Jorge Segura Gisbert, José Antonio Alvarez Jareño, José Pavía Miralles and Patricia García Torres

Detección del riesgo de fuga de clientes de una entidad aseguradora mediante algoritmos de machine learning

Executive Summary

Lapse risk detection in a life insurance company based on machine learning algorithms

Many authors have researched and analysed the risk of client departure. In this paper, we focus on the detection and prediction of the risk of departure based on the premise that the cost of capturing a new customer is greater than the cost of retaining them. In this way, detecting and predicting the risk of departure in a more precise way will help the Insurer to focus on areas which promote client retention. The objective of the current paper is twofold:

• To detect and predict which clients are thinking of leaving.

• To interpret and explain the selected model for the risk of departure.

The question is not *who* is going to leave the company but *why* they are thinking of leaving. Machine learning methodologies help to identify people with a high probability of leaving the company; in other words, who? However, if the algorithms used, at least those that normally obtain better results, are complex, there is little possibility of being able to interpret the model and explain in a logical and reasoned way the behaviour of the clients; in other words, why?

Based on the real database of an insurance company with a commercial portfolio in non-life insurance, an analysis of the data is carried out. The Insurance Company provided a database of car policies for the period 2012-2016 for the realization of the study. The database is a flat text file (.csv), composed of 173 variables and 473,356 instances.

It is a *conditio sine quan non* to identify the different variables as a starting point to the modelling process. The time horizon collected is sufficient and meets the requirements established in the initial period of analysis. The information collected is made up of risk factors used in pricing models, behaviour and payment methods, Bonus-Malus systems, etc. The raw material is the data, which means the final result will depend on the quality of the raw material. In this first phase, the database is treated and cleaned, including the transformation or generation of new variables in the database. We are aware that if the data are too noisy they could lead to erroneous decisions and conclusions. An indispensable aspect is the determination of the objective variable classifying the different reasons for annulment. The problem is

diametrically opposed depending on whether the departure is voluntary or not. Synergistically, we analyse all the variables. However, only those that are essentially representative for the purpose of this research are included in the paper. As we have said, the main variable under study is the voluntary departure of the clients from the company.

In the second phase, variables are selected. The principle of parsimony or Ockham´s knife is imposed. Some variables may be important, others may provide some information and many others may be irrelevant; some may even provide only noise, which would be counterproductive. Although not in line with certain theoretical approaches, practice has taught us that sometimes a subset of variables obtains better results than all available variables. The introduction of a lot of noise can mean the algorithm is not able to detect the signal.

Several methodologies and techniques are applied to enable selection of those variables that will provide more efficient models which will work better. First, we consider indirect methods that are generally used as a step for pre-processing the data. The selection of variables is independent of any machine learning methodology that is applied to a data set. The variables are selected on the basis of the valuations obtained in several statistical correlation tests with the dependent variable. The correlation will depend on the type of variables that are being used. An important issue in this section is to remember that filtering methods do not eliminate multicollinearity, so it is necessary to address the collinearity problems of the variables before modelling.

Secondly, we consider direct methods consisting of selecting a subset of variables to train a model and, depending on the results, adding or removing certain variables. The difficulty is finding the best subset of variables. These methods are usually computationally intensive. One of the best ways to implement the selection of variables with direct methods is to use the R package Boruta.

Thirdly, we consider integrated methods that combine direct and indirect methods. They are implemented through algorithms that have their own methods of selecting built-in variables. Some of the most popular are the LASSO and RIDGE regression that have intrinsic penalty functions to reduce over-adjustment. Once the set of variables that will be part of the analysis has been selected, the process of extracting the information from the data can begin. The next stage to be covered is the division of the data set into two subsets; one of training and the other of testing.

Next, the algorithms are selected with which the models will be first trained and then the instances of the test set classified. As already indicated, the data set is unbalanced in its dependent variable, so the methodologies used to solve the problem, if not completely at least partially, are revised. This section ends by establishing what the evaluation measures will be of the different proposed models. Being able to compare the results will enable selection of the best, given that the spatial characteristics of the sample means that a single measure of precision would not enough.

The usual procedure in automatic learning or machine learning is the division of the data into two subsets: one of training and another of checking or testing. When dealing with data that have a temporality, the division is made according to the periods requiring prediction. Depending on the algorithms used, we can find models that explain the training data very well, having a poor predictive capacity. Our main objective is to be able to predict in advance which customers have a greater risk of departure in order to target commercial actions to retain the client. Once the training and prediction periods are determined, the two subsets are generated to construct the models to be used to explain the departure of the clients. By having a database that spans 3 years, a cascade analysis is used. An initial period of two years is used as training data, and predictions are then made using the quarter immediately after this period.

This procedure is repeated four times, with different sets of training and testing, as if the analysis was moving forward in time. This procedure can show whether the model varies with time and temporal behaviour. The predictive capacity of the model is obtained by comparing the real results of the sample with the predictions of the model.

When you have a variable with two categories where one is much larger than the other, the data is said to be unbalanced. This problem is quite common in machine learning because the group to be identified is very small within a huge collective. To avoid this problem, there are different methods available to balance the data. A modification of the size of the original sample can help to generate a new balanced data set. These methods have become very important in light of many researchers proving that the results of the balanced data in a classification process improve substantially compared to the unbalanced data. With the SMOTE technique (Synthetic Minority Oversampling TEchnique), the data set is balanced by generating artificial data, similar to a form of oversampling but with better conditions. This technique generates a random set of observations of the minority class in order to change the learning bias of the classifier towards the minority class.

Automatic learning methods are used to solve problems when managing customers from different perspectives. The difficulty with this research project is the classification of the different clients: customers with probability of departure and those who plan to remain with the company. To complete this classification, different algorithms are analysed. Firstly, a literature review is carried out of the methods that previously obtained the best results in analogous problems.

Based on this review, we proceed to apply the different algorithms which obtained good results for similar problems. However, in this paper only the results of three algorithms will be shown: Decision tree with pruning, Random Forest and Extreme Gradient Boosting.

The analysis is based on data with a total of 53 variables (52 independent variables and one endogenous variable). Traditionally, random forest algorithms have excelled in the data science field but have recently been superseded by the Extreme Gradient Boosting algorithm. Its success is due not only to its precision in classification of instances but to its processing speed. Although there is no literature that supports the application of this methodology for the resolution of departure risk problems, we decided to include it due to the good results obtained.

It is almost impossible to build a perfect model that allows the correct classification of all the instances of the test set. So, we must choose the classification model that best suits the needs of the problem and works best in the domain. The confusion matrix or classification matrix is useful in machine learning to compare the real values of a variable with the values estimated by the model built with the training data. When constructed as a matrix, the rows represent the prediction values of the model and the columns represent the actual values of the variable. In this way four categories can be identified: false positive, true positive, false negative and true negative.

The importance of the matrix is that it not only tells us which instances are correctly classified but also where those that have been classified correctly are classified.

The models are trained with the data corresponding to a period of two years and, subsequently, the predictions are checked on the quarter immediately after the training period. In addition, the models are generated with unbalanced data and, then, with the inclusion of synthetic data through the SMOTE algorithm. The first subsection presents the results of the three models with the unbalanced or, in other words, the original data and the second subsection shows only the results for the model made with the XGBoost algorithm with SMOTE data.

The comparison is made on two levels. The first level, in terms of the descriptive capacity of the models, does not present substantial differences and all have an overall precision similar to each other. On the second level, in relation to the predictive capacity, the SMOTE methodology with the xgboost algorithm presents an important improvement compared to the rest of the models. The objective is to correctly classify the maximum number of insured persons at "risk of departure". The algorithm xgboost enables us to obtain a probability of the variable relevant to our objective: that is, the probability that each client is going to leave at the next due date, although ultimately each client is classified as departure (probability greater than 50%) or permanence.

The fact that the models are not interpretable causes some problems. To overcome these, a series of tools are being developed that enables comprehension and interpretation of the most complex models. An insured individual can request an explanation of how the model affects him/her according to the values of the variables of each individual and this meets the requirements posed by the new General Regulation of Data Protection. However, it is unlikely that many policyholders will ask the company to explain how the model being applied affects them.

The marketing department of an insurer that is interested in retaining those insured individuals who are profitable yet have a high probability of departure will want to know the most important variables for each of them, their sign and their weighting . The objective is to carry out a personalized retention strategy based on the individual results for each insured individual at risk of departure. By being able to interpret the ensemble of models, the marketing department will be able to carry out individualized actions and, subsequently, look at more appropriate ways of measuring the efficiency of these commercial actions.