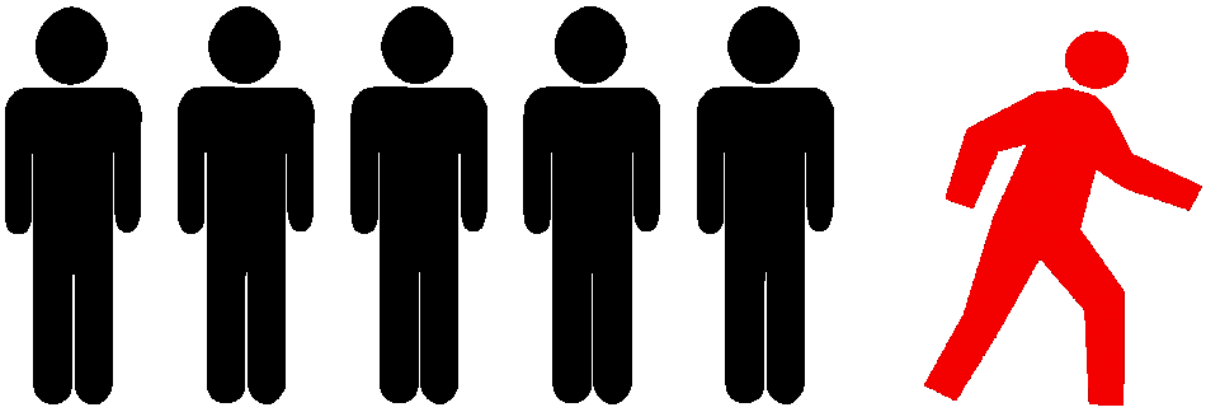

Report on Telecom Customer Churn

WRITTEN BY: AJINKYA ASHOK JADHAV

Problem Statement:

In the telecom industry, customers are able to choose from multiple service providers and actively switch from one telecom operator to another. In this highly competitive market, the telecom industry experiences on average of 15-20% annual churn rate. Given the fact that it costs 5-10 times more resources to acquire a new customer than to retain an existing one, so customer retention has become even more important than customer acquisition.

For many incumbent operators, **retaining high profitable customers is the number one business goal**. So to reduce those customer churn, **companies need to predict which customers are at high risk of churn beforehand**.

Objective:

- To analyze telecom user data of leading telecom firm
- To build predictive models to identify customer at high risk of churn and identify main indicators of churn.

Data Understanding:

- Dataset contains almost 1 lac observations with 226 feature columns.
- Most of them are float datatype, some of them are integer and object datatype.
- 4 months of data is given for each customer; minutes of usage for all type services as well as number of recharge per month and average revenue per user per month, age on network , data consumption of user etc. is given.
- There is no any duplicate entries present; means all the users are unique.
- Missing values are present in the some feature columns; 30 feature columns have around 75% missing, rest of them have missing percentage in between 3-5% and rest do not have missing values.
- I checked for any relationship between columns to have missing values and treat them accordingly.

- After doing it still some columns have high percentage of missing values; which simply can drop.
- Then for small percentage of missing values iterative imputer is used to fill those columns.

Since all the customers are not revenue generating entities so we need to focus on only those which are using services of the company which is generating revenue. So we need to pick those high value customers (HVCs). I have used usage based churn approach to identify such customers.

HVCs are 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 i.e. 6th and 7th months (the good phase). Total of data and calling recharge for these months. And take the 70% percentile value to tagging customer as HVC.

Target Variable:

To tagged the customer as whether churn or not; I use 9 months which last month in the dataset total incoming and outgoing minutes of usage and data volume of 2G or 3G consumption. Churn customers will be those who would have zero usage of above services for this month's else not churn.

Then all columns containing information of 9 months are dropped.

Feature Engineering:

Addition of roaming incoming and outgoing with respective total incoming and outgoing minutes of usage for respective months. And Change in average revenue per user between consecutive months are derived from the given feature columns.

Some of the feature columns are drop to reduce multi-collinearity.

Exploratory data Analysis:

EDA is done to see the nature of some feature columns. Boxplot is used to see data distribution over different quartiles and if there is outliers present then capping is done.

In bivariate analysis; I observe that some derived features are helping to predict target variable. Such as change in total recharge amount for 6 and 7 months shows that the median value for the customer who has churned have negative value that would mean services were not satisfactory for him/her. Similar observation were obtained from bivariate analysis for other columns also.

Total outgoing and incoming minutes of usage also separate churned user from not churn user.

Model Building:

Interpretable Model:

To make interpretable model I build decision tree classifier. Because we can interpret it easily.

Observation for this model:

- If roaming outgoing minutes of usage for 8 month is greater than 0.005 then most of the customer will churn
- Else if it is less than or equal to 0.005 then most of the customer will not churn.

As we know to get interpretable model we have to compromise with model evaluation metrics. But if we want more accurate model we need to give up on the interpretability of the model. So to build more complex model I decide to ignore about interpretability.

More Complex model:

Since it is classification model, and our prime concern is to predict who will get churn in future beforehand means we need to build model whose false negative; means a customer who would be churned but our model could not predict them accurately as churned, has to be minimum that means model's recall must be higher. But while trying to improve recall of the model precision of the model goes down. There is trade-off between those two metrics. So while building model I tried to maintain both this metric at optimum level.

Out of model given in the notebook RandomForest with Neighbourhood cleaning rule is the optimum model with recall 0.75 on test data.