Prepaid Churn Prediction

By Michael Constantinou

BUSINESS SCIENCE CORPORATION

Overview

The purpose of this paper is to outline the process and churn prediction model for a telecommunications operating company. The chosen model will be formulated using advanced data mining techniques, where the overall model performance criterion will be discussed and evaluated.

For the successful implementation of a data mining model, there needs to be alignment between business and data mining objectives. With this in mind, a proposed framework will be outlined and the data mining process documented.

Churn Types

It is a well-documented fact that retention cost is substantially lower than the acquisition of a new subscriber. In addition to the cost-saving benefit in churn prevention, there is the realization of a long-term continuous stream of revenue which would have otherwise been lost by increasing the customer lifetime value.

There are two basic categories of churn, voluntary and involuntary. Involuntary churners are the subscribers that the telecommunication company decides to remove for reasons such as fraud and non-payment. On the other hand, voluntary churn can be described as the termination of service by the subscriber. This is further categorized into two types, namely, deliberate and incidental. Incidental churn is unplanned and could be related to factors such as financial circumstances or relocation.

The focus of this paper is on the latter type. Deliberate churn occurs due to a number of factors which include, amongst others, price dynamics i.e. price sensitivity to competitor offerings and overall quality of service.

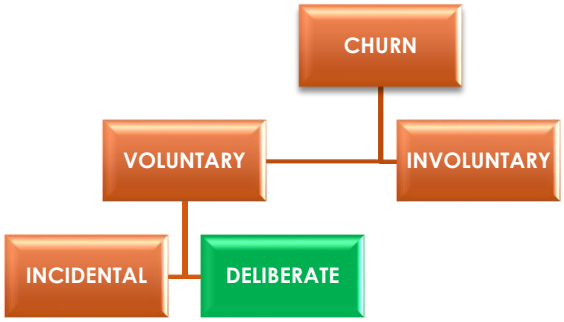

Figure 1: Breakdown of churn

Objectives & Framework

In order to successfully create and implement a churn prediction model, there needs to be a process and framework in place. The process begins with a clear definition of the business objectives. These objectives need to be realistic but at the same time quantifiable e.g. reduce prepaid churn in 3 months by 15 %. Coupled with this is the goal of the prediction model e.g. have an accuracy rate of 80% in the prediction of subscribers likely to churn.

Once objectives have been decided, consideration is then given to factors such as data availability, cleansing and final transformation. Data preparation is by far the most time consuming element in the

entire process. The required data is often located in disparate locations which need to be integrated into a central source. If the organization has a relational database management system, this will require a specialist with a strong SQL (Structured Query Language) background to extract the necessary information. Research suggests that as much as 70% of the entire prediction model development is spent on data preparation.

The model construction element is an iterative process and various models based on different algorithms need to be tested and compared. Not all factors used in the model will be beneficial and hence the need to run numerous iterations. There are several types of models used in churn prediction, and some of the more commonly used are Decision Trees, Logistic Regressions, Neural Networks and K-means Clustering.

Before the modelling begins, data is typically divided into two distinct groups; the training and testing set. The split is approximately 70% training and 30% test. The model is constructed using the training data and then verified on the unseen test data to evaluate model performance.

A gains chart / lift curve is a graphical interpretation of model performance. For our purpose, it illustrates what percentage of subscribers would need to be targeted in order to reach a certain percentage of all likely churners. In figure 2, the X-axis shows the percentage of the population we plan to target while the Y-axis displays the percentage of churners we are likely to contact.

The central curve is known as the random guess model i.e. there is a 50% probability of identifying all the churners in the population if you select 50% of the population.

The upper curve is indicative of the perfect model that would achieve 100 % accuracy by selecting only churners from the population, in this case, by targeting 30% of the population. The aim is to develop a model with a curve as close to the ideal model as possible and above the central curve.

Once the model has been tested, it will need to undergo a final verification called cross-validation. This technique assists in the development and fine-tuning of the model by partitioning the training data set into cross-sections, using one of the partitions as a new test set, and the remaining partitions as the training set. This process is repeated several times confirming model robustness.

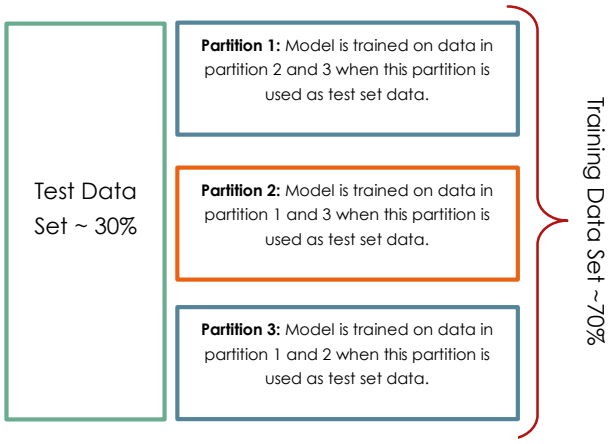

Figure 3: Cross-validation methodology

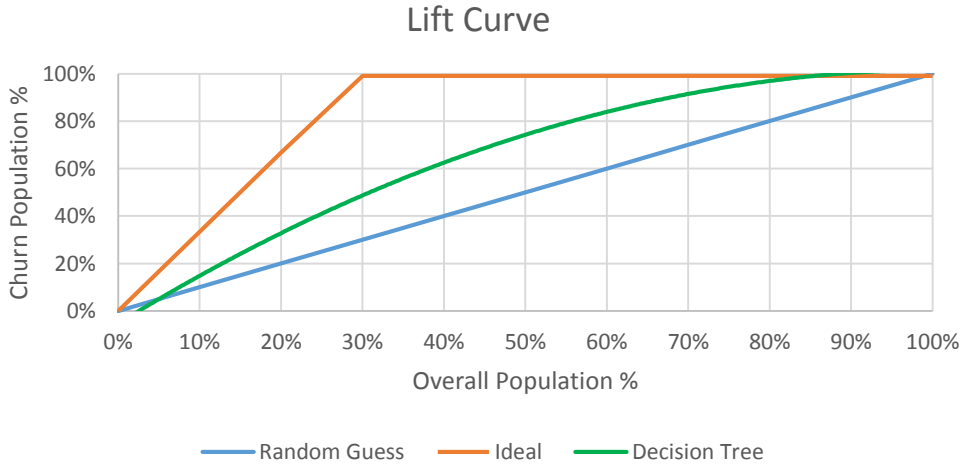

Figure 2: Lift Chart of mining model

The final two phases of the process are deployment and monitoring. At this stage, the model will be applied to the existing subscriber base where each individual MSISDN will be assigned a likelihood or probability of churning. There will be a minimum threshold applied to the probability, and anything over and above will be flagged as likely to churn. With this information, the Customer Value Management (CVM) team will develop campaigns to entice the subscriber to remain on the network.

It should be noted that not all likely churners should be treated equally. Additional segmentation models should be run to profile the likely churners into different value groups to access the type of campaign that should be executed, if any, bearing in mind the inherent costs of running a campaign. Campaigns should be designed not only to retain the subscriber but also to seek potential cross-sell and upsell opportunities to drive revenue and increase tenure.

The monitoring of any deployed model is critical. Model performance may change with time due to exogenous factors which could change subscriber behavior, rendering the model ineffective. Constant monitoring and recalibration is necessary to ensure the churn prediction model remains relevant.

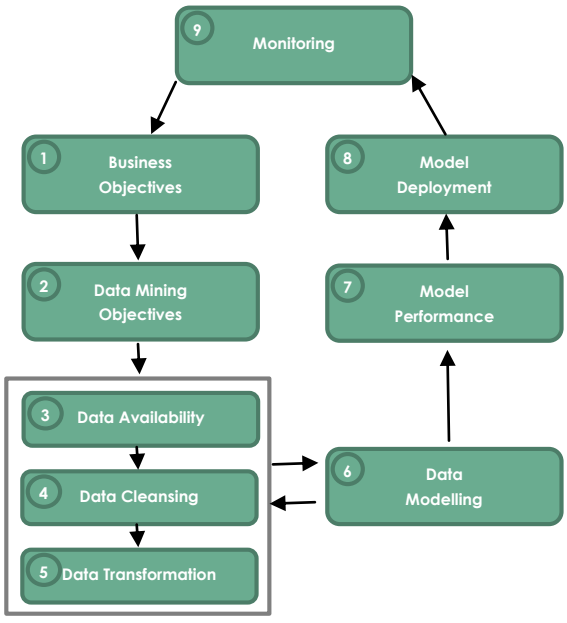

Figure 4: Framework for development of churn prediction model

Churn Classification

Unlike postpaid subscribers, the prepaid base has no contractual obligation to the telecom operator. With this in mind, the postpaid churn date is the date at which the customer disconnects from the network. In contrast, the deactivation date for prepaid subscribers is not necessarily the churn date. In many instances the definition of prepaid churn is subjective and could be classified as a period of inactivity on the network spanning 90 days. If this is the case, then the deactivation date is not a suitable indicator for churn date. The definition of prepaid churn is the date at which the subscriber indefinitely stopped using their SIM-card.

Model Attributes

The dataset used in the model prediction is aggregated monthly, and is classified into different groups i.e. subscriber details, usage traffic by bearer and revenue (recharges). The attributes of interest can vary between telecomm operators and not all will be applicable and available. A technique to assist in deciding what attributes will be valuable is to formulate different hypothesis as why subscribers are churning. For example, "*Subscribers who have a distinct change in off-net behavior are likely to churn.*" If you call more subscribers off-net, this could be an early indication for a change in calling circle behavior, and ultimately the decision to leave the network because of the higher off-net costs incurred.

Below is an example of the types of attributes one could utilize for prediction.

Subscriber Details:

- subscriber MSISDN
- tenure on network
- no. of days active / inactive in a month
- prepaid tariff name
- location (region, municipality, territory)
- most commonly used cell site
- segment (based on internal rule set)
- distinct number of MSISDNS called / received calls (using CDR information)
- connecting channel
- churn indicator (Y/N)

Recharge Detail:

- maximum recharge amount
- total recharge amount
- average recharge amount
- no. of recharges

Usage Details:

- total minutes of use (MOU)
- total MOU mobile originating (on-net / off-net)
- total MOU mobile terminating (on-net / off-net)
- total MOU in-bundle / out-bundle
- no. of voice bundles purchased

- total SMS's sent
- total SMS mobile originating (on-net / off-net)
- total SMS mobile terminating (on-net / off-net)
- total SMS in-bundle / out-bundle
- no. of SMS bundles purchased

- total data usage (upload and download)
- total data in-bundle / out-bundle
- no. of data bundles purchased

Time Period Outline

In order to understand the factors driving churn, subscriber behavior over time must be documented. For the purpose of the model, a minimum of 6 months data from the month preceding the churn month (as defined above) should be extracted. A condition of tenure greater than or equal to 6 months also needs to be imposed to improve data accuracy. The 6 month time period must be divided into two sub-periods each three months in length. At this stage, per time period, each usage and recharge attribute needs to be averaged i.e. a rolling 3 month average. The delta % between the time periods then needs to be computed.

For example:

$$\%\Delta MOU = \frac{Ave\ MOU\ (period\ 2) - Ave\ MOU\ (period\ 1)}{Ave\ MOU\ (period\ 1)}$$

Following the computation of the deltas, appropriate BIN's should be created and each subscriber's delta apportioned to the correct BIN. This is the information that will be divided into the training and test datasets for model computation.

Model Evaluation

Once the model has been created and implemented, the performance is then scrutinized. Lift curves (as described above) can be used as a graphic interpretation. However, it is useful to compute a single metric for overall model performance. A useful tool to aid in this computation is a confusion matrix. This matrix allows for the visualization of the performance of the data mining algorithm by summarizing the results of the model.

For example:

Say we had 360 cases in the test dataset to which the model was applied, and the results of the confusion matrix were as follows:

		Model Prediction	
	Actual	Churner	Non-churner
	Churner	100	12
	Non-churner	13	235

These results as percentages would translate to:

		Model Prediction	
	Actual	Churner	Non-churner
	Churner	89%	11%
	Non-churner	5%	95%

The above does give a useful breakdown per model. However, it would be more beneficial to have the ability to compare multiple models. For this we use two performance metrics i.e. Accuracy and Error rate.

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

$$Error\ rate = \frac{\text{Number of incorrect predictions}}{\text{Total number of predictions}}$$

Model	Accuracy	Error rate
Decision Tree	93%	7%
Logistic Regression	87%	13%
Neural Network	83%	17%

Conclusion

Having the capability to accurately predict subscribers at risk of churn, with a high degree of certainty is invaluable to telecom companies. Data mining and predictive analytics is becoming more vital in assisting companies to remain relevant and competitive. Being able to target the correct individual, at the best possible time with the most attractive offering will assist in customer retention and revenue generation.

This paper provides one of many different data mining approaches to prepaid churn prediction within the telecoms industry. A key component to any successful data mining project is the establishment of the correct framework. Data mining has the ability to pick up trends and patterns that would otherwise be unattainable. However, not all such trends are useful to an organization, hence the need for alignment between business and data mining objectives from the outset.

References

[1] –Essam Shaaban, Yehia Helmy, Ayman Khedr, Mona Nasr "A proposed churn prediction model", International Journal of Engineering

[2] - Goran Kraljevi´c, Sven Gotovac, "Modeling Data Mining Applications for Prediction of Prepaid Churn in Telecommunication Services"

[3] - Rahul J. Jadhav, Usharani T. Pawar, "Churn Prediction in Telecommunication Using Data Mining Technology"

[4] – Ali Tamaddoni Jahromi, "Predicting Customer Churn in Telecommunications Service Providers"

[5] – Han Lai, "PayPal Survey Analysis & Churn Risk Detection"

[6] – Khalida binti Oseman, Sunarti binti Mohd Shukor, Norazrina Abu Haris1, Faizin bin Abu Bakar, "Data

Mining in Churn Analysis Model for Telecommunication Industry"

http://technet.microsoft.com, "Cross-Validation (Analysis Services Data Mining)

BUSINESS SCIENCE CORPORATION

Contact Us:

BSC provides leading analytical and modelling services to help organisations grow their revenues with deliberate precision.

Email us today to show you how we can do the same for your business:

Email: info@BSCglobal.com