753
$mean_{KS}$	0.1121	0.0012	0.0114	0.8753
$mean_{MER}$	0.1113	0.0012	0.0122	0.8753
$mean_{MWL}$	0.1107	0.0012	0.0127	0.8753
$mean_{SpSe99}$	0.1111	0.0012	0.0124	0.8753
$mean_{SeSp99}$	0.1113	0.0012	0.0122	0.8753
$mean_{\frac{1}{AUC}}$	0.1109	0.0012	0.0126	0.8753
$mean_M$	0.1126	0.0011	0.0109	0.8754
$mean_{\frac{1}{M}}$	0.1076	0.0015	0.0159	0.8751
$mean_{RP}$	0.1026	0.0011	0.0109	0.8754

Table 4.45: Confusion matrix for all models at cut-off=0.6 for churn analysis (T=3)

Model	TP	TN	FP	FN
Cox	0.1062	0.0027	0.0173	0.8739
RSF	0.1044	0.0032	0.0191	0.8733
$mean$	0.1060	0.0026	0.0175	0.8739
$mean_C$	0.1059	0.0026	0.0175	0.8739
$mean_{\frac{1}{C}}$	0.1061	0.0026	0.0174	0.8739
$mean_H$	0.1055	0.0026	0.0180	0.8739
$mean_{Gini}$	0.1059	0.0026	0.0176	0.8739
$mean_{AUC}$	0.1063	0.0026	0.0171	0.8739
$mean_{AUCH}$	0.1063	0.0025	0.0171	0.8740
$mean_{KS}$	0.1060	0.0026	0.0175	0.8739
$mean_{MER}$	0.1061	0.0026	0.0174	0.8739
$mean_{MWL}$	0.1059	0.0026	0.0176	0.8739
$mean_{SpSe99}$	0.1059	0.0026	0.0176	0.8739
$mean_{SeSp99}$	0.1059	0.0026	0.0175	0.8739
$mean_{\frac{1}{AUC}}$	0.1059	0.0026	0.0176	0.8739
$mean_M$	0.1088	0.0024	0.0147	0.8741
$mean_{\frac{1}{M}}$	0.1035	0.0029	0.0199	0.8736
$mean_{RP}$	0.1088	0.0024	0.0147	0.8741

Table 4.46: Confusion matrix for all models at cut-off=0.7 for churn analysis (T=3)

Model	**TP**	**TN**	**FP**	**FN**
Cox	0.1030	0.0063	0.0205	0.8702
RSF	0.1030	0.0043	0.0205	0.8722
$mean$	0.1015	0.0043	0.0219	0.8723
$mean_C$	0.1016	0.0043	0.0219	0.8723
$mean_{\frac{1}{C}}$	0.1015	0.0043	0.0219	0.8722
$mean_H$	0.1014	0.0043	0.0221	0.8722
$mean_{Gini}$	0.1012	0.0044	0.0223	0.8721
$mean_{AUC}$	0.1015	0.0044	0.0220	0.8721
$mean_{AUCH}$	0.1015	0.0044	0.0220	0.8721
$mean_{KS}$	0.1012	0.0044	0.0223	0.8721
$mean_{MER}$	0.1015	0.0043	0.0219	0.8722
$mean_{MWL}$	0.1020	0.0043	0.0215	0.8723
$mean_{SpSe99}$	0.1020	0.0043	0.0215	0.8723
$mean_{SeSp99}$	0.1016	0.0043	0.0219	0.8723
$mean_{\frac{1}{AUC}}$	0.1020	0.0043	0.0215	0.8723
$mean_M$	0.1062	0.0037	0.0173	0.8729
$mean_{\frac{1}{M}}$	0.1007	0.0057	0.0227	0.8708
$mean_{RP}$	0.1062	0.0037	0.0173	0.8723

Table 4.47: Confusion matrix for all models at cut-off=0.8 for churn analysis (T=3)

Chapter 5

Conclusion and future research

5.1 Summary and future development

Literature on combining forecasts is very our extensive, the common thread of most of the work is that through the combination of models is improved the accuracy of predictions. In addition, a review of the literature shows that the simplest methods of combination provide better results than more complex. In fact, in many cases it was possible to improve the performance of the models through the simple medium of forecasts obtained by single models. In chapter 1 we have presented two main categories of models: time-dependent and time-independent. With regard to the first category of models, we focus our attention on the model of Cox proportional hazards model and Random Survival Forest. With regard to the time-independent model,we will discuss the Logistic Regression and Classification Trees.
We have described, also, the performance measures chosen to assess the model in terms of discriminatory power and predicted capability. Also the performance measures are of double nature: for time-dependent and time-independent models. The characteristic of all performance measures used is that are releted with the Roc curve.
In the Chapter 2 we have presented the concept of forecast combination. The forecast combination are born at the moment when we can have two or more predictions of the same event. Usually, the interest of the analyst is to identify the *best* forecast. When this is identified is used by the analyst while the others forecasts are discarded. However, we must consider that the forecasts discarded may contain useful information especially when the purpose is to determine the best possible forecast. We can desume that if we take into account all the forecasts and we do a combination we can come to a better and more robust forecasting performance than what we can achieve in the case of

predictions generated by individual models. The existing literature concerning forecasts combination suggests that even when it is possible to identify the best model may be convenient to combine forecasts since the combination could lead to an increase in the accuracy of the forecast. We ave described many aspects that emphasize the usefulness of recourse to forecasts combination with respect to forecasts from individual models, for example forecast combinato leads to overcomes misspecification problem. In presence of two or more forecasts of the same variable the first question to answer is whether to combine or not to combine and we treat this aspect in Chapter 2. We emphasized that there are different combinato schemes. In the literature they are found in the majority of cases, patterns of combination of the linear type. In essence, if we assume as a loss function the MSE proceed to a combination of linear forecast models. However, you can also encounter cases where build combinations of non-linear type or time-varying combination methods. In our opinion the starting point, when thinking whether to build or not predictions combined, is the statement of Clemen (1989): "Combining forecasts has been shown to be practical, economical and useful". In fact, the forecast combination allows us not only to use a broader range of information than what we allow the individual models but makes possible to overcome, for example, the misspecification of single models. We will present the method studied by us to combine models of different nature in order to improve the prediction of the PD. Specifically, we will see that we determined averaged models using weights as performance measures and distance measures, to our knowledge, is an innovation in the literature of forecast combination. The goal of our work is to be able to correctly classify statistics units. We choose two doefferent approaches: time-dependent and time-independent, in both cases we apodo the same combination scheme as described in Chapter 3. The initial idea was to compare the time-dependent models with those time-independent in terms of AUC to establish the single model that performs better and then use the same measure as weight for the combination. The choice of using a measure of performance as weight arises from the consideration that the literature suggests that use weights without any constraint leads to combination methods more accurate than those that pose conditions on the characteristics of the weights. Since the performance measures can be incoherent, in particular the AUC, we decided to calculate and use as weight distance measure, specifically we have focused our attention on the Mahalanobis distance. The choice of this distance measure is due to the following reflections. First, in multivariate statistics this measure is used to capture the similarity between two objects and we are interested to measure the similarity that exists between the response variable observed and forecast the

same made by the different individual models considered. Second, we wanted a measure that could be calculated in the same way for all the models under study view of the different nature of these models. Last but not least, we wanted a simple measure to calculate to meet one of the main characteristics of the methods of combination. We determine the combination models as a weigthed average models in which the weigths are performance and distance measures. We evaluate the performance of all models considered in our study, single and combained, based on the units correctly classified. Therefore we determine the confusion matrix and we can see the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). In this case the best model is the one that will have a higher percentage of units are correctly classified and therefore true positives and true negatives.
In Chapter 4 we presented two examples of application of the method of combination designed to provide empirical evidence of the results that this method leads. We presented two applications of the method to problems of prediction of dichotomous data. In the first case we have applied the method to a problem of credit risk with the aim to predict the probability of default, while in the second case we have focused our attention on the churn analysis with the aim to predict the likelihood of abandonment. Otteuti The results have been described in detail in Chapter 4, it is pointed out that empirical evidence suggests that the approach seems to work better than time-independent time-dependent. Weights that determine the best combination models are those based on the Mahalanobis distance.
As regards future developments, the idea that can be realized is that which is based on the use of concordance measures for dichotomous data for the evaluation of the models. The concordance should be calculated between the response variable of interest and forecasts of that obtained through the combined models with certain weights introduced in this study.

Appendix A

Confusion matrix for simulated data

Model	TP	TN	FP	FN
Cox	0.0013	0.9623	0.0094	0.0270
RSF	0	0.9043	0.0108	0.0795
CT	0	0.9097	0.0108	0.0121
LR	0.0013	0.9690	0.0994	0.0202
$mean$	0	0.9596	0.0108	0 .0296
$mean_H$	0	0.9367	0.0108	0.0526
$mean_{Gini}$	0	0.9420	0.0108	0.0472
$mean_{AUC}$	0	0.9528	0.0108	0.0364
$mean_{AUCH}$	0	0.9582	0.0108	0.0310
$mean_{KS}$	0	0.9420	0.0108	0.0472
$mean_{MER}$	0.0013	0.9650	0.0013	0.0243
$mean_{MWL}$	0	0.9879	0.0108	0.0013
$mean_{SpSe99}$	0	0.9367	0.0108	0.0526
$mean_{SeSp99}$	0	0.9191	0.0108	0.0101
$mean_{\frac{1}{AUC}}$	0.0013	0.9650	0.0094	0.0243
$mean_M$	0	0.9232	0.0108	0.0660
$mean_{\frac{1}{M}}$	0	0.9636	0.0108	0.0256
$mean_{RP}$	0	0.9636	0.0108	0.0256

Table A.1: Confusion matrix for all models at cut-off=0.7 (P=0.01)

Model	TP	TN	FP	FN
Cox	0.0027	0.9474	0.0243	0.0256
RSF	0	0.8881	0.0270	0.0849
CT	0.0013	0.8949	0.0256	0.0782
LR	0.0013	0.9528	0.0256	0.0202
$mean$	0	0.9730	0.0270	0
$mean_H$	0	0.9730	0.0270	0
$mean_{Gini}$	0	0.9730	0.0270	0
$mean_{AUC}$	0	0.9730	0.0270	0
$mean_{AUCH}$	0	0.9730	0.0270	0
$mean_{KS}$	0	0.9730	0.0270	0
$mean_{MER}$	0	0.9730	0.0270	0
$mean_{MWL}$	0	0.9730	0.0270	0
$mean_{SpSe99}$	0.0027	0.9730	0.0243	0
$mean_{SeSp99}$	0.0040	0.9730	0.0229	0
$mean_{\frac{1}{AUC}}$	0	0.9730	0.0270	0
$mean_M$	0	0.9730	0.0270	0
$mean_{\frac{1}{M}}$	0	0.9730	0.0270	0
$mean_{RP}$	0	0.9730	0.0270	0

Table A.2: Confusion matrix for all models at cut-off=0.7 (P=0.02)

Model	TP	TN	FP	FN
Cox	0.0013	0.9340	0.0377	0.0270
RSF	0.0027	0.8787	0.0364	0.0822
CT	0.0054	0.8868	0.0337	0.0741
LR	0	0.9394	0.0391	0.0216
$mean$	0	0.9609	0.0391	0
$mean_H$	0.0013	0.9609	0.0270	0.0377
$mean_{Gini}$	0	0.9609	0.0391	0
$mean_{AUC}$	0	0.9609	0.0391	0
$mean_{AUCH}$	0	0.9609	0.0391	0
$mean_{KS}$	0	0.9609	0.0391	0
$mean_{MER}$	0	0.9609	0.0391	0
$mean_{MWL}$	0	0.9609	0.0391	0
$mean_{SpSe99}$	0.0013	0.9609	0.0377	0
$mean_{SeSp99}$	0.0040	0.9609	0.0350	0
$mean_{\frac{1}{AUC}}$	0	0.9609	0.0391	0
$mean_M$	0	0.0148	0.0387	0.9461
$mean_{\frac{1}{M}}$	0	0.9609	0.0391	0
$mean_{RP}$	0.0391	0	0	0.9609

Table A.3: Confusion matrix for all models at cut-off=0.7 (P=0.03)

Model	TP	TN	FP	FN
Cox	0.0013	0.9461	0.0256	0.0270
RSF	0.0040	0.8922	0.0229	0.0809
CT	0.0027	0.8962	0.0243	0.0741
LR	0	0.9515	0.0256	0.0216
$mean$	0	0.9730	0.0270	0
$mean_H$	0	0.9730	0.0270	0
$mean_{Gini}$	0	0.9730	0.0270	0
$mean_{AUC}$	0	0.9730	0.0270	0
$mean_{AUCH}$	0	0.9730	0.0270	0
$mean_{KS}$	0	0.9730	0.0270	0
$mean_{MER}$	0	0.9730	0.0270	0
$mean_{MWL}$	0	0.9730	0.0270	0
$mean_{SpSe99}$	0	0.9730	0.0270	0
$mean_{SeSp99}$	0	0.9730	0.0270	0
$mean_{\frac{1}{AUC}}$	0	0.9730	0.0270	0
$mean_M$	0	0.9730	0.0270	0
$mean_{\frac{1}{M}}$	0	0.9730	0.0270	0
$mean_{RP}$	0	0.9730	0.0270	0

Table A.4: Confusion matrix for all models at cut-off=0.7 (P=0.04)

Model	TP	TN	FP	FN
Cox	0.0013	0.9380	0.0337	0.0270
RSF	0.0013	0.8814	0.0337	0.0836
CT	0.0040	0.8895	0.0310	0.0755
LR	0.0013	0.9447	0.0337	0.0202
$mean$	0	0.9650	0.0350	0
$mean_H$	0.0013	0.9650	0.0337	0
$mean_{Gini}$	0	0.9650	0.0350	0
$mean_{AUC}$	0	0.9650	0.0350	0
$mean_{AUCH}$	0	0.9650	0.0350	0
$mean_{KS}$	0	0.9650	0.0350	0
$mean_{MER}$	0	0.9650	0.0350	0
$mean_{MWL}$	0	0.9650	0.0350	0
$mean_{SpSe99}$	0	0.9650	0.0350	0
$mean_{SeSp99}$	0.0054	0.9650	0.0296	0
$mean_{\frac{1}{AUC}}$	0	0.9650	0.0350	0
$mean_M$	0	0.9650	0.0350	0
$mean_{\frac{1}{M}}$	0	0.9650	0.0350	0
$mean_{RP}$	0	0.9650	0.0350	0

Table A.5: Confusion matrix for all models at cut-off=0.7 (P=0.05)

Model	TP	TN	FP	FN
Cox	0.0013	0.9245	0.0472	0.0270
RSF	0.0040	0.8706	0.0445	0.0809
CT	0.0081	0.8801	0.0404	0.0714
LR	0.0013	0.9313	0.0472	0.0202
$mean$	0	0.9515	0.0485	0
$mean_H$	0.0485	0	0.9515	0
$mean_{Gini}$	0.0418	0.9515	0.0067	0
$mean_{AUC}$	0.0081	0.9515	0.0404	0
$mean_{AUCH}$	0.0081	0.9515	0.0404	0
$mean_{KS}$	0.0418	0.9515	0.0067	0
$mean_{MER}$	0	0.9515	0.0485	0
$mean_{MWL}$	0	0.9515	0.0485	0
$mean_{SpSe99}$	0.0485	0.9515	0	0
$mean_{SeSp99}$	0.0485	0.9515	0	0
$mean_{\frac{1}{AUC}}$	0	0.9515	0.0485	0
$mean_M$	0.0040	0.9501	0.0445	0.0013
$mean_{\frac{1}{M}}$	0	0.9515	0.0485	0
$mean_{RP}$	0	0.9515	0.0485	0

Table A.6: Confusion matrix for all models at cut-off=0.7 (P=0.06)

Model	TP	TN	FP	FN
Cox	0	0.9003	0.0714	0.0283
RSF	0.0067	0.8504	0.0647	0.0782
CT	0.0081	0.8571	0.0633	0.0714
LR	0	0.9070	0.0714	0.0216
$mean$	0	0.9286	0.0714	0
$mean_H$	0.0580	0.9286	0.0135	0
$mean_{Gini}$	0.0216	0.9286	0.0499	0
$mean_{AUC}$	0	0.9286	0.0714	0
$mean_{AUCH}$	0	0.9286	0.0714	0
$mean_{KS}$	0.0202	0.9286	0.0512	0
$mean_{MER}$	0	0.9286	0.0714	0
$mean_{MWL}$	0	0.9286	0.0714	0
$mean_{SpSe99}$	0.0633	0.9285	0.0081	0
$mean_{SeSp99}$	0.0687	0.9285	0.0027	0
$mean_{\frac{1}{AUC}}$	0	0.285	0.0714	0
$mean_M$	0	0.9272	0.0714	0.0013
$mean_{\frac{1}{M}}$	0.0013	0.9286	0.0701	0
$mean_{RP}$	0.0013	0.9286	0.0701	0

Table A.7: Confusion matrix for all models at cut-off=0.7 (P=0.07)

Model	TP	TN	FP	FN
Cox	0.0027	0.8787	0.0930	0.0256
RSF	0.0081	0.8275	0.0876	0.0768
CT	0.0067	0.8315	0.0889	0.0728
LR	0	0.8827	0.0957	0.0216
$mean$	0	0.9043	0.0957	0
$mean_H$	0.0189	0.9043	0.0768	0
$mean_{Gini}$	0.0121	0.9043	0.0836	0
$mean_{AUC}$	0	0.9043	0.0957	0
$mean_{AUCH}$	0	0.9043	0.0957	0
$mean_{KS}$	0.0121	0.9043	0.0836	0
$mean_{MER}$	0	0.9043	0.0957	0
$mean_{MWL}$	0	0.9043	0.0957	0
$mean_{SpSe99}$	0.0283	0.9043	0.0674	0
$mean_{SeSp99}$	0.0431	0.9030	0.0526	0.0013
$mean_{\frac{1}{AUC}}$	0	0.9043	0.0957	0
$mean_M$	0.0054	0.9030	0.0903	0.0013
$mean_{\frac{1}{M}}$	0	0.9043	0.0957	0
$mean_{RP}$	0	0.9043	0.0957	0

Table A.8: Confusion matrix for all models at cut-off=0.7 (P=0.08)

Model	TP	TN	FP	FN
Cox	0.0040	0.8976	0.0741	0.0243
RSF	0.0094	0.8466	0.0687	0.0755
CT	0.0108	0.8531	0.0674	0.0687
LR	0.0027	0.9030	0.0755	0.0189
$mean$	0	0.9218	0.0782	0
$mean_H$	0	0.9218	0.0782	0
$mean_{Gini}$	0	0.9218	0.0782	0
$mean_{AUC}$	0	0.9218	0.0782	0
$mean_{AUCH}$	0	0.9218	0.0782	0
$mean_{KS}$	0	0.9218	0.0782	0
$mean_{MER}$	0	0.9218	0.0782	0
$mean_{MWL}$	0	0.9218	0.0782	0
$mean_{SpSe99}$	0	0.9218	0.0782	0
$mean_{SeSp99}$	0	0.9218	0.0782	0
$mean_{\frac{1}{AUC}}$	0	0.9218	0.0782	0
$mean_M$	0	0.9218	0.0782	0
$mean_{\frac{1}{M}}$	0	0.9218	0.0782	0
$mean_{RP}$	0	0.9218	0.0782	0

Table A.9: Confusion matrix for all models at cut-off=0.7 (P=0.09)

Model	TP	TN	FP	FN
Cox	0.0027	0.8787	0.0930	0.0256
RSF	0.0108	0.8302	0.0849	0.0741
CT	0.0081	0.8329	0.0876	0.0714
LR	0.0027	0.8854	0.0957	0.0189
$mean$	0	0.9043	0.0782	0
$mean_H$	0.0148	0.9043	0.0809	0
$mean_{Gini}$	0.0040	0.9043	0.0916	0
$mean_{AUC}$	0	0.9043	0.0782	0
$mean_{AUCH}$	0	0.9043	0.0782	0
$mean_{KS}$	0.0040	0.9043	0.0916	0
$mean_{MER}$	0	0.9043	0.0782	0
$mean_{MWL}$	0	0.9043	0.0782	0
$mean_{SpSe99}$	0.0189	0.9043	0.0768	0
$mean_{SeSp99}$	0.0296	0.9043	0.0660	0
$mean_{\frac{1}{AUC}}$	0	0.9043	0.0782	0
$mean_M$	0.0013	0.9043	0.0943	0
$mean_{\frac{1}{M}}$	0.0782	0.9016	0	0.0027
$mean_{RP}$	0	0.9043	0.0782	0

Table A.10: Confusion matrix for all models at cut-off=0.7 (P=0.10)

Model	TP	TN	FP	FN
Cox	0.0027	0.9346	0.1631	0.0256
RSF	0.0121	0.7615	0.1536	0.0728
CT	0.0148	0.7695	0.1509	0.0647
LR	0.0040	0.8167	0.1617	0.0175
$mean$	0.1334	0.8315	0.0323	0.0027
$mean_H$	0.1658	0.6456	0	0.1887
$mean_{Gini}$	0.1658	0.7210	0	0.1132
$mean_{AUC}$	0.1644	0.8208	0.0013	0.0135
$mean_{AUCH}$	0.1644	0.8235	0.0013	0.0108
$mean_{KS}$	0.1658	0.7453	0	0.0889
$mean_{MER}$	0	0.8342	0.1658	0
$mean_{MWL}$	0	0.8342	0.1658	0
$mean_{SpSe99}$	0.1658	0.6321	0	0.2022
$mean_{SeSp99}$	0.1658	0.5916	0	0.2426
$mean_{\frac{1}{AUC}}$	0.0310	0.8329	0.1348	0.0013
$mean_M$	0.0216	0	0.0013	0.7776
$mean_{\frac{1}{M}}$	0.1240	0.8302	0.0418	0.0040
$mean_{RP}$	0.1348	0.8302	0.0310	0.0040

Table A.11: Confusion matrix for all models at cut-off=0.7 (P=0.15)

Model	TP	TN	FP	FN
Cox	0.0067	0.79116	0.15911	0.0216
RSF	0.0189	0.7466	0.1685	0.0660
CT	0.0202	0.7534	0.1671	0.0593
LR	0.0040	0.7951	0.1833	0.0175
$mean$	0.0094	0.8019	0.1780	0.0108
$mean_H$	0.0162	0.7884	0.1712	0.0243
$mean_{Gini}$	0.0458	0.7197	0.1415	0.0930
$mean_{AUC}$	0.0175	0.7978	0.1698	0.0148
$mean_{AUCH}$	0.0108	0.7978	0.1765	0.0256
$mean_{KS}$	0.0175	0.7871	0.1698	0.0889
$mean_{MER}$	0.0108	0.8019	0.1765	0.0108
$mean_{MWL}$	0.0067	0.8032	0.1806	0.0094
$mean_{SpSe99}$	0	0.8127	0.1873	0
$mean_{SeSp99}$	0	0.8127	0.1873	0
$mean_{\frac{1}{AUC}}$	0.0054	0.8032	0.1819	0.0094
$mean_M$	0.0148	0.7210	0.1725	0.0916
$mean_{\frac{1}{M}}$	0.0040	0.7628	0.1833	0.0499
$mean_{RP}$	0.0027	0.7776	0.1846	0.0350

Table A.12: Confusion matrix for all models at cut-off=0.7 (P=0.20)

Model	TP	TN	FP	FN
Cox	0.0094	0.6927	0.2790	0.0189
RSF	0.0310	0.6577	0.2574	0.0539
CT	0.0229	0.6550	0.2655	0.0566
LR	0.0108	0.7008	0.2776	0.0108
$mean$	0.2884	0.6792	0	0.0323
$mean_H$	0.2884	0.2372	0	0.4744
$mean_{Gini}$	0.2884	0.2830	0	0.4286
$mean_{AUC}$	0.2884	0.5768	0	0.1348
$mean_{AUCH}$	0.2884	0.5916	0	0.1199
$mean_{KS}$	0.2884	0.3059	0	0.4056
$mean_{MER}$	0	0.7116	0.2884	0
$mean_{MWL}$	0	0.7116	0.2884	0
$mean_{SpSe99}$	0.2884	0.4717	0	0.2399
$mean_{SeSp99}$	0.2884	0.4582	0	0.2534
$mean_{\frac{1}{AUC}}$	0.1860	0.7075	0	0.0040
$mean_M$	0.2857	0.6078	0.0027	0.1038
$mean_{\frac{1}{M}}$	0.2884	0.1186	0	0.5930
$mean_{RP}$	0.2655	0.5606	0.0229	0.1509

Table A.13: Confusion matrix for all models at cut-off=0.7 (P=0.25)

Model	TP	TN	FP	FN
Cox	0.0094	0.6846	0.2871	0.0189
RSF	0.0202	0.6388	0.2763	0.0647
CT	0.0189	0.6429	0.2766	0.0606
LR	0.0067	0.6887	0.2898	0.0148
$mean$	0.0472	0.7035	0.2493	0.0620
$mean_H$	0.2965	0.6415	0	0.4744
$mean_{Gini}$	0.2965	0.6495	0	0.0539
$mean_{AUC}$	0.2439	0.7008	0.0526	0.0027
$mean_{AUCH}$	0.2049	0.7035	0.0916	0
$mean_{KS}$	0.2965	0.6725	0	0.0310
$mean_{MER}$	0	0.7035	0.2965	0
$mean_{MWL}$	0	0.7035	0.2965	0
$mean_{SpSe99}$	0.2925	0.6806	0.0040	0.0229
$mean_{SeSp99}$	0.2911	0.6819	0.0054	0.0216
$mean_{\frac{1}{AUC}}$	0.0013	0.7035	0.2951	0
$mean_M$	0.2156	0.6698	0.0809	0.0337
$mean_{\frac{1}{M}}$	0.1644	0.7035	0.1321	0
$mean_{RP}$	0.1051	0.7035	0.1914	0

Table A.14: Confusion matrix for all models at cut-off=0.7 (P=0.30)

Model	TP	TN	FP	FN
Cox	0.0094	0.6456	0.3261	0.0189
RSF	0.0323	0.6119	0.3032	0.0526
CT	0.0267	0.6119	0.3086	0.0526
LR	0.0081	0.6509	0.3275	0.0135
$mean$	0.0094	0.6590	0.3261	0.0054
$mean_H$	0.0027	0.6361	0.3329	0.0283
$mean_{Gini}$	0.0121	0.5674	0.3235	0.0970
$mean_{AUC}$	0.0094	0.6146	0.3162	0.0499
$mean_{AUCH}$	0.0081	0.6469	0.3275	0.0175
$mean_{KS}$	0.0067	0.6051	0.3288	0.0593
$mean_{MER}$	0.0108	0.6509	0.3248	0.0135
$mean_{MWL}$	0.0108	0.6590	0.3248	0.0054
$mean_{SpSe99}$	0	0.6644	0.3356	0
$mean_{SeSp99}$	0	0.6644	0.3356	0
$mean_{\frac{1}{AUC}}$	0.0081	0.6590	0.3275	0.0054
$mean_M$	0.0660	0.5674	0.2695	0.0970
$mean_{\frac{1}{M}}$	0.0229	0.5243	0.3127	0.1402
$mean_{RP}$	0.0148	0.5377	0.3208	0.1267

Table A.15: Confusion matrix for all models at cut-off=0.7 (P=0.35)

Model	TP	TN	FP	FN
Cox	0.0121	0.5916	0.3801	0.0162
RSF	0.0431	0.5660	0.3491	0.0418
CT	0.0364	0.5647	0.3558	0.0431
LR	0.0094	0.5957	0.3827	0.0121
$mean$	0.3922	0.3989	0	0.2089
$mean_H$	0.3922	0.0795	0	0.5283
$mean_{Gini}$	0.3922	0.1024	0	0.5054
$mean_{AUC}$	0.3922	0.2520	0	0.3558
$mean_{AUCH}$	0.3922	0.2588	0	0.3491
$mean_{KS}$	0.3922	0.1119	0	0.4960
$mean_{MER}$	0	0.6078	0.3922	0
$mean_{MWL}$	0	0.6078	0.3922	0
$mean_{SpSe99}$	0.3922	0.1846	0	0.4232
$mean_{SeSp99}$	0.3922	0.0863	0	0.5216
$mean_{\frac{1}{AUC}}$	0.3895	0.5526	0.0027	0.0053
$mean_M$	0.0660	0.4879	0	0.1105
$mean_{\frac{1}{M}}$	0.3733	0.3679	0.0189	0 .2399
$mean_{RP}$	0.3720	0.3693	0.0202	0.2385

Table A.16: Confusion matrix for all models at cut-off=0.7 (P=0.40)

Model	TP	TN	FP	FN
Cox	0.0135	0.5296	0.4420	0.0148
RSF	0.0364	0.4960	0.4191	0.0485
CT	0.0350	0.5000	0.4205	0.0445
LR	0.0094	0.5323	0.4461	0.0121
$mean$	0.4394	0.1213	0.0162	0.4232
$mean_H$	0.4542	0.0620	0.0013	0.4825
$mean_{Gini}$	0.4555	0.0485	0	0.4960
$mean_{AUC}$	0.4501	0.0849	0.0054	0.4596
$mean_{AUCH}$	0.4488	0.0849	0.0067	0.4596
$mean_{KS}$	0.4542	0.0633	0.0013	0.4811
$mean_{MER}$	0.4191	0.2183	0.0364	0.3261
$mean_{MWL}$	0.4151	0.2345	0.0404	0.3099
$mean_{SpSe99}$	0.4286	0.2318	0.0270	0.3127
$mean_{SeSp99}$	0.3113	0.4286	0.1442	0.1159
$mean_{\frac{1}{AUC}}$	0.4245	0.2022	0.0309	0.3423
$mean_M$	0.4555	0.1253	0	0.4191
$mean_{\frac{1}{M}}$	0.4084	0.0970	0.0472	0.4474
$mean_{RP}$	0.4111	0.0916	0.0445	0.4528

Table A.17: Confusion matrix for all models at cut-off=0.7 (P=0.45)

Model	**TP**	**TN**	**FP**	**FN**
Cox	0.0162	0.4744	0.4973	0.0121
RSF	0.0391	0.4407	0.4744	0.0458
CT	0.0377	0.4447	0.4757	0.0418
LR	0.0148	0.4798	0.4987	0.0067
$mean$	0.5121	0.0984	0.0013	0.4420
$mean_H$	0.5135	0.0445	0	0.4825
$mean_{Gini}$	0.5135	0.0175	0	0.4960
$mean_{AUC}$	0.5135	0.0728	0	0.4137
$mean_{AUCH}$	0.5135	0.0741	0.0067	0.4124
$mean_{KS}$	0.5135	0.0148	0	0.4717
$mean_{MER}$	0.4973	0.1712	0.0162	0.3154
$mean_{MWL}$	0.4960	0.1725	0.0175	0.3140
$mean_{SpSe99}$	0.5135	0.0458	0	0.4407
$mean_{SeSp99}$	0.5135	0.0526	0	0.4340
$mean_{\frac{1}{AUC}}$	0.3868	0.1066	0.1267	0.3801
$mean_M$	0.4946	0.1429	0.0189	0.3437
$mean_{\frac{1}{M}}$	0.4784	0.0836	0.0350	0.4030
$mean_{RP}$	0.4798	0.0768	0.0337	0.4097

Table A.18: Confusion matrix for all models at cut-off=0.7 (P=0.50)

Appendix A

Combination formulas

$$C_{w_H} = \frac{f_{Cox}H_{Cox} + f_{RSF}H_{RSF} + f_{LR}H_{LR} + f_{CT}H_{CT}}{H_{Cox} + H_{RSF} + H_{LR} + H_{CT}}$$

$$C_{w_{Gini}} = \frac{f_{Cox}Gini_{Cox} + f_{RSF}Gini_{RSF} + f_{LR}Gini_{LR} + f_{CT}Gini_{CT}}{Gini_{Cox} + Gini_{RSF} + Gini_{LR} + Gini_{CT}}$$

$$C_{w_{AUCH}} = \frac{f_{Cox}AUCH_{Cox} + f_{RSF}AUCH_{RSF} + f_{LR}AUCH_{LR} + f_{CT}AUCH_{CT}}{AUCH_{Cox} + AUCH_{RSF} + AUCH_{LR} + AUCH_{CT}}$$

$$C_{w_{KS}} = \frac{f_{Cox}KS_{Cox} + f_{RSF}KS_{RSF} + f_{LR}KS_{LR} + f_{CT}KS_{CT}}{KS_{Cox} + KS_{RSF} + KS_{LR} + KS_{CT}}$$

$$C_{w_{MER}} = \frac{f_{Cox}MER_{Cox} + f_{RSF}MER_{RSF} + f_{LR}MER_{LR} + f_{CT}MER_{CT}}{MER_{Cox} + MER_{RSF} + MER_{LR} + MER_{CT}}$$

$$C_{w_{MWL}} = \frac{f_{Cox}MWL_{Cox} + f_{RSF}MWL_{RSF} + f_{LR}MWL_{LR} + f_{CT}MWL_{CT}}{MWL_{Cox} + MWL_{RSF} + MWL_{LR} + MWL_{CT}}$$

$$C_{w_{SpSe99}} = \frac{f_{Cox}SpSe99_{Cox} + f_{RSF}SpSe99_{RSF} + f_{LR}SpSe99_{LR} + f_{CT}SpSe99_{CT}}{SpSe99_{Cox} + SpSe99_{RSF} + SpSe99_{LR} + SpSe99_{CT}}$$

$$C_{w_{SeSp99}} = \frac{f_{Cox}SeSp99_{Cox} + f_{RSF}SeSp99_{RSF} + f_{LR}SeSp99_{LR} + f_{CT}SeSp99_{CT}}{SeSp99_{Cox} + SeSp99_{RSF} + SeSp99_{LR} + SeSp99_{CT}}$$

$$C_{w_{\frac{1}{AUC}}} = \frac{f_{Cox}\frac{1}{AUC}{}_{Cox} + f_{RSF}\frac{1}{AUC}{}_{RSF} + f_{LR}\frac{1}{AUC}{}_{LR} + f_{CT}\frac{1}{AUC}{}_{CT}}{\frac{1}{AUC}{}_{Cox} + \frac{1}{AUC}{}_{RSF} + \frac{1}{AUC}{}_{LR} + \frac{1}{AUC}{}_{CT}}$$

$$C_{w_M} = \frac{f_{Cox}M_{Cox} + f_{RSF}M_{RSF} + f_{LR}M_{LR} + f_{CT}M_{CT}}{M_{Cox} + M_{RSF} + M_{LR} + M_{CT}}$$

$$C_{w_{\frac{1}{M}}} = \frac{f_{Cox}\frac{1}{M}{}_{Cox} + f_{RSF}\frac{1}{M}{}_{RSF} + f_{LR}\frac{1}{M}{}_{LR} + f_{CT}\frac{1}{M}{}_{CT}}{\frac{1}{M}{}_{Cox} + \frac{1}{M}{}_{RSF} + \frac{1}{M}{}_{LR} + \frac{1}{M}{}_{CT}}$$

$$C_{w_{RP}} = \frac{f_{Cox}RP_{Cox} + f_{RSF}RP_{RSF} + f_{LR}RP_{LR} + f_{CT}RP_{CT}}{RP_{Cox} + RP_{RSF} + RP_{LR} + RP_{CT}}$$

$$C_{w_C} = \frac{f_{Cox}C_{Cox} + f_{RSF}C_{RSF} + f_{LR}C_{LR} + f_{CT}C_{CT}}{C_{Cox} + C_{RSF} + C_{LR} + C_{CT}}$$

$$C_{w_{\frac{1}{C}}} = \frac{f_{Cox}\frac{1}{C}{}_{Cox} + f_{RSF}\frac{1}{C}{}_{RSF} + f_{LR}\frac{1}{C}{}_{LR} + f_{CT}\frac{1}{C}{}_{CT}}{\frac{1}{C}{}_{Cox} + \frac{1}{C}{}_{RSF} + \frac{1}{C}{}_{LR} + \frac{1}{C}{}_{CT}}$$

Bibliography

[1] Aiolfi, M., Capistrán, C., and Timmermann, A., Forecast combinations. Unpublished manuscrip, 2010.

[2] Aiolfi ,M., and Timmermann, A., Persistence of forecasting performance and combination strategies. *Journal of Econmetrics*, 135:31–53, 2006.

[3] Altman, E., and Sabato, G., Modelling Credit Risk for SMEs: Evidence from the US Market, , 19(6): 716-723, 2006.

[4] Bates, J., and Granger, C., The combination of forecasts. *Operations Research Quarterly*, 20:451–468, 1969.

[5] Breiman, L., Random forest. *Machine Learning*, 45:5–32, 2001.

[6] Breiman, L., Friedman, J., and Stone, C. J., *Classification and Regression Trees*. First edition, Ganuary 1984.

[7] Cao, J., and Xie, X. J., and Zhang, S., and Whitehurst, A., and White, M. A., Bayesian optimal discovery procedure for simultaneous significance testing. *BMC Bioinformatics* 10:5. doi:10.1186/1471-2105-10-5.2009

[8] Capistran, C., and Timmermann, A., Forecast combination with entry and exit of experts. *Journal of Business and Economic Statistics*, 27:428–440, 2009.

[9] Chong, Y. Y., and Hendry, D.F., Econometric evaluation of linear macro-economic models. *The Review of Economic Studies*, 53(4):671–690, 1986.

[10] Clemen, R., Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5:559–581, 1989.

[11] Clemen, R., and Winkler, R.L., Combining economic forecasts. *Journal of Business and Economic Statis-ics*, 4:39–46, 1986.

[12] Cox, D. R., and Hinkley, D. V., *Theoretical Statistics*. Chapman and Hall, 1974.

[13] Cressie, N., and Read, T. R. C., Spatial modelling of regional variables. *Journal of the American Statistical Association*, 84(406):393–401, 1985.

[14] DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L., Correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, September 1988.

[15] de Menezes, L. M., Bunn, D.W., and Taylor, J. W., Review of Guidelines for the Use of Combined Forecasts. *Europeans Journal of Operational Research*, 120:190–204, 2000.

[16] Diebold, F.X., and Pauly, P., The use of prior information in forecast combination. *Board of Governors of the Federal Reserve System, Special Studies Paper-Division of Research and Statistics*, 218, 1987c.

[17] Dobson, A. J., and Barnett, A. G., *An introduction to Generalized Linear Models*. CRC Texts in Statistical Science, third edition, May 2008.

[18] Edgerton, H.A., and Kolbe, L.E., The method of minimum variation for the combination of criteria. *Psychometrika* , 1:183-188, 1936.

[19] Elliott, G., and Timmermann, A., Optimal forecast combinations under general loss functions and forecast error distributions. *Journal of Econometrics*, 122:47–79, 2004.

[20] Granger, C. W. J., and Newbold, P., Spurious regressions in econometrics. *Journal of Econometrics*, 2:111–120, 1974.

[21] Goin, J.E., ROC curve estimation and hypothesis testing: Applications to breast cancer detection, *Journal of the Pattern Recognition Society*, 15: 263-269, 1982.

[22] Granger, C., and Ramanathan, R., Improved methods of combining forecasts. *Journal of Forecasting*, 3:197–204, 1984.

[23] Fielding, A.H., and Bell, J.F., A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24: 38-49, 1997.

[24] Figini, S., and Fantazzini, D., Random Survival Forest models for SME Credit Risk Measurement. *Methodology and computing in applied probability*, 11: 29-45, 2009.

[25] Figini, S., and Gigliarano, C., and Muliere, P., Making classifier performance comparisons when Receiver Operating Characteristic curves intersect. *Quaderni di Statistica*, 14, 2012.

[26] Figini, S., and Giudici, P., Merging of Rating Models. *Journal of the Operational Research Society*, 62; 1067–1074, 2011.

[27] Geisser, S., Bayes approach for combining correlated estimates. *ournal of the American Statistical Association, Series A*, 60: 602–607, 1965.

[28] Halperin, M., Almost linearly-optimum combi- nation of unbiased estimates. *Journal of the American Statistical Association*, 56:36–43, 1961.

[29] Hand, D. J., Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine Learning*, 77:103–123, 2009.

[30] Heagerty, P. J., and Zheng, Y., Survival model predictive accuracy and roc curves. *Biometrics*, 61:92–105, March 2005.

[31] Horst, P., a composite measure from a number of different measures of the same attribute. *Psychometrika*, 1:53–60, 1938.

[32] Hosmer, D., and Lemeshow, S., and May, S., *Applied Survival Analysis: Regression Modeling of Time to Event Data* (second ed.). Wiley-Interscience, 2008

[33] Iswaran H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S., Random survival forest. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

[34] Kass, G. V., An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29:119–127, 1980.

[35] Kleinbaum, D. G., and Klein, M., *Survival Analysis*. Springer Science Business Media, Ganuary 2005.

[36] Makridakis, S., and Winkler, R. L., The combination of forecasts. *Juornal of the Royal Statistical Society*, 146:150–157, 1983.

[37] Makridakis, S., and Andersen, A., and Carbone, R., Fildes, R., and Hibon, M., and Lewandowski, R., and Newton, J., and Parzen, E., and Winkler, R., The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, 1:111-153, 1982.

[38] Mogensen, U. B., Ishwaran, H., Gerds, T. A., Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. *Journal of Statistical Software*, 50(11):1–23, 2012.

[39] Narain, B., *Survival analysis and the credit granting decision*, 1992.

[40] Newbold, P., and Granger, C.W. J., Experience with forecasting univariate time series and the combination of forecasts. *Juornal of the Royal Statistical Society*, 137:131–164, 1974.

[41] Reid, D.J., Combining three estimates of gross domestic product. *Economica*, 35:431-444, 1968.

[42] Ripley, B. D., Recognition and Neural Networks. Cambridge: Cambridge University Press, 1996.

[43] Satchell, S., and Xia, W., Analytic models of the ROC curve : applications to credit rating model validation. Sydney, August 2006.

[44] Schoenfeld, D., Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241, April 1982.

[45] Scott, M.J. J., Niranjan, M., and Prager, R. W., Realisable classifiers: Improving operating performance on variable cost problems. In *British Machine Vision Conference*, pages 304–315, 1998b.

[46] Stepanova, M., and Thomas, L., Survival analysis methods for personal loan data. *Operations Research*, 50(2): 277–289, 2002

[47] Stigler, S.M., Laplace, Fisher, and the discovery of the concept of sufficiency. *Biometrika*, 60: 439-445, 1973

[48] Stock, J., and Watson, M., A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series, *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive Granger*. Oxford University Press, Oxford, 2001.

[49] Swanson, N. R., and Zeng, T., Choosing among competing econometric forecasts: Regression-based forecast combination using model selection. *Journal of Forecasting*, 20:425–440, 2001.

[50] Timmermann, A., Forecast combinations. In G. E. et. al, editor, *Handbook of Forecasting*, volume 1, pages 135–196. Elsevier, Amsterdam, 2006.

[51] Thomas, L., and Banasik, J., and Crook ,J., Not if but when loans default. *Journal Operational Research Society*, 50:1185–1190,1999.

[52] Yang, Y., Combining forecasting procedures: some theoretical results. *Econometric Theory*, 20:176–222, 2004.

[53] Zani, S., *Analisi dei dati statistici*, volume 2, *Osservazioni multidimensionali.* Giuffré, Milano, 2000.