Telecom customer churn prediction

Telecom Customer Churn Prediction

Abhijit Sundararajan

Department of MSIS, Rutgers University
E-mail: abhijit.sundararajan@rutgers.edu

Kemal Gursoy

Department of MSIS, Rutgers University,
E-mail: kgursoy@business.rutgers.edu

Abstract

Customer churn is often referred to as customer attrition, or customer defection which is the rate at which the customers are lost. Telecom companies often use customer churn as a key business metrics to predict the number of customers that will leave a telecom service provider. Churn is significant in the telecommunication industry because it directly affects the competitiveness of the service provider. Churn is the proportion of clients leaving the service provider. The service provider should therefore find new clients to preserve profitability. This is not possible due to the high and difficult cost of acquiring new customers. The cost of retaining an existing customer is far less than acquiring a new one, so telephone service companies, Internet service providers, pay TV companies, insurance firms and alarm monitoring services use the customer churn to predict profitability. Businesses basically believe that a customer has churned with the amount that has passed since the customer's last interaction with the site or service. The full cost of customer churn includes both lost revenue and the marketing costs involved with replacing those customers with new ones. Reducing customer churn is important because cost of acquiring a new customer is higher than retaining an existing one. This case is related to telecom industry where organizations want to know that for given certain parameters whether a person will churn or not.

1. Introduction

Customer churn is a common problem across several business in different industries. There is a substantial amount of financial loss associated with churning, as the businesses must invest huge sum of money in getting new customers. Time and efforts need to be invested into replacing the customers that are lost. Businesses need to be able to predict when a client is likely to leave and offer them lucrative incentives to stay. The most important essence of using predictive analysis is to determine the customers that are likely to churn.

Companies typically make a difference between voluntary churn and involuntary churn. Voluntary churn occurs because of a selection by customer to replace to any other organisation or provider, involuntary churn happens because of instances inclusive of a client's relocation for an extended-time period in a care facility, loss of life, or the relocation to a distant region. In maximum packages, involuntary reasons for churn are excluded for analytical purposes. Analysts tend to concentrate on voluntary churn, as it normally occurs because of factors of the corporation-customer courting which agencies manage, which includes how billing interactions are treated or how after-income assistance is provided. Predictive analytics use churn prediction fashions that expect purchaser churn through assessing their propensity of chance to churn. Seeing that these fashions generate a small prioritized list of ability defectors, they're effective at focusing consumer retention advertising packages at the subset of the consumer base who're most liable to churn.

There are 2 types of churners: Voluntary churner and Involuntary Churner.

Involuntary Churners: This type includes the customers that are removed by the service

providers from the subscription list. The customers that come under this category are the customers who cheat or are churned for fraudulent activities, the customers who do not pay their subscription charges and the customers who do not use the services.

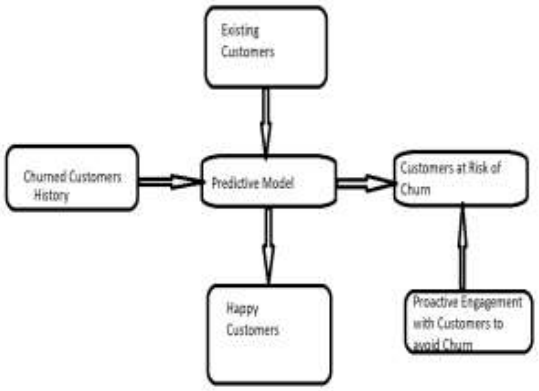
Voluntary Churners: This type includes the customers that voluntarily decide to leave and terminate their services. Voluntary churn can be subdivided into 2 categories: Incidental and Deliberate churn. Incidental churn occurs due to location change, financial issues and can happen even due to bad experiences. Deliberate churn occurs when the customer wants to upgrade to new service providers with better technologies. It also happens due to price sensitivity, quality of the services provided, social and psychological factors and convenience factors.

Through means of a Kaggle competition, the objective is to find an appropriate model to predict whether a customer may churn from a particular service provider in the near future. Having this model in place can ensure that Santander can take proactive steps to improve a customer's happiness before they would take their business elsewhere. First the paper will discuss related work done on this. Secondly we do an extensive exploratory analysis of the data, analysing groups of variables and individual features to give us insight in what is relevant. Thirdly several cleaning procedures that were employed to lead to better results are outlined. Fourthly we explain the performance measure of this competition and the three models to tackle the problem.

2. Today's Approach

The data set includes information about:
Customers that churned during the last month – the column is called Churn.
Services that each customer has been using– phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.
Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges.
Demographic information about the customers – gender, age range, and if they have partners and dependents

Figure 1.



Dataset is from the IBM Watson Repository and available in Kaggle also. The data set consists of 7043 rows & 21 columns. It has 17 categorical variables, 3 numerical variables and 1 unique variable i.e. customerID. The row consists of Customers, columns consists of Customer Attributes and Target Variable consists Churn
The questions we would like to answer through Machine Learning algorithms include:
What are the demographics of people who choose to stay/leave? What kind of services do the existing customers subscribe to? Does loyalty and money billed influence the churn? Growing companies and those expanding their product range usually segment their customers using previously defined and selected features. Customers can be divided into subgroups based on their lifecycle stage, needs, used solutions, level of engagement, monetary value, or basic information. Since every customer category shares common behaviour patterns, it's possible to increase prediction accuracy through the use of ML models trained specifically on datasets representing each segment.

3. Models

1. Logistic regression

This is used for classification problem. We can use logistic regression for this problem. Logistic Regression is used when the dependent variable is categorical. It has a bias towards classes which have large number of instances. It tends to only predict the majority class data. The features of the minority class are treated as noise and are often ignored. Thus, there is a high probability of misclassification of the minority class as compared to the majority class. Suppose in a dataset there are

2 categories, then logistic regression models the probability that Y belongs to a particular category.

2. Random Forest

Random is used for classification problems. It is a combination of tree predictors such that each tree depends on the value of a vector randomly sampled and sampled independently and with the distribution of all the trees in the forest. After many trees are generated, they vote for the most popular class. This procedure is called random forests. There have been significant improvements in the classification accuracy from growing an ensemble of trees and letting them vote for the most popular class.

3. Support Vector Machine

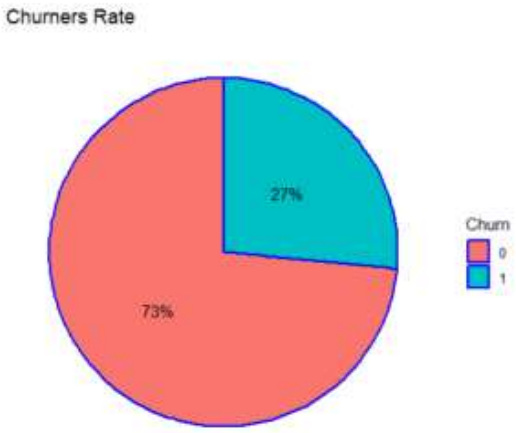
Support Vector Machine is a supervised learning algorithm which is used for both classification and regression challenges. it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space. where n is number of features you have) with the value of each feature being the value of a coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes

4. Decision Tree

Decision trees can be applied to both classification and regression problems. It is used to predict a qualitative response rather than a quantitative response. We predict that each of the observation belongs to the most commonly occurring class. It is a type of supervised learning algorithm with a predefined target variable. While mostly used in classification tasks, it can handle numeric data as well. This algorithm splits a data sample into two or more homogeneous sets based on the most significant differentiator in input variables to make a prediction. With each split, a part of a tree is being generated. As a result, a tree with decision nodes and leaf nodes (which are decisions or classifications) is developed. A tree starts from a root node – the best predictor. Out of three variables we use, Contract is the most important variable to predict customer churn or not churn. If a customer in a one-year or two-year contract, no matter he (she) has Paperless billing or not, he (she) is less likely to churn. On the other hand, if a customer is in a month-to-month contract, and in the tenure group of 0–12 month, and using Paperless billing, then this customer is more likely to churn.

4. Experiments

Figure 2.



We can infer from the above figure that 27% of the customers in the dataset from Kaggle have churned.

5. Comparison of Methods

Models	Accuracy Score
Random Forest Classifier	0.9355
SVM	0.8192
Decision Tree	0.762
Logistic Regression	0.7894

6. Results and Discussion

We observe that the Support Vector Machine model is a good fit. This is due to the AUC values for training and test are high and similar. We may consider improving the SVM model by performing a grid search for values of C, gamma and degree of the kernel that gives higher accuracy using k-fold cross-validation.

Conclusion

From the above example, we can see that SVM, Logistic Regression and Random Forest performed better than Decision Tree for customer churn analysis for this dataset. The following things can be observed from the dataset:

1. Attributes and features such as tenure group, Contract, Paperless Billing, Monthly Charges and Internet Service appear to play a role in customer churn.
2. There seems to be no relationship between the gender and the churn rate.
3. Customers having a service plan of month-to-month contract, with Paperless Billing and are within 12 months tenure, are more likely to churn. On the other hand, customers with one- or two-year contract, with longer than 12 months tenure, that are not using Paperless Billing, are less likely to churn.

References

1. https://www.datasciencecentral.com/profiles/blogs/customer-churn-logistic-regression-with-r
2. https://pdfs.semanticscholar.org/75d3/73f987be5c2fb5a3cb1830f417c63f09a68d.pdf
3. https://datascienceplus.com/predict-customer-churn-logistic-regression-decision-tree-and-random-forest/
4. https://www.kaggle.com/pavanraj159/telecom-customer-churn-prediction
5. https://towardsdatascience.com/hands-on-predict-customer-churn-5c2a42806266