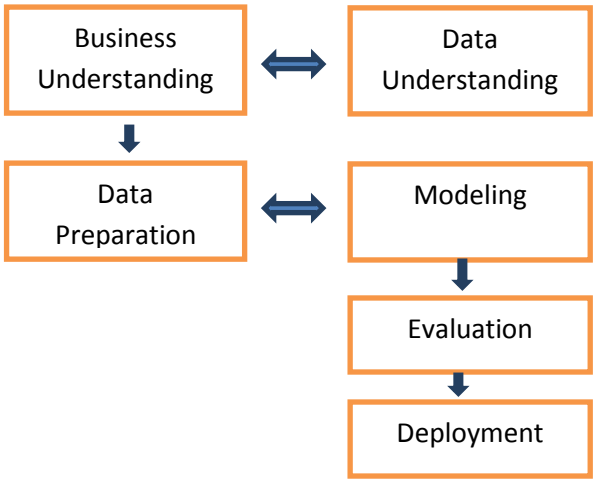
Case Study: Customer churn analysis In Telecommunication sector

Introduction

Churn in the broadest sense is a measure of the number of individuals or items moving out of a collective system over a specific period of time. It is one of two primary factors that determine the steady-state level of customers a business supports. Churn prediction aims to detect customers intended to leave a service provider. Churn prediction is of high importance because cost of losing a customer is 5 to 10 times greater than gaining a new customer. Predictive models can provide correct identification of possible churners in the near future to design a retention solution.

The relationship between customer and company is one of the prominent determinants of the progress of any company. To be able to retain the old customers is a challenge in itself and the addition of new customers partially depends on the feedback of old customers. Hence, it is imperative to predict the churn count and make necessary reforms for a better customer satisfaction. This case study explains how to predict the number of churners and their reasons for churning.

Stages of Implementation



Business & data Understanding

Before proceeding to the prediction of customer churn, one needs to understand the business and the environment. After understanding the business and the services it provides, the data is understood with the help of the documentation provided.

Data Preparation

Data preparation is not only the most important phase but also the most time-consuming phase in a data mining process. Data selected should represent enough quantity of data in a given period of time. In this phase, data is collected, integrated and filtered. Integration of data may require extracting from multiple sources. The data is in a fully characterized state and is ready to model once it is in a tabular format. Data is cleaned by resolving any ambiguities, errors, and by removing redundant and problematic data. Finally, the table is divided, if required, into subsets in order to optimize the performance of the database. It simplifies the analysis and enhances the performance to perform queries.

Modeling Phase

In this phase, an appropriate model is developed for future predictions that satisfy the main objectives. Once the model has been selected, different parameters are obtained to make improvement on the results. Then the model is evaluated in the next phase.

Evaluation Phase

For a model to be rendered fit, evaluation is important. Evaluation of the response time, confidence level, cost, error rate and the fitness of the model for the defined objectives is done. Based on the accuracy of prediction, the model is refined and re-evaluated.

Deployment Phase

After the evaluation phase, the model is then deployed to the real time environment to make predictions.

Implementation

Problem Statement

The aim is to predict the number of churners and the reasons for the customers to churn in a telecommunication sector.

Dataset

The dataset consists of all the new customers that have been added and the customers that have churned to move to other telecommunication companies. The dataset includes the demographic data, account information, and the complaint data.

Dataset Preparation

First we have chosen Billing Account No as unique and customer ID is linked to Billing Account No to access a particular customer. Then the data in different files are concatenated in one single file for modelling.

Demographic Data

This data file consists of demographic information like age, Gender etc., and are differentiated using the unique Customer ID.

Account Information data

This consists of variables related to Account information data like average call rate, average number of call per month etc. and are differentiated using Billing Account Number

Complaint data

Variables related to complaint information's like complaints, complaints etc. and are again identifies using the Billing Account Number.

The data is combined provided in different tables using Excel with the help of the function *Vlookup.* The complaints are counted using *count if*.

The final data has a compiled version of all the variables and is ready for modelling.

Logistic Regression Model

Logistic regression or Logit regression is a type of regression analysis used for predicting the outcome of a categorical dependent variable on one or more independent variables.

Training Set

If the prediction has to be done for a particular year, all the non-churned customers prior to the year and the newly added customers in the first three quarters of the year is considered as the training set. The duplicate records which are characterised by the repeated Customer ID of the data are removed thereby leaving the data with only unique records. Then the logistic regression for this data set is performed with the billing account no as the Unique ID.

Model Test

Model Testing is another important aspect of model building. To test the contribution of each variable in the model, Type III (Wald) tests have been adopted. These tests help in assessing the importance of each variable in the model independently.

The model is tested by type **III (Wald) tests method** which gives ChiSq value as an output which is used as the filter the variables. The variable selection criterion is to accept those variables with ChiSq values near to zero. A threshold is chosen to filter according to the model which is generally of the order 10^{-3}. The variables satisfying the criterion are accepted and the model equation is generated.

Model Equations

The model equation is generated by using R, open source software that has an inbuilt toolbox for Logistic Regression. A typical model equation generated will be a representative of the variables along with their weighted contribution to the prediction of customer churn.

Testing Set

The testing set is used to validate the model accuracy performed on the training data. The testing set is prepared by considering the non-churned customers prior to the year for which the prediction has to be done, as well as the first three quarters appended to the newly added customers in the last quarter. By using training set we predict the testing set data.

Results

Formula for Accuracy & Error Rate:

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

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

Using the above formulas tested the model we got 95.13% Accuracy & 4.97% Error rate.

Conclusion

Many churn prediction models and techniques have been presented till date. However, a simple model is required to distinguish churners from non-churners then clustering the resulted churners for providing retention solutions. In This Project, a simple model based on Logistic Regression techniques was introduced to help a CRM department to keep track its customers and their behaviour against churn. A data set of one year instances with some attributes is used to train and test the model. Using Logistic regression techniques the next stage research will involve performing a deeper analysis into the customer data to try to establish new churn prediction retention model that will use the predicted data to assign a suitable retention strategies for each churner type.