Business Problem Overview

Let us say that Reliance Jio Infocomm Limited approached us with a problem. There is a general tendency in the telecomm industry that customers actively switch from one operator to another. As the telecomm industry is highly competitive, the telecommunications industry experiences an average of 18% - 27% annual churn-rate. Since, it costs 7 -12 times more to acquire a new customer as compared to retaining an existing one, customer retention is an important aspect when compared with customer acquisition which is why our client, Jio, wants to retain their high profitable customers and thus, wish to predict those customers which have a high risk of churning.

Also, since a post-paid customer usually informs the operator prior to shifting their business to a competitor's platform, our client is more concerned regarding its prepaid customers that usually churn or shift their business to a different operator without informing them which results in loss of business because Jio couldn't offer any promotional scheme in time to prevent churning.

As per Jio, there are two kinds of churning - revenue based and usage based. Those customers who have not utilized any revenue-generating facilities such as mobile data usage, outgoing calls, caller tunes, SMS etc. Over a given period of time. To determine such a customer, Jio usually uses an aggregate metric like 'customers who have generated less than ₹7 per month in total revenue'. However, the disadvantage of using such a metric would be that many of Jio customers who use their services only for incoming calls will also be counted/treated as churn since they do not generate direct revenue. In such scenarios, revenue is generated by their relatives who also uses Jio network to call them. For example, many users in rural areas only receive calls from their

wage-earning siblings in urban areas.

The other type of Churn, as per our client, is usage based which consists of customers who do not use any of their services i.e., no calls (either incoming or outgoing), no internet usage, no SMS, etc. The problem with this segment is that by the time one realizes that a customer is not utilizing any of the services, it may be too late to take any corrective measure since the said customer might already switched to another operator. Currently, our client, Reliance Jio Infocomm Limited, have approached us to help them in predicting customers who will churn based on the usage-based definition

Another aspect that we have to bear in mind is that as per Jio, 80% of their revenue is generated from 20% of their top customers. They call this group High-valued customers.Thus,if we can help reduce churn of the high-value customers, we will be able to reduce significant revenue leakage and for this they want us to define high-value customers based on a certain metric based on usage-based churn and predict only on high-value customers for prepaid segment.

Understanding the Data-set

The data-set 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. The business objective is to predict the churn in the last (i.e. the ninth) month using the data (features) from the first three months. To do this task well, understanding the typical customer behavior during churn will be helpful. You can download your data-set here.

Understanding Customer Behavior During Churn

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). In churn prediction, we assume that there are three phases of customer life-cycle:

1) The 'good' phase: In this phase, the customer is happy with the service and behaves as usual.

2) The 'action' phase: The customer experience starts to sore in this phase, for e.g. he/she gets a compelling offer from a competitor, faces unjust charges, becomes unhappy with service quality, etc. In this phase, the customer usually shows different behavior than the 'good' months. Also, it is crucial to identify high-churn-risk customers in this phase, since some corrective actions can be taken at this point (such as matching the competitor's offer/improving the service quality, etc.)

3) The 'churn' phase: In this phase, the customer is said to have churned. You define churn based on this phase. Also, it is important to note that at the time of prediction (i.e. the action months), this data is not available to you for prediction. Thus, after tagging churn as 1/0 based on this phase, you discard all data corresponding to this phase.

In this case, since you are working over a four month window, the first two months are the 'good' phase, the third month is the 'action' phase, while the fourth month is the 'churn' phase.

Data Dictionary

The data-set is available in a .CSV file named as “Company Data.csv” and the data dictionary has been provided in a separate file named as “Data Dictionary.xlsx” The data dictionary contains meanings of abbreviations as well. Some frequent ones are loc (local), IC (incoming), OG (outgoing), T2T (telecom operator to telecomoperator), etc. The attributes containing 6, 7, 8, 9 as suffix imply that those correspond to the months 6, 7, 8, 9 respectively.

The following data preparation steps are crucial for this problem:

1. Derive new features

Use your business understanding to derive features you think could be important indicators of churn. This is one of the most important parts of data preparation since good features are often the differentiating factors between good and bad models.

2. Filter 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 a month in the first two months (the good phase).

3. 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 you 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, remove all the attributes corresponding to the churn phase (all attributes having '9', etc. In their names).

Project:

i. Clean the Data-set and prepare it for model building

ii. Derive at least 2 new features based on available variables in the data-set. Justify your reasons on how the derived features will add-on to your model prediction

iii. There might be class imbalance present when you define your target variable. Use any technique necessary to handle it.

iv. Perform EDA and visualize it using matplotlib, seaborn and plotly (at least 3 different graphs from each library) and present the insights drawn from each of them

v. Clearly define which metric or metrics will you use to evaluate your model. Justify your answer

vi. Create 2 logistic models - one with PCA for dimensionality reduction and another without PCA using all features and compare their results w.r.t accuracy and your chosen metrics

vii. In part (v) above, using PCA will improve computational time but will have an impact on interpret-ability of the model. Hence, in your

opinion, should we use PCA or not. Justify your answer

viii. Now, train a model using SVM and compare it with Logistic Regression model built with PCA from the above step and summarize your findings

ix. Train a Decision Tree, a Random Forest and a Neural Network on the data-set. Tune the hyper-parameters using GridSearchCV(), if required, and evaluate the models on your chosen metrics as per (v) above

x. Provide a list of at least 4 most important and least important features that contribute in predicting churn

xi. Finally,compare all the models (logistic with PCA, logistic without PCA, SVM, Decision Tree, Random Forest and Neural Network) and recommend one final model for deployment. Base your recommendation on your chosen metric, computational resources required, interpret-ability, ease of usage, feature importance, etc.

NOTE: *A good code is readable, well commented and properly structured with all your assumptions defined properly.*