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. Given 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, you 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 - postpaid (customers pay a monthly/annual bill after using the services) and prepaid (customers pay/recharge with a certain amount in advance and then use the services). In the postpaid model, when customers want to switch to another operator, they usually inform the existing operator to terminate the services, and you directly know that this is an instance of churn. However, in the prepaid model, customers who want to switch to another network can simply stop using the services without any notice, 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, and the term 'churn' should be defined carefully. Also, prepaid is the most common model in India and southeast Asia, while postpaid is more common in Europe in North America. This project is based on the Indian and Southeast Asian market.

Definitions of churn

There are various ways to define churn, such as:

Revenue-based churn: Customers who have not utilised 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/SMSes 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 you 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.

In this project, you will use the **usage-based definition** to define churn.

High-value 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.

In this project, you will define high-value customers based on a certain metric (mentioned later below) and predict churn only on high-value customers.

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.

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 behaviour during churn will be helpful.

Understanding customer behaviour 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 lifecycle** :

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

- 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.

Analysis Steps

Data Cleaning and EDA

We have started with importing Necessary packages and libraries. We have loaded the dataset into a dataframe. We have checked the number of columns, their data types, Null count and unique value_value_count to get some understanding about data and to check if the columns are under correct data-type. Checking for duplicate records (rows) in the data. There were no duplicates. Since 'mobile_number' is the unique identifier available, we have made it our index to retain the identity. Have found some columns that do not follow the naming standard, we have renamed those columns to make sure all the variables follow the same naming convention. Following with column renaming, we have dealt with converting the columns into their respective data types. Here, we have evaluated all the columns which are having less than or equal to 29 unique values as categorical columns and rest as continuous columns. The date columns were having 'object' as their data type, we have converted to the proper datetime format. Since, our analysis is focused on the HVC(High value customers), we have filtered for high value customers to carryout the further analysis. The metric of this filtering of HVC is such that all the customers whose 'Average_rech_amt' of months 6 and 7 greater than or equal to 70th percentile of the 'Average_rech_amt' are considered as High Value Customers. Checked for missing values. Dropped all the columns with missing values greater than 50%. We have been given 4 months data. Since each months revenue and usage data is not related to other, we did month-wise drill down on missing values. Some columns had similar range of missing values. So, we have looked at their related columns and checked if these might be imputed with zero. We have found that 'last_date_of_the_month' had some missing values, so this is very meaningful and we have imputed the last date based on the month. We have found some columns with only one unique value, so it is of no use for the analysis, hence we have dropped those columns. Once after checking all the data preparation tasks, tagged the Churn variable(which is our target variable). After imputing, we have dropped churn phase columns (Columns belonging to month - 9). After all the above processing, we have retained 30,011 rows and 126 columns. Exploratory Data Analysis

Analysis Results

The telecom company has many users with negative average revenues in both phases. These users are likely to churn. Most customers prefer the plans of '0' category. The customers with lesser 'aon' are more likely to Churn when compared to the Customers with higer 'aon'. Revenue generated by the Customers who are about to churn is very unstable. The Customers whose arpu decreases in 7th month are more likely to churn when compared to ones with increase in arpu. The Customers with high total_og_mou in 6th month and lower total_og_mou in 7th month are more likely to churn compared to the rest. The Customers

with decrease in rate of total_ic_mou in 7th month are more likely to churn, compared to the rest. Customers with stable usage of 2g volume throughout 6 and 7 months are less likely to churn. Customers with fall in usage of 2g volume in 7th month are more likely to Churn. Customers with stable usage of 3g volume throughout 6 and 7 months are less likely to churn. Customers with fall in consumption of 3g volume in 7th month are more likely to Churn. The customers with lower total_og_mou in 6th and 8th months are more likely to Churn compared to the ones with higher total_og_mou. The customers with lesser total_og_mou_8 and aon are more likely to churn compared to the one with higher total_og_mou_8 and aon. The customers with less total_ic_mou_8 are more likely to churn irrespective of aon. The customers with total_ic_mou_8 > 2000 are very less likely to churn.

Important Steps carried out

Correlation analysis has been performed. We have created the derived variables and then removed the variables that were used to derive new ones. Outlier treatment has been performed. We have looked at the quantiles to understand the spread of Data. We have capped the upper outliers to 99th percentile. We have checked categorical variables and contribution of classes in those variables. The classes with less ccontribution are grouped into 'Others'. Dummy Variables were created.

Pre-processing Steps

Train-Test Split has been performed. The data has high class-imbalance with the ratio of 0.095 (class 1 : class 0). SMOTE technique has been used to overcome class-imbalance. Predictor columns have been standardized to mean - 0 and standard_deviation- 1.

Modelling

Model 1 : Logistic Regression with RFE & Manual Elimination (Interpretable Model) Most important predictors of Churn , in order of importance and their coefficients are as follows :

loc_ic_t2f_mou_8 -1.2736 total_rech_num_8 -1.2033 total_rech_num_6 0.6053 monthly_3g_8_0 0.3994 monthly_2g_8_0 0.3666 std_ic_t2f_mou_8 -0.3363 std_og_t2f_mou_8 -0.2474 const -0.2336 monthly_3g_7_0 -0.2099 std_ic_t2f_mou_7 0.1532 sachet_2g_6_0 -0.1108 sachet_2g_7_0 -0.0987 sachet_2g_8_0 0.0488 sachet_3g_6_0 -0.0399 PCA: PCA : 95% of variance in the train set can be explained by first 16 principal components and 100% of variance is explained by the first 45 principal components.

Model 2 : PCA + Logistic Regression

```
Train Performance :


Accuracy : 0.627

Sensitivity / True Positive Rate / Recall : 0.918

Specificity / True Negative Rate :  0.599

Precision / Positive Predictive Value : 0.179

F1-score : 0.3


Test Performance :


Accuracy : 0.086

Sensitivity / True Positive Rate / Recall : 1.0
```

Specificity / True Negative Rate : 0.0

Precision / Positive Predictive Value : 0.086

F1-score : 0.158

Model 3 : PCA + Random Forest Classifier

Train Performance :

Accuracy : 0.882

Sensitivity / True Positive Rate / Recall : 0.816

Specificity / True Negative Rate : 0.888

Precision / Positive Predictive Value : 0.408

F1-score : 0.544

Test Performance :

Accuracy : 0.86

Sensitivity / True Positive Rate / Recall : 0.80

Specificity / True Negative Rate : 0.78

Precision / Positive Predictive Value :0.37

F1-score :0.51

Model 4 : PCA + XGBoost

Train Performance :

Accuracy : 0.873

Sensitivity / True Positive Rate / Recall : 0.887

Specificity / True Negative Rate : 0.872

Precision / Positive Predictive Value : 0.396

F1-score : 0.548

Test Performance :

Accuracy : 0.086

Sensitivity / True Positive Rate / Recall : 1.0

Specificity / True Negative Rate : 0.0

```
Precision / Positive Predictive Value : 0.086

F1-score : 0.158
```

Recommendations :

Following are the strongest indicators of churn

Customers who churn show lower average monthly local incoming calls from fixed line in the action period by 1.27 standard deviations , compared to users who don't churn , when all other factors are held constant. This is the strongest indicator of churn. Customers who churn show lower number of recharges done in action period by 1.20 standard deviations, when all other factors are held constant. This is the second strongest indicator of churn. Further customers who churn have done 0.6 standard deviations higher recharge than non-churn customers. This factor when coupled with above factors is a good indicator of churn. Customers who churn are more likely to be users of 'monthly 2g package-0 / monthly 3g package-0' in action period (approximately 0.3 std deviations higher than other packages), when all other factors are held constant.

Based on the above indicators the recommendations to the telecom company are :

Concentrate on users with 1.27 std devations lower than average incoming calls from fixed line. They are most likely to churn. Concentrate on users who recharge less number of times (less than 1.2 std deviations compared to avg) in the 8th month. They are second most likely to churn. Models with high sensitivity are the best for predicting churn. Use the PCA + Logistic Regression model to predict churn. It has an ROC score of 0.87, test sensitivity of 100%.