Telecom Churn Case Study

Business Problem Overview

- In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another.
- In this highly competitive market, the telecommunications industry experiences an average of 15-25% annual churn rate.
- The fact that it costs 5-10 times more to acquire a new customer than to retain an existing one, customer retention has now become even more important than customer acquisition.
- For many incumbent operators, retaining high profitable customers is the number one business goal.
- To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.
- In this project, we will analyse customer-level data of a leading telecom firm, build predictive models to identify customers at high risk of churn and identify the main indicators of churn.

Understanding and Defining Churn

•There are two main models of payment in the telecom industry –

• In the postpaid model: when customers want to switch to another operator, they usually inform the existing operator to terminate the services, and we directly know that this is an instance of churn. (Here, customers pay a monthly/annual bill after using the services) and

•However, in the prepaid model, customers who want to switch to another network can simply stop using the services without any notice (customers pay/recharge with a certain amount in advance and then use the services)., and it is hard to know whether someone has actually churned or is simply not using the services temporarily (e.g. someone may be on a trip abroad for a month or two and then intend to resume using the services again).

•Thus, churn prediction is usually more critical (and non-trivial) for prepaid customers.

•There are various ways to define churn, such as:

•Revenue-based churn**:** Customers who have not utilized any revenue-generating facilities such as mobile internet, outgoing calls, SMS etc. over a given period of time. One could also use aggregate metrics such as ‘customers who have generated less than INR 4 per month in total/average/median revenue’. The main shortcoming of this definition is that there are customers who only receive calls/messages from their wage-earning counterparts, (i.e. they don’t generate revenue but use the services.) For example, many users in rural areas only receive calls from their wage-earning siblings in urban areas.

•Usage-based churn**:** Customers who have not done any usage, either incoming or outgoing - in terms of calls, internet etc. over a period of time.

A potential shortcoming of this definition is that when the customer has stopped using the services for a while, it may be too late to take any corrective actions to retain them. For e.g., if we define churn based on a ‘two-months zero usage’ period, predicting churn could be useless since by that time the customer would have already switched to another operator.

Understanding the Business Objective and the Data

- The dataset contains customer-level information for a span of four consecutive months - June, July, August and September. The months are encoded as 6, 7, 8 and 9, respectively.
- Here, in this project, we will use the usage-based definition to define churn.
- In the Indian and the southeast Asian market, approximately 80% of revenue comes from the top 20% customers (called high-value customers). Thus, if we can reduce churn of the high-value customers, we will be able to reduce significant revenue leakage. Hence, this case study focuses on high value customers only.
- Customers usually do not decide to switch to another competitor instantly, but rather over a period of time (this is especially applicable to high-value customers).The business objective is to predict the churn in the last month using the data (features) from the first three months.

The following data problem:

•Derive new features : This is one of the most important parts of data preparation since good features are often the differentiators between good and bad models. We will use our business understanding to derive features that we think could be important indicators of churn.

• Filter high-value customers : As mentioned above, we need to predict churn only for the high-value customers. Define high-value customers as follows: Those who have recharged with an amount more than or equal to X, where X is the 70th percentile of the average recharge amount in the first two months (the good phase).

•Tag churners and remove attributes of the churn phase : Now tag the churned customers (churn=1, else 0) based on the fourth month as follows: Those who have not made any calls (either incoming or outgoing) AND have not used mobile internet even once in the churn phase. The attributes we need to use to tag churners are:

total_ic_mou_9
total_og_mou_9
vol_2g_mb_9
vol_3g_mb_9

After tagging churners, we need to remove all the attributes corresponding to the churn phase (all attributes having ‘ _9’, etc. in their names).

Modelling

•Build models to predict churn. The predictive model that we are going to build will serve two purposes:

•It will be used to predict whether a high-value customer will churn or not, in near future (i.e. churn phase). By knowing this, the company can take action steps such as providing special plans, discounts on recharge etc.

•It will be used to identify important variables that are strong predictors of churn. These variables may also indicate why customers choose to switch to other networks.

•In some cases, both of the above-stated goals can be achieved by a single machine learning model. But here, we have a large number of attributes, and thus we should try using a dimensionality reduction technique such as PCA and then build a predictive model. After PCA, we can use any classification model.

•Also, since the rate of churn is typically low (about 5-10%, this is called class-imbalance) – by using techniques to handle class imbalance.

We can take the following suggestive steps to build the model:

•Preprocess data (convert columns to appropriate formats, handle missing values, etc.)

•Conduct appropriate exploratory analysis to extract useful insights (whether directly useful for business or for eventual modelling/feature engineering).

•Derive new features.

•Reduce the number of variables using PCA.

•Train a variety of models, tune model hyperparameters, etc. (handle class imbalance using appropriate techniques).

•Evaluate the models using appropriate evaluation metrics. Note that it is more important to identify churners than the non-churners accurately - choose an appropriate evaluation metric which reflects this business goal.

•Finally, choose a model based on some evaluation metric.

Therefore, we can build another model with the main objective of identifying important predictor attributes which help the business understand indicators of churn. A good choice to identify important variables is a logistic regression model or a model from the tree family and to handle multi-collinearity.

After identifying important predictors, display them visually - we can use plots, summary tables etc.

Finally, recommend strategies to manage customer churn based on our observations.

Recommendations

- Target the customers, whose minutes of usage of the incoming local calls and outgoing ISD calls are less in the action phase (mostly in the month of August).
- Target the customers, whose outgoing others charge in July and incoming others on August are less.
- Also, the customers having value based cost in the action phase increased are more likely to churn than the other customers. Hence, these customers may be a good target to provide offer.
- Cutomers, whose monthly 3G recharge in August is more, are likely to be churned.
- Customers having decreasing STD incoming minutes of usage for operators T to fixed lines of T for the month of August are more likely to churn.
- Cutomers decreasing monthly 2g usage for August are most probable to churn.
- Customers having decreasing incoming minutes of usage for operators T to fixed lines of T for August are more likely to churn.
- That means for the customers, whose roaming outgoing minutes of usage is increasing are more likely to churn.

Conclusion

•After trying several models we can see that for achieving the best sensitivity, which was our ultimate goal, the classic Logistic regression or the SVM models performs well. For both the models the sensitivity and accuracy was good.

•We can see that the logistic model has good sensitivity and accuracy, which are comparable. So, we can go for the more simplistic model such as logistic regression as it explains the important predictor variables as well as the significance of each variable. The model also helps us to identify the variables which should be act upon for making the decision of the to be churned customers. Hence, the model is more relevant in terms of explaining to the business.