Survival analysis in a business context: how to control the abandons of my subscribers

Analisi di sopravvivenza in contesto aziendale: come controllare gli abbandoni fra i miei clienti

Andrea Marletta and Marco Morandi

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.

Abstract *La letteratura statistica ha proposto molti lavori riguardanti l'analisi di sopravvivenza nella ricerca medica, in questo lavoro questo approccio è proposto in un contesto azeindale. Lo scopo del contributo è modellare la mortalità degli utenti appartenenti ad una lista di iscritti via mail per una azienda operante nel settore dell'informazione medico-scientifica. Avendo a disposizione i tempi di sopravvivenza per ogni iscritto, la scelta è stata orientata su modelli di sopravvivenza per valutare gli abbandoni dei clienti. L'analisi di sopravvivenza è stata effettuata attraverso un modello di Cox considerando alcuni fattori di rischio per i clienti. Il modello finale selezionato ha portato all'identificazione di categorie di clienti più a rischio rappresentando situazioni differenti in termini di probabilità di abbandono.*

Key words: Survival analysis, e-mail marketing, Cox model

Andrea Marletta
University of Milano-Bicocca e-mail: andrea.marletta@unimib.it

Marco Morandi
PKE e-mail: m.morandi@pke.it

1 Introduction

During last years, among the marketing strategies, one of the most used is the e-mail marketing. Companies are using email marketing to engage with customers and encourage active transactional behavior. Extant research either focuses only on how customers respond to email messages or looks at the average effect of email on transactional behavior [7].

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, this strategy is faced from a statistical point of view. Other authors tried to deal with this issue using a Bayesian approach [1, 6], here the used technique is the survival analysis. To apply this method, it is necessary to have a time variable measuring the difference between the birth and the death of the phenomenon. About the e-mail marketing, the birth time could be represented by the entrance of the customer in the subscriber list and the death the exit from this list.

Starting from this analogy between the concept of death in a natural population and the e-mail marketing, the idea is to use some statistical techniques usually used for survival analysis as models able to predict information on the subscribers. Following this approach, parametric (Weibull) and semi-parametric models (Cox) have been applied for the available data. The aim of this paper is to create an useful tool to follow the temporal evolution of the profiles. This could be done creating different strategies on the basis of the features of the customer.

The paper is structured as follows: after the introduction, a second section is dedicated to the methodologies used to answer the research objectives. A third section will show the description of the dataset and some preliminary results.

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 data own some features that need the use of some tailored statistical procedures. The first one is their distribution, generally survival data are not symmetrically distributed: an histogram based on survival times will tend to be positively asymmetric; this means that all classical models as linear regression are not suitable for these data. The second one is the presence of censored data, that is to say, statistical observation that did not experiment the time event. In medical research, they are all patient are not dead at the end of the experiment or dead for alternative causes or retired from the treatment.

Survival analysis can be treated using non-parametric, parametric or semi-parametric models. The first non-parametric 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 [2].

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. This approach allows to determine possible combinations of explanatory variables or risk factors conditioning the risk and the survival function. Once the probability distribution function $(f(t))$ is chosen, then it is possible to obtain the survival function $(S(t))$, the hazard risk function $(h(t))$ and the cumulative hazard risk function $(H(t))$. The most used probability function for time variables are the exponential and the Weibull distribution.

Finally, semi-parametric models were introduced by Cox [3] in 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 $h_i(t)$ as a product of two components:

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

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.

Survival analysis is principally used in medical statistics, but there are a lot of applications of this method in economic issues. The application of survival analysis in economic field could have companies or obligors as statistical units.

Giambona and Vassallo [4] in 2007 built hazard risk profiles for Italian banking credits using a survival model at discrete time on loan data from Italian banking system. They want to observe the hazard risk function in the first decade after the loan assignment. The hazard risk profiles were built using the risk levels compared to the baseline profile through odds ratios. Giambona also studied the death rate of Italian banking credits using a non-proportional hazard logistic model [5].

3 Application

In this paper, the dataset was available thanks to PKE and it is composed by all the subscribers in their e-mail marketing list. PKE sends over 18 million emails 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.

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.

The risk factors included in the Cox model are the number of received mails and the feature of the subscriber. The available information are gender, age, workplace, the dummy variable about activity profile (1 for active, 0 for non-active), the belonging to a category of the target audience and their specialization.

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.

References

1. Ansari, A., Mela, C. F. (2003). E-customization. Journal of marketing research, 40(2), 131-145.
2. Collett D (1994). Modelling Survival Data in Medical Research. Chapman & Hall: London.
3. Cox DR (1972). Regression models and life-tables (with discussion). Journal of Royal Statistical Society, Series B 74: 187–220.
4. Giambona F., Vassallo E., (2007), Profili di rischio dei crediti bancari italiani: un'analisi per generazioni di finanziamenti, Rivista Minerva Bancaria, 2, 9-46.
5. Giambona F., (2007), Mortalità dei crediti bancari italiani: Altre evidenze empiriche, Rivista Minerva Bancaria, 5, 1-16.
6. Wu, J., Li, K. J., Liu, J. S. (2018). Bayesian inference for assessing effects of email marketing campaigns. Journal of Business & Economic Statistics, 36(2), 253-266.
7. Zhang, X. (2015). Managing a Profitable Interactive Email Marketing Program: Modeling and Analysis. Georgia State University.