Telecom Churn Case Study

Problem Description:

- 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 become even more important than customer acquisition.
- Here we are given with 4 months of data related to customer usage. In this case study, we analyze 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.
- Churn is predicted using two approaches. Usage based churn and Revenue based churn. Usage based churn: Customers who have zero usage, either incoming or outgoing - in terms of calls, internet etc. over a period of time.
- This case study only considers usage based 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.
- 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**.
- This is a classification problem, where we need to predict whether the customer is about to churn or not. We have carried out Baseline Logistic Regression, then Logistic Regression with PCA, PCA + Random Forest, PCA + XGBoost.

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 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 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.)
- 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.
- The attributes containing 6, 7, 8, 9 as suffixes imply that those correspond to the months 6, 7, 8, 9 respectively

Analysis Steps

Data Cleaning and EDA

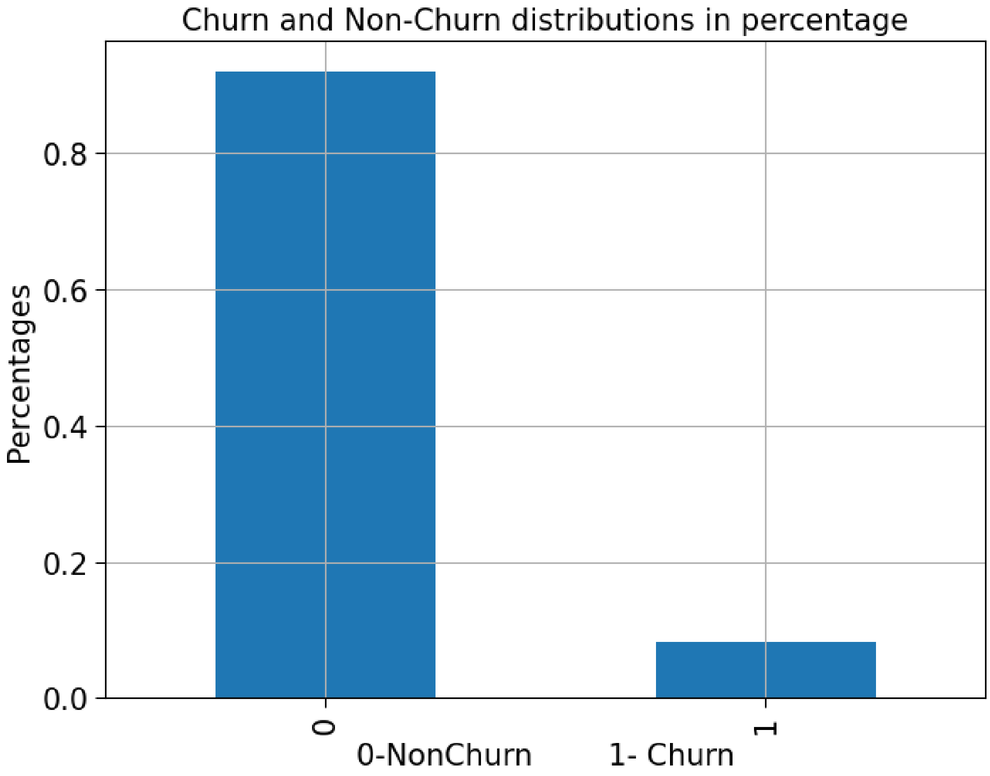
1. We have started with importing Necessary packages and libraries.
2. We have loaded the dataset into a dataframe.
3. We have checked the number of columns, their data types, Null count and unique value_value_count to get some understanding about data and to check if the columns are under correct data-type.
4. Checking for duplicate records (rows) in the data. There were no duplicates.
5. Since 'mobile_number' is the unique identifier available, we have made it our index to retain the identity.
6. 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.
7. 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.
8. The date columns were having 'object' as their data type, we have converted to the proper datetime format.
9. Since, our analysis is focused on the HVC (High value customers), we have filtered for high value customers to carry out 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.
10. Checked for missing values.
11. Dropped all the columns with missing values greater than 50%. We have removed 30 columns from the dataframe
12. 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.

13. 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.
14. We have found that 'last_date_of_the_month' had some misisng values, so this is very meaningful and we have imputed the last date based on the month.
15. 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.
16. Once after checking all the data preparation tasks, tagged the Churn variable (which is our target variable).
17. After imputing, we have dropped churn phase columns (Columns belonging to month - 9).
18. After all the above processing, we have retained 30,011 rows and 126 columns.

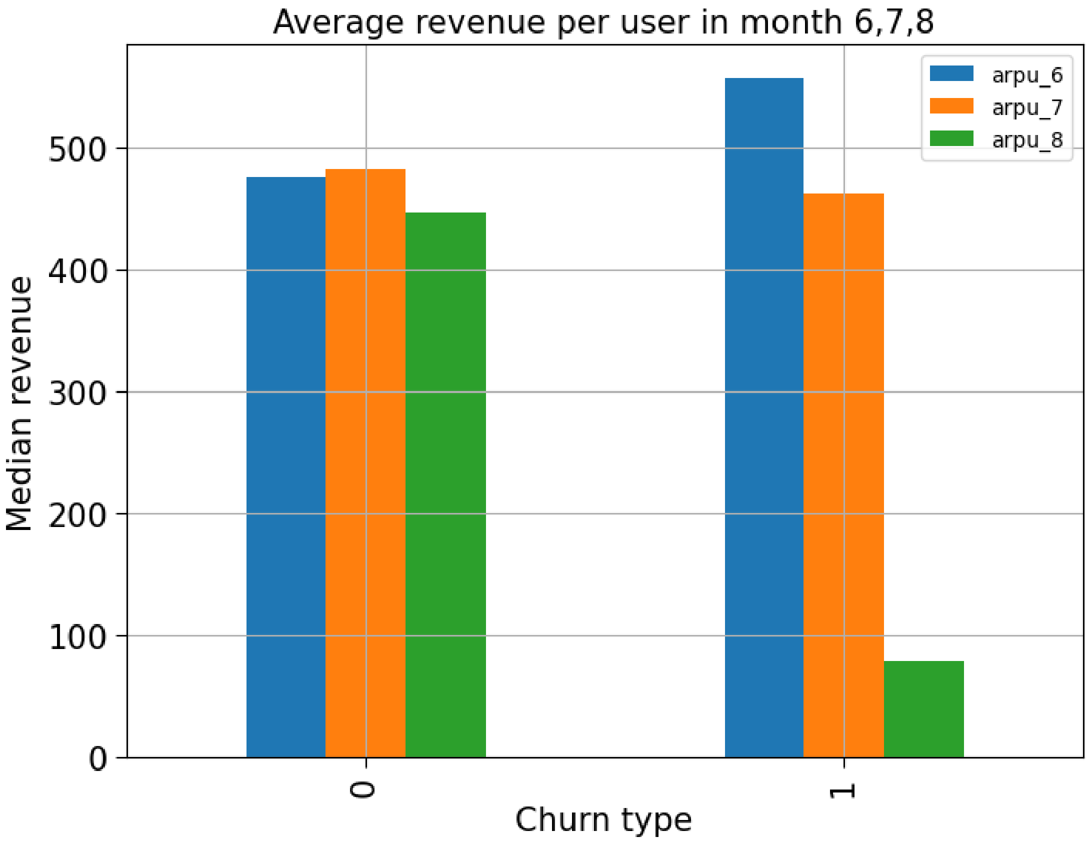
Exploratory Data Analysis

- 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.

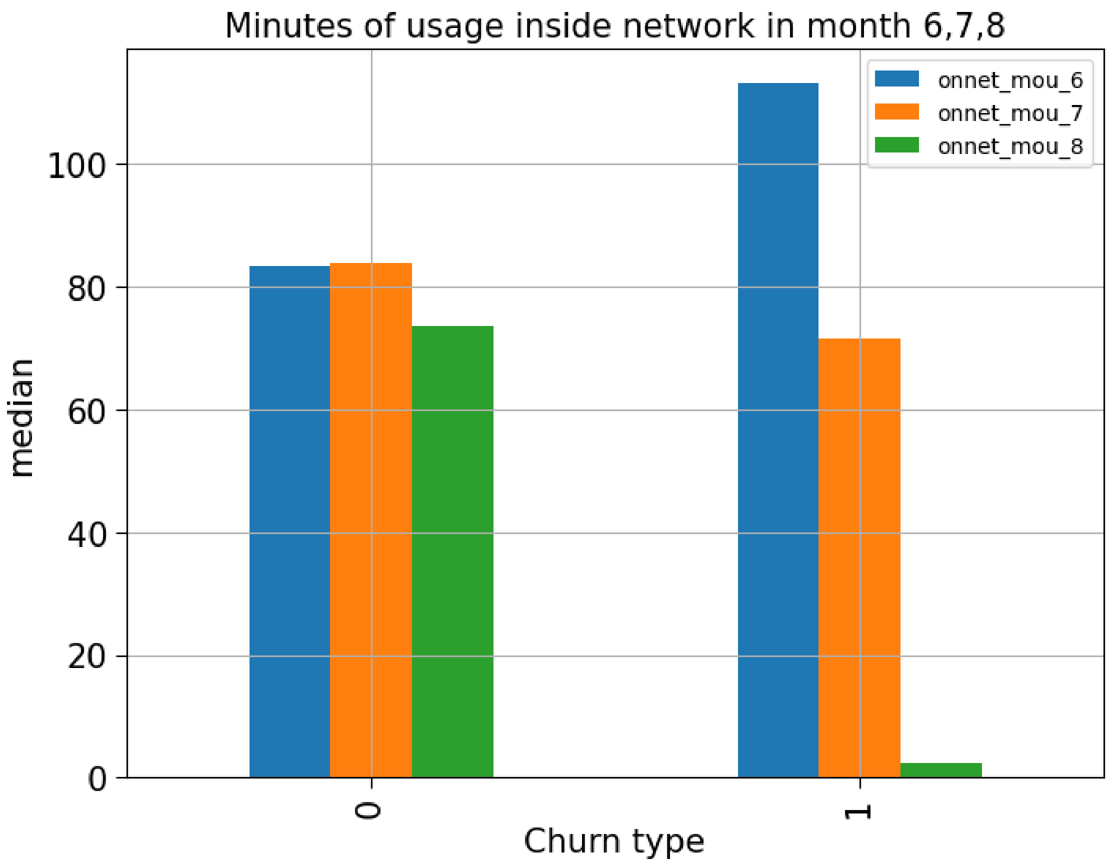
- The customers with total_ic_mou_8 > 2000 are very less likely to churn.



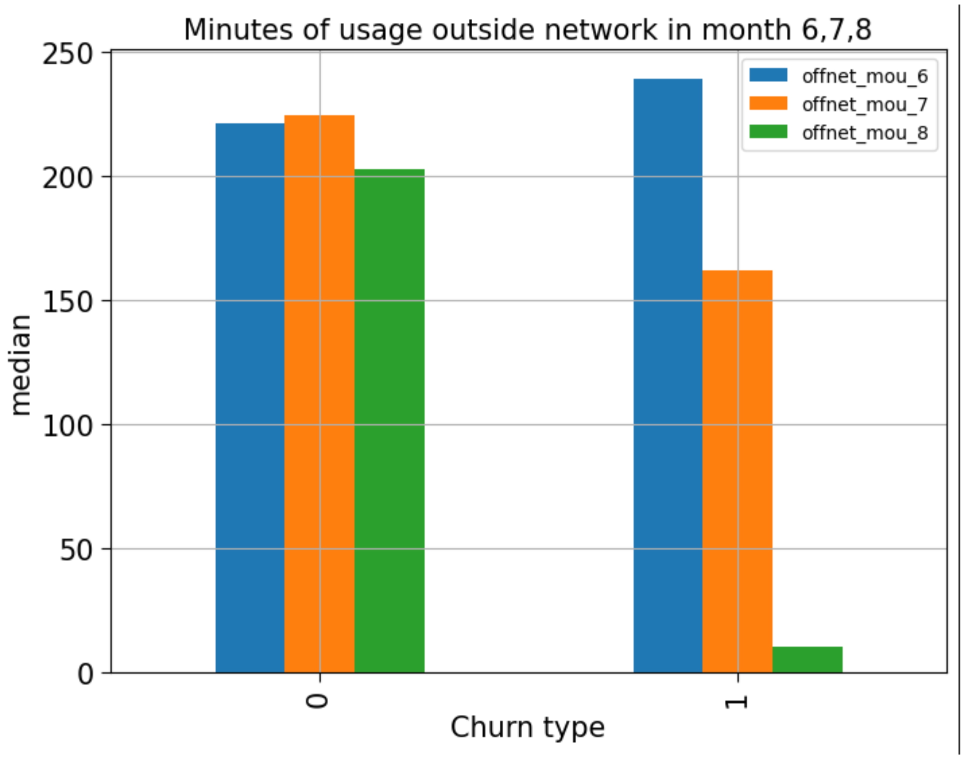
- **We have 92% customers belong non-churn and 8% customers belong to Churn type.**



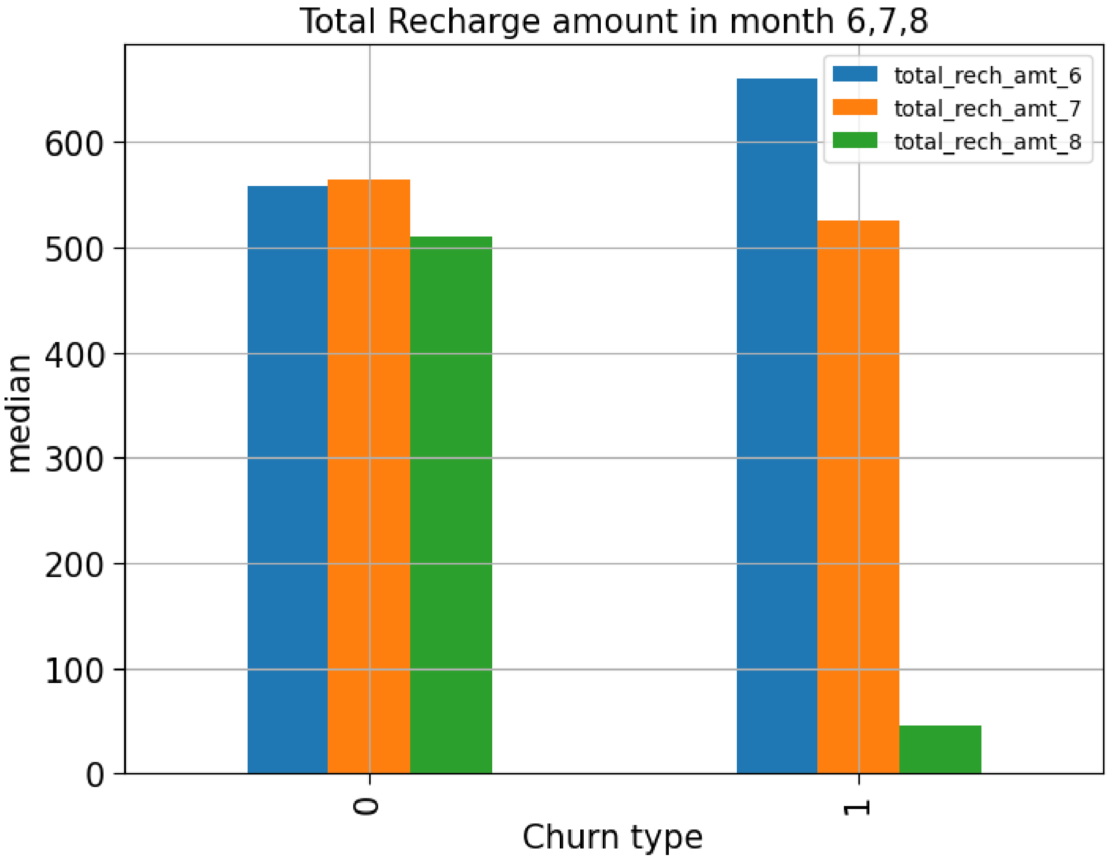
- Average revenue per user more in month 6 means, if they are unsatisfied, those useres are more likely to churn



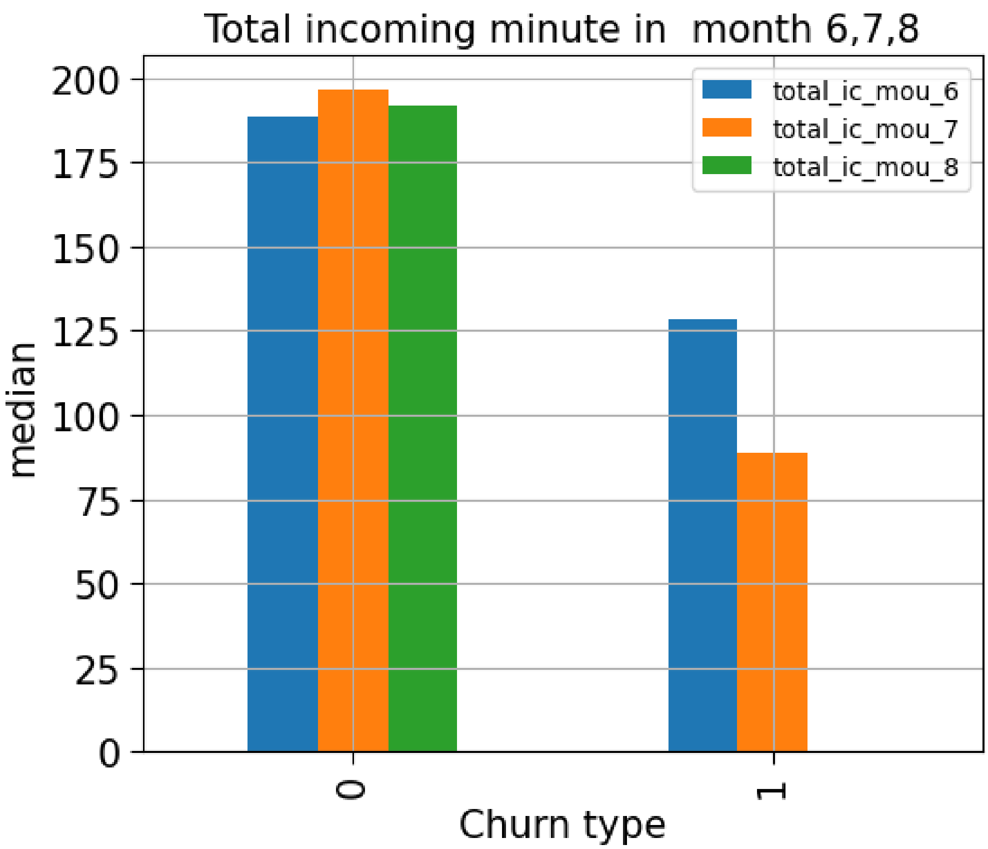
- **Users whose minutes of usage are more in month 6, they are more likely to churn.**



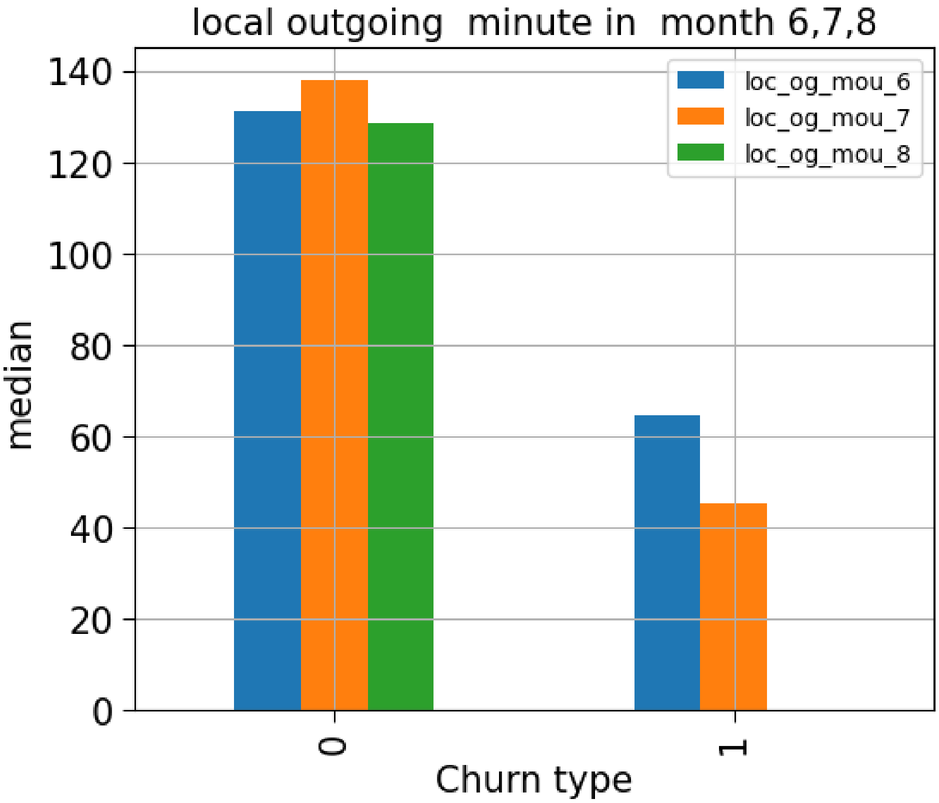
- **The users who have big difference of minutes of call duration to other network between month 6 and month 7,are likely to churn.**



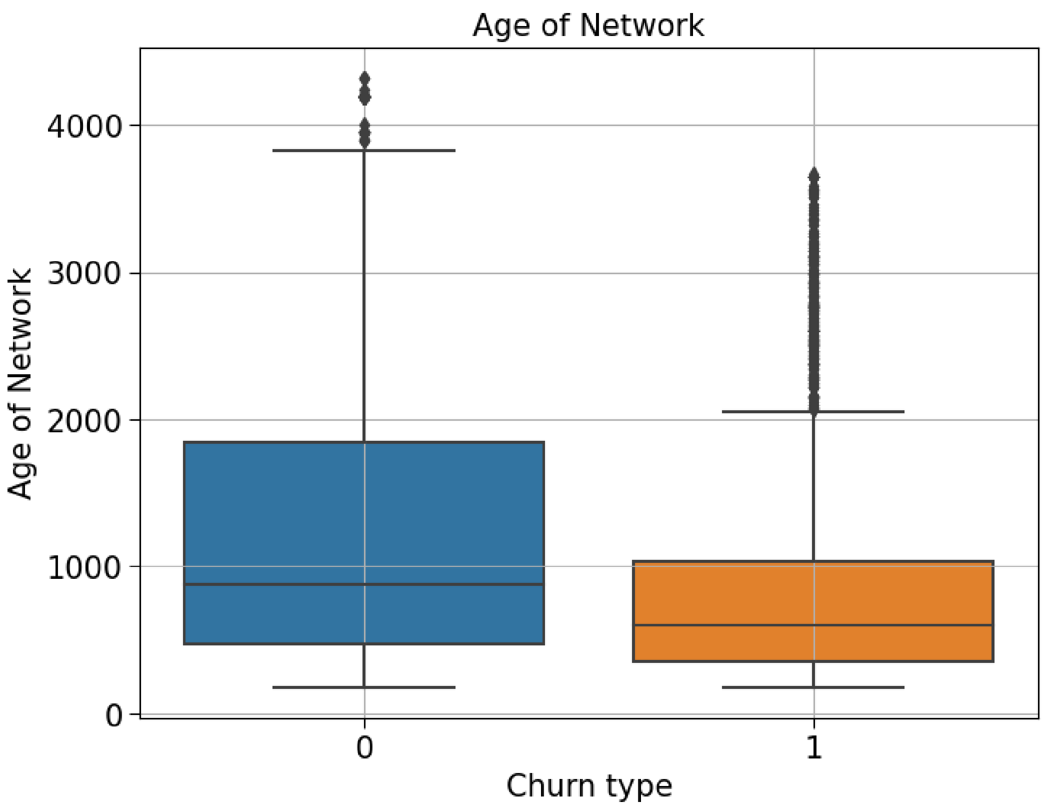
- **when the difference of total recharge amount is more, those users are more likely to churn.**



- Users who have more difference in Total incoming minutes in month 6,7,8 are more likely to churn



- local outgoing minute are less, users are more likely to churn



- Median Age of network less,more likely to churn

Key Steps -

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

Pre-processing Steps

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

Modelling

Model 1 -

Logistic Regression: Test Data

precision: 0.09159005788219271

recall: 0.558091286307054

f1_score: 0.15735595203275812

roc_auc: 0.576069917651508

Model 2 –

Decision Tree: Test Data

precision: 0.33690360272638753

recall: 0.7178423236514523

f1_score: 0.4585818422796554

roc_auc: 0.8510624180969703

Model 3 –

Random Forest: Test Data

precision: 0.5862708719851577

recall: 0.6556016597510373

f1_score: 0.6190009794319294

roc_auc: 0.9244024227132374

Model 4 –

GradientBoosting: Test data

precision: 0.48326055312954874

recall: 0.6887966804979253

f1_score: 0.5680068434559452

roc_auc: 0.9197948016621568

Model 5 –

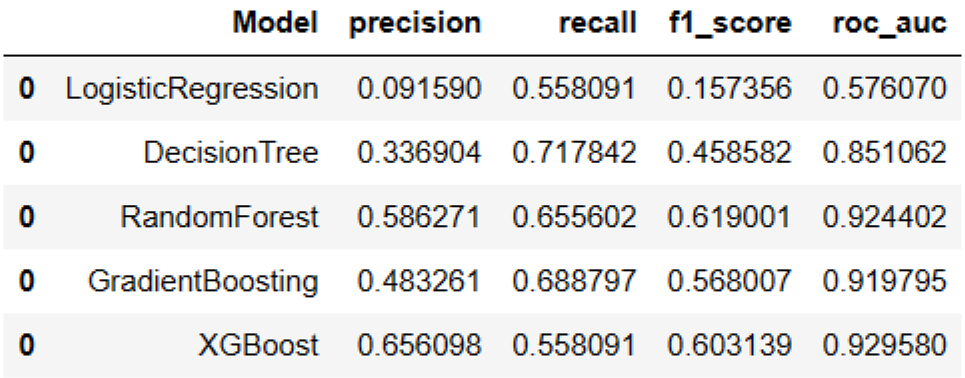
XGBoost: Test Data

precision: 0.6560975609756098

recall: 0.558091286307054

f1_score: 0.6031390134529148

roc_auc: 0.9295804986019627



	Model	precision	recall	f1_score	roc_auc
0	LogisticRegression	0.091590	0.558091	0.157356	0.576070
0	DecisionTree	0.336904	0.717842	0.458582	0.851062
0	RandomForest	0.586271	0.655602	0.619001	0.924402
0	GradientBoosting	0.483261	0.688797	0.568007	0.919795
0	XGBoost	0.656098	0.558091	0.603139	0.929580

- **The randomforest worked well on this data in churn with precision close to 59%, recall close to 65% and f1_score close to 61%.**
- In Logistic regression we have used PCA.
- In this scenario, Without PCA model works well.

Feature importance –

columns	feature_importance	
80	total_ic_mou_8	0.066874
95	total_rech_amt_8	0.043899
134	fb_user_8	0.038620
65	loc_ic_mou_8	0.037469
119	night_pck_user_8	0.037256
11	roam_ic_mou_8	0.035839
59	loc_ic_t2m_mou_8	0.033583
29	loc_og_mou_8	0.027910
2	arpu_8	0.027545
156	total_rech_amt_diff	0.026513
14	roam_og_mou_8	0.024042
104	total_rech_data_8	0.021991
144	roam_og_mou_diff	0.018811
143	roam_ic_mou_diff	0.018789
20	loc_og_t2m_mou_8	0.015012
98	max_rech_amt_8	0.013761
150	loc_ic_mou_diff	0.013625

columns	feature_importance	
53	total_og_mou_8	0.013502
154	total_ic_mou_diff	0.013340
110	av_rech_amt_data_8	0.013241
56	loc_ic_t2t_mou_8	0.012584
107	max_rech_data_8	0.012510
101	last_day_rch_amt_8	0.012273
155	total_rech_num_diff	0.011773
140	arpu_diff	0.011626
149	total_og_mou_diff	0.010407
141	onnet_mou_diff	0.008210
17	loc_og_t2t_mou_8	0.007867
145	loc_og_mou_diff	0.007731
157	max_rech_amt_diff	0.007625
77	std_ic_mou_8	0.006873
8	offnet_mou_8	0.006631
146	std_og_mou_diff	0.006177
92	total_rech_num_8	0.005596

columns	feature_importance	
46	spl_og_mou_7	0.005352
158	total_rech_data_diff	0.005312
113	vol_2g_mb_8	0.005133
153	spl_ic_mou_diff	0.005132
118	night_pck_user_7	0.005101
135	aon	0.00496

Recommendations:

Customers who churn show lower average monthly local incoming calls from fixed line in the action period by 1.27 standard deviation, 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.
- Use the Random Forest model to predict churn. It has an ROC score of 0.92.
- Users who are using more Roaming in Outgoing and Incoming calls, are likely to churn. Company can focus on them too.