Predicting Customer Churn in the Insurance Industry: A Data Mining Case Study

Francesco Schena

Introduction

In markets with fierce competition and low switching costs customer attrition (or churn) is a key concern. In the past marketers addressed the problem primarily by making the customer loyal to the brand. More recently, roles are reverting and brands are being committed to demonstrate loyalty to the customer, through virtuous customer relationship management (CRM) that pre-empts churn before it occurs. As the business know-how is becoming increasingly data-literate and technology breaking down the barriers for IT infrastructure costs, analytics is proved effective in anticipating churn behaviour, enabling targeted, cost-effective retention campaigns towards customers at risk of defection.

The purpose of this research is to develop an effective and easy-to-interpret model for predicting churn in commercial line insurance, an industry that, particularly in mature markets, strongly relies on customers' lifetime.

Methodology

The research methodology followed one of the most popular frameworks for data mining projects: CRISP-DM (*Cross Industry Standard Process for Data Mining)* framework (Chapman et al. 2000; Shearer 2000). It consists of the following six phases: (1) Business Understanding; (2) Data Understanding; (3) Data Preparation; (4) Modelling; (5) Evaluation and (6) Deployment.

F. Schena (✉)
Etihad Airways, Khalifa City, United Arab Emirates
e-mail: francesco.schena@gmail.com

L. Petruzzellis, R.S. Winer (eds.), *Rediscovering the Essentiality of Marketing*,
Developments in Marketing Science: Proceedings of the Academy of Marketing
Science, DOI 10.1007/978-3-319-29877-1_141

Business Understanding

In conjunction with business experts, it has been defined that the data mining project should answer the question: ‘which policies will be cancelled within the next 12 months?’ through binary classifiers that output 1 if the contract will be cancelled, 0 otherwise.

Giving a definition to the word “cancelled” is non-trivial. According to Hadden (2007), deliberate churn (i.e. churn due to voluntary and explicit decision of the customer) is the only case of attrition that companies can avoid. In our case, we modelled any cancellation event, irrespective of whether it was deliberate or not, because the quality of the data was not good enough to infer a reliable classification of deliberate churners. Unsurprisingly, any attempts of modelling deliberate churn had no success.

Data Understanding and Preparation

The data set was made up of over 600 k records, which were randomly split among training, validation and test sets. Each record had 410 attributes. These can be grouped into the following six categories (Schena 2014): (a) product specifications; (b) premium, (c) customer demographics, (d) customer behaviour, (e) claims history and (f) customer relationship.

The main challenge at this phase was to deal with the rarity of the positive class, as cancelled policies were less than 10 % of the data set. This is a recurring issue in churn modelling (Burez and Van den Poel 2009) that biases the model in favour of the majority class (Longadge and Dongre 2013; Weiss 2004), since there would be more chances to classify correctly a non-churner than a churner. To address this limitation we evaluated three approaches: undersampling, which means removing instances from the majority class at random until obtaining a 50:50 balanced data set; cost-sensitive learning, which would weigh the accuracy of churn classification more than the one of non-churn classification; rule induction tree pruning, a tree ensemble algorithm that removes the purest splits recursively to minimise class imbalance bias.

Another problem at this stage was to reduce the number of features, being the data too wide. The disadvantages of handling too many features are (Yu and Liu 2004): (a) high computational time, (b) difficulties in model interpretation and, above all, (c) overfitting. To work around this problem, we combined expert judgement with the following methods: R^2-based selection and decision tree-based selection.

Modelling

As the literature confirms, the most widespread models in churn prediction are decision tree and logistic regression. These were the preferred choice also on this study because of the ease of understanding and interpretation for decision-makers.

Rule Induction Pruning (RIP) Ensemble

This algorithm builds rule classifiers in a tree-based fashion (de Ville 2007; SAS-Institute 2014) with the aim of improving learning from the rare class. The mechanism is the following: leaves purer than a certain threshold are removed from the training set and a new tree is built on the remaining instances. This step is reiterated until no more leaves reach the given purity threshold. Finally, residuals are cleared through an artificial neural network.

Model Evaluation

After discarding invalid or poor performing models, the choice for the final model was drawn among eight competing options. The models were assessed against (1) predictive performance, (2) robustness and (3) interpretability criteria, as shown below.

Predictive performance. We used the following metrics: *accuracy*, the percentage of correct predictions, *sensitivity,* the percentage of positive instances classified as such and *specificity,* the percentage of negative instances classified as such. From a managers' view, accuracy, sensitivity and specificity were found extremely intuitive and thus preferred to other performance indicators. Nonetheless, the champion model was tested against other metrics as well: AUC/ROC and cumulative lift.

The *ROC chart* plots sensitivity in the vertical axis versus false positive rate (i.e. 1—specificity) in the horizontal axis, at each percentile of the population. The broader the underlying area (called AUC or ROC index), the more accurate the predictions.

The *cumulative captured response* is the percentage of positive instances captured up to a given percentile of the population. The *cumulative captured response* chart draws the cumulative response captured for each percentile. The *cumulative lift chart* is a variant that benchmarks the cumulative response to the random classification.

Robustness. The robustness is the ability of the model to resist external disturbances. Clemente et al. (2010) suggest the difference in absolute values between the AUC for the validation set and the AUC for the training set as a measure of the model's resistance to noise.

Interpretability. There is often trade-off between predictive accuracy and ease of interpretation, and the challenge for the modeller is to determine how to weight both factors so that the model chosen is the best suited for the problem at hand (Feelders, Daniels and Holsheimer 2000). Overly sophisticated models make the models less understandable, therefore less successfully deployable by the business. Generally, classifiers based on 'if-then' rules like decision trees are the most intuitive, whereas models like artificial neural networks require more abstraction. In this study, the firm was very concerned about the deployability of the model, to such an extent that comprehensible classifiers were always preferred to *black box* models, irrespective of their accuracy.

Findings

The choice of the model to deploy was drawn among four decision trees, three logistic regression and one ensemble model. All models were trained by cost-sensitive algorithms that penalise the misclassification of positives. Plus, some models were fitted to an undersampled data set.

Overall, logistic regression models consistently outperformed in estimating the correct churn probability. Nonetheless, the model chosen was a decision tree because of its great simplicity and comprehensibility, having only 21 leaves derived from nine variables. What is more, this model had acceptable accuracy and the highest robustness (lowest $\left|\mathrm{AUC}_{\mathrm{valid}} - \mathrm{AUC}_{\mathrm{train}}\right|$) among the options available.

Undersampling the training data little improved decision tree models, whereas no benefits were seen on logistic regression models.

Testing the champion model it was found, as expected, very small I-type error and considerable type-II error: almost every non-churner (98.8 %) was predicted as such, versus 11.5 % of churners correctly classified. This is due to the strong imbalance between the two classes.

The cumulative lift chart demonstrated reasonably good performance at the top percentiles. At the 5th percentile, it was 3.88. This result was judged satisfactory from both managers and business experts, being it an improvement over previous in-house attempts. The most important attribute by far was the difference between the premium at the current year and the premium at the previous year. Businesswise, this variable can be translated into the amount of upsells/downsells year-to-date: the more often the customer withdraws its insurance coverages from its portfolio, the more likely it will churn. Other important variables are: the time since the last record mutation, interpreted as the number of months since the last time the contract has been changed; and a Boolean variable that gives information on whether the contract is about to expire or not.

Conclusion

In saturated markets such as the commercial line insurance, predictive analytics tools for churn behaviour are important support for marketing decision-making. In order to predict which contracts will be cancelled within the next 12 months in an insurance company, there have been built decision trees and logistic regression models. The key challenges that arose for the problem at hand were the following:

1. **Giving a definition of “churn” that is consistent with the project objectives**. There is not a universal definition of churn: the definition depends on company-specific business objectives, but also technical constraints. In this study we mapped all possible churn events, distinguishing between deliberate and non-deliberate churn. However, this theoretical distinction did not find practical application due to the poor quality of the data stored in the database.

2. **Dealing with class imbalance**. Churn is a rare—yet not negligible—event. From a modelling perspective this entails working with data sets biased towards the non-churn event. This is considered a problem because the cost of misclassifying a churner (false negative) is much higher than the cost of misclassifying a non-churner (false positive). For this reason, three approaches have been explored: cost-sensitive learning, undersampling, ensemble learning algorithms. It is worth mentioning that undersampling improved decision tree models, but not logistic regression models, which are by their nature more resistant to these biases.
3. **Selecting relevant variables**. The insurance business is characterised by large volumes of data, spanning from customer demographics to their behaviour and claims history. Appropriate algorithms for dimensions reduction were forward selection methods and tree-based approaches.
4. **Coping with trade-off decisions between ease of interpretation and accuracy**. In business contexts, the most accurate model is not necessarily the best suited to the problem at hand. Rather, predictive performance and deployability of the model should be considered together and weighted each according to the client needs and expertise. Indeed, the model chosen in this study was very simple and intuitive, as it learned from only nine variables. Since final users, i.e. underwriters and salespersons, are not expected to have a statistical modelling background, an easy-to-interpret model increases dramatically the chances of success of the retention strategy.

The model constructed was a cost-sensitive tree based on a CS-C4.5 algorithm. It fulfilled both requirements of performance and deployability. On the one hand, its performance was better than any previous model developed in the past by the company. On the other, it was pruned in such a way that the structure was simple and intuitive.

References available upon request.