Monitoring the survival of subscribers in a marketing mailing list

Andrea Marletta

Abstract

The statistical literature proposed many contributions about survival analysis in medical research, in this work this approach is proposed in a business context. The aim of this paper is to control the mortality of the users belonging to an e-mail subscribers list for a company operating in the healthcare information sector. Having available the survival times for each subscriber, the choice was oriented to survival models to evaluate the abandon of the customers. A survival analysis was conducted through a Cox model considering some risk factors of the subscriber. The selected Cox model carried to the identification of risk profiles representing different situations in terms of probability of abandon.

Keywords: *Mailing list marketing; Healthcare information; Survival analysis; Cox Model*

1. Introduction

During last years, the marketing strategies experienced many changes and the introduction of new technological tools favoured innovative approaches more hinged on a direct relationship with customers. One of these approaches is based on the use of the e-mail. This tool passed from a simple communication media to a privileged channel for the direct attainment of a possible list of customers. The user could enter in a marketing mailing list after a volunteer subscription, indeed during an on-line purchase, the entering of an e-mail address in a mandatory information to complete the process. This information should be used for a faster interaction between the customer and the seller and it is inserted into a business database. But even after the accomplishment of the purchase process, the communication with the seller proceeds with the sending of the advertising material in order to make easier new purchases in the future. In other cases, the addition in the e-mail database occurs indirectly, for example after the transfer of the information by a third stakeholder, according to provided modalities by the contract. The user often accepts unconditionally these terms, above all during e-commerce purchases, without giving importance to the transfer of personal data, generating problems in terms of data privacy.

The introduction of the direct mailing among the marketing strategies generated a raising of the available data. This tool led to creation of a business databases that starting from the e-mail address, it joined personal information of the customer (name, surname, age, gender, professional role, telephone number, address,...) and other information related to the interaction business-customer (date of entering, last mail sent, number of sent mails, last click on e-mail, ...). Using this database, the companies could obtain useful information to profile the customer list, customizing the sending of personalized contents during time. Beyond the profiling, the firms could be interested to monitor the abandons, due to voluntary cancellations from the mailing list or for the closure or disuse of the e-mail address automatically reported after the shipping.

The existent literature showed recently interest for this issue focusing on how customers respond to email messages or looks at the average effect of email on transactional behaviour (Zhang, 2015) or investigating the effectiveness of triggered e-mail marketing (Goic et al., 2021). In this paper, this strategy is faced from a statistical point of view. Here, the statistical approach is focused on survival analysis, a method usually applied in medical research. This technique is well-known and described in Clark et al. (2003); Collett (2015); Cox and Oakes (2018); Hougaard (2000).

The term survival is here intended as synonym of permanence of an individual in the mailing list with the registered e-mail address. This approach is data-driven since this technique needs a survival time variable obtained as difference between the date in which the e-mail address entered the list and the date in which the last mail was sent. In this date, the customer chose

to not receive e-mails. Another similarity with the medical approach is in the concept of censored data, that is to say, the statistical unit that did not experiment the death effect. In this case, the censored unit is the customer that received the last e-mail when the data collection period is over.

Applying this technique in this context could have multiplex aims: firstly, to monitor from a descriptive point of view the effectiveness of the direct mailing; secondly to compute the abandon rates of the mailing list; finally, to detect customer profiles more at risk. The aim of this work is to verify the applicability of these models in a context different from the usual one and to provide to the companies a new tool suitable to check the reliability of this marketing strategy.

2. Survival analysis

Survival analysis contains all the techniques and statistical models designed for the description and the analysis of time events of a statistical unit. It is necessary to identify the unit exposed at risk respect to this event and the measure of the time duration and the end of these event. Survival is therefore characterised by a time variable with a start-up and an end-point. In medical research, start-up corresponds to time in which an individual has been introduced in the experimental study or a clinical treatment or the start of a particular condition for a disease. On the other hand, if the end-point is the death of the patient, data are referred to the death time. The end-point could be not necessarily the death, but also the end of a pathological state. For this work, the start-up is the date in which the customer was subscribed in the e-mail lists and the end-point is represented by the exit of the customers from the list.

Survival analysis can be treated using non-parametric, parametric or semi-parametric models. The first nonparametric approach considers the estimate of the survival function of a t time variable using the life-tables. These tables are obtained dividing the observation period in temporal intervals (Collett, 2015). Non-parametric models are very flexible but they do not guarantee consistent and precise estimates. This is why they are usually as exploratory tools. For this reason, parametric models have been proposed proposing that the time variable assumes a probability distribution depending on some parameters. Once the function probability distribution function $h(t)$ and the cumulative hazard risk function $f(t)$ is chosen, then it is possible to obtain the survival function $H(t)$. Finally, semi-parametric models were introduced by Cox (1972) and it is so defined because even if it is based on the hypothesis of proportional hazards, it makes no assumption about a probability distribution for the survival times. The Cox model assumes the hazard risk function $h(t)$ as a product of two components:

$$h(t) = h_0(t) * exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}) \quad (1)$$

The first component $h_0(t)$ is named baseline hazard risk function, the second one is the exponential of the sum of the combination terms $\beta_i x_i$ extended to all p explanatory variables. The computed model allows to identify some categories of users more at risk.

3. Application

The analysis was based on research proposed by PKE, Professional Knowledge Empowerment, a company created to manage Italian healthcare databases. Over time, the areas of expertise have expanded, thus specialising both in data management and communication. In the communication area, one of the services is the e-mail marketing. From an increasingly digital standpoint, communication strategies must also take into account the change that PKE reinterprets, by making email marketing projects available that guarantee precious and exclusive value: in-depth knowledge of the health professional and in particular of doctors.

In this paper, the dataset is composed by all the subscribers in the PKE e-mail marketing list until 2021. PKE sends over 18 million e-mails every month, this tool has allowed it to perfect communication models for promotion of drugs in launch, mature or in decline, in siding or replacing the local pharmaceutical representative. The target audience is made of pharmaceutical companies, medical device companies, certification bodies, scientific societies, patient associations, insurance, technology companies, public/private bodies of the NHS, CME providers, publishing companies, public utilities. The type of subiscription is volunteer since the user takes part to some events related to the world of healthcare information. The professional background of these users is mainly represented by professionals in the medical sector.

Several models could be obtained considering different dependent variables of the Cox model. A time variable could be computed as difference between the subscription in the list and the last received e-mail. Another time variable could be the difference between the subscription and the last time the subscriber opened or clicked the e-mail. Once these Cox models are estimated, it is possible to define some risk profiles and determine the categories of target audience more inclined to abandon the e-mail marketing strategy.

The variables considered as potential risk factors for the abandon state are referred to the personal information of the subscriber:

- gender;
- age;
- workplace;
- main professional figure;

- medical specialization.

The full dataset is composed by 651.783 mailing lists for the PKE subscribers. First of all, it could be interesting to compute the percentage of abandoned e-mails over to the total. The number of abandons is equal to 162.960. So the abandon rate could be easily computed:

$abandon\ rate$ = not-active e-mails/ $total\ email\ addresses$ = 162.960/651.783 = 25,0%

Some operations of aggregation and encoding were applied on the original variables in order to better organize the data. The age of the subscriber was categorized in three slots: young (until 35 years old), medium-age (between 35 and 65 years old) and retired (more than 65 years old). About the workplace of the subscriber, this information is available at different levels according to the Istat classification: NUTS1 (macro-regions), NUTS2 (regions) and NUTS3 (provinces). The number of professional figure considered are 55, they are grouped in medical and non-medical area. For individuals in medical area, it is available the sub-category of specialization. It was divided into 5 macro-categories: Medical area, Clinical Services, Surgery, Dentistry and Other. Specialization in general medicine groups internal and specialized medicine, psychiatry and pediatrics. Clinical services gathers radio diagnostics, anaesthesia, rehabilitation medicine and public health. Surgery specialization includes general surgery, neurosurgery and heart surgery.

In table 1, the frequency distribution for the risk factors are shown. There is equal distribution between men and women among subscribers (51,8% vs 48,2%). The majority of subscribers is in the middle age class (2 over 3), more than 1 over 4 is more than 65 years old and only 7,5% of the customers is under 35 years old. For this representation, workplace variable is represented by the higher level, NUTS 1, Northwest and Central Italy are the most present area with respectively (28,0% and 26,4%), followed by Northeast and South Italy (17,2% and 18,0%). Only 10,4% of the subscribers works in Insular Italy. Since the source of the dataset is a company specialized in healthcare information, most of the subscribers (87,3%) belongs to the medical area as main professional figure. For this customers, it is also available the information about the specialization. For subjects in medical area, 45,6% is represented by physicians operating in general medicine specializations followed by clinical services (23,3%) and surgery (22,8%), finally 6,5% of subscribers are in the dentistry area.

The frequency distribution allows to build the baseline profile, joining the typical features of the subscriber. This customer is a man of medium-age with a workplace in Central Italy belonging to medical area specialized in general medicine. Central Italy has been chosen instead of Northwest Italy as a reference level because when spatial variables are considered, it is preferable to choose a central area. To measure the abandon risk, a semi-parametric Cox model was adopted as described in the previous section. Estimated coefficients β_i, the hazard ratio relative risk $e(\beta_i)$ for all risk factors are presented in table 2. Applying this model, it is possible to create the profile of a customer with higher or lower abandon risk.

Table 1. Risk factors frequency distribution

Risk factor	Percentage (%)
Gender	
Men	51,8%
Women	48,2%
Age of the subscriber	
Young	7,5%
Medium-age	66,6%
Retired	25,8%
Workplace (NUTS1)	
Northwest	28,0%
Northeast	17,2%
Central	26,4%
South	18,0%
Insular	10,4%
Professional figure	
Medical area	87,3%
Non-medical area	12,7%
Medical Specialization	
General medicine	45,6%
Clinical services	23,3%
Surgery	22,8%
Dentistry	6,5%
Other	1,8%

In table 2, the underlined levels of risk factors represented the baseline profile, the βi coefficients could be interpreted in terms of significance and in terms of abandon risk considering the hazard ratio relative risk $exp(\beta_i)$. Last column reports the p-value, associated to the hypothesis test. Since all p-values are under the threshold of 5%, then all the explanatory variables have significant coefficients, this means that no risk factors have to be deleted from the full model.

The estimated parameters could be interpreted in terms of sign and value. The positive sign implicates an higher risk in comparison with the baseline level. The hazard ratio HR $e(\beta_i)$

indicates how much increase the risk for the subscriber with that level of the explanatory variable.

Table 2. Full semi-parametric Cox Model

Risk factor	β_i	$exp(\beta_i)$	P-value
Gender			
Men	0,000		
Women	0,074	1,077	0,000
Age of the subscriber			
Young	0,522	1,685	0,000
Medium-age	0,000		
Retired	0,238	1,269	0,000
Workplace (NUTS1)			
Northwest	0,098	1,103	0,000
Northeast	0,251	1,285	0,000
Central	0,000		
South	0,017	1,017	0,217
Insular	-0,112	0,894	0,000
Professional figure			
Medical area	0,000		
Non-medical area	0,140	1,150	0,000
Medical Specialization			
General medicine	0,000		
Clinical services	0,146	1,157	0,000
Surgery	0,090	1,094	0,000
Dentistry	-0,407	0,666	0,000
Other	-0,035	0,966	0,489

For example, the value $\beta_i = 0{,}074$ for women reports an higher risk of abandon of the mailing list for this category. The value $e(\beta_i) = 1{,}077$ means that there is a +7% ($exp(\beta_i)-1$) of risk for these users. Young and retired subscribers are more at risk of abandon compared to medium-age subscribers (+68% and + 27%). Users with workplace in Northeast of Italy are the more at risk (+28% vs the Central Italy). In the medical area, physicians specialized in

clinical services are more inclined to leave the list. Missing values for this variable are residuals, so they did not affect the consistency of the model.

4. Conclusions

The proposed approach combines the use of survival analysis in a business context for measuring abandon risk in e-mail marketing. The availability of information about time days between first and last e-mail sent suggested the use of a semi-parametric Cox model. The results for this model for abandon rate are satisfactory detecting low and high risk profiles. These results could be very useful for the company owner of the mailing list. For example, starting from the Cox Model, it is possible to compute a risk score for each profile. Using this tool, when a new user enters the list, it could be immediately classified as user more or less inclined to abandon. The age of the subscribers seems to be a significant risk factor, considering young and retired users as more at risk individuals. Some territorial differences are present using NUTS1 and specialization, a possible enhancement could regard the use of NUTS2 and NUTS 3 variables focusing the attention on a smaller area.

Future works could regard different survival analysis models, for example comparing these preliminary results with methods as Kaplan-Meier survival curves or with exponential or Weibull models. These methods are parametric and usually proposed for descriptive issues. Finally, the presented model could be enhanced using the number of emails sent during the considered period or different time variables such as time between first e-mail sent and last click.

References

Clark, T.G., Bradburn, M.J., Love, S.B., Altman, D.G., (2003). Survival analysis part i: basic concepts and first analyses. British journal of cancer 89, 232–238.

Collett, D., (2015). Modelling survival data in medical research. CRC press.

Cox, D.R., (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) 34, 187–202.

Cox, D.R., Oakes, D., (2018). Analysis of survival data. Chapman and Hall/CRC.

Goic, M., Rojas, A., Saavedra, I., (2021). The effectiveness of triggered email marketing in addressing browse abandonments. Journal of Interactive Marketing 55, 118–145.

Hougaard, P., (2000). Analysis of multivariate survival data. volume 564. Springer.

Wu, J., Li, K.J., Liu, J.S., (2018). Bayesian inference for assessing effects of email marketing campaigns. Journal of Business & Economic Statistics 36, 253–266.

Zhang, X., (2015). Managing a Profitable Interactive Email Marketing Program: Modeling and Analysis. Georgia State University.Blazquez, D., & Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, 130, 99-113.