Telecom Churn Prediction

Case Study by -
Sayali Sanjay Deshpande

Business Problem Overview

- 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 analyse customer - level data of a leading telecom firm, build predictive models to identify customers at high risk of churn and the main indicators of churn.

Business Problem Overview

- 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.

Business Problem Overview

- 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.

Business Objective

To predict the churn in the last (i.e. the ninth) month using the data (features) from the first three months.

Analysis Approach

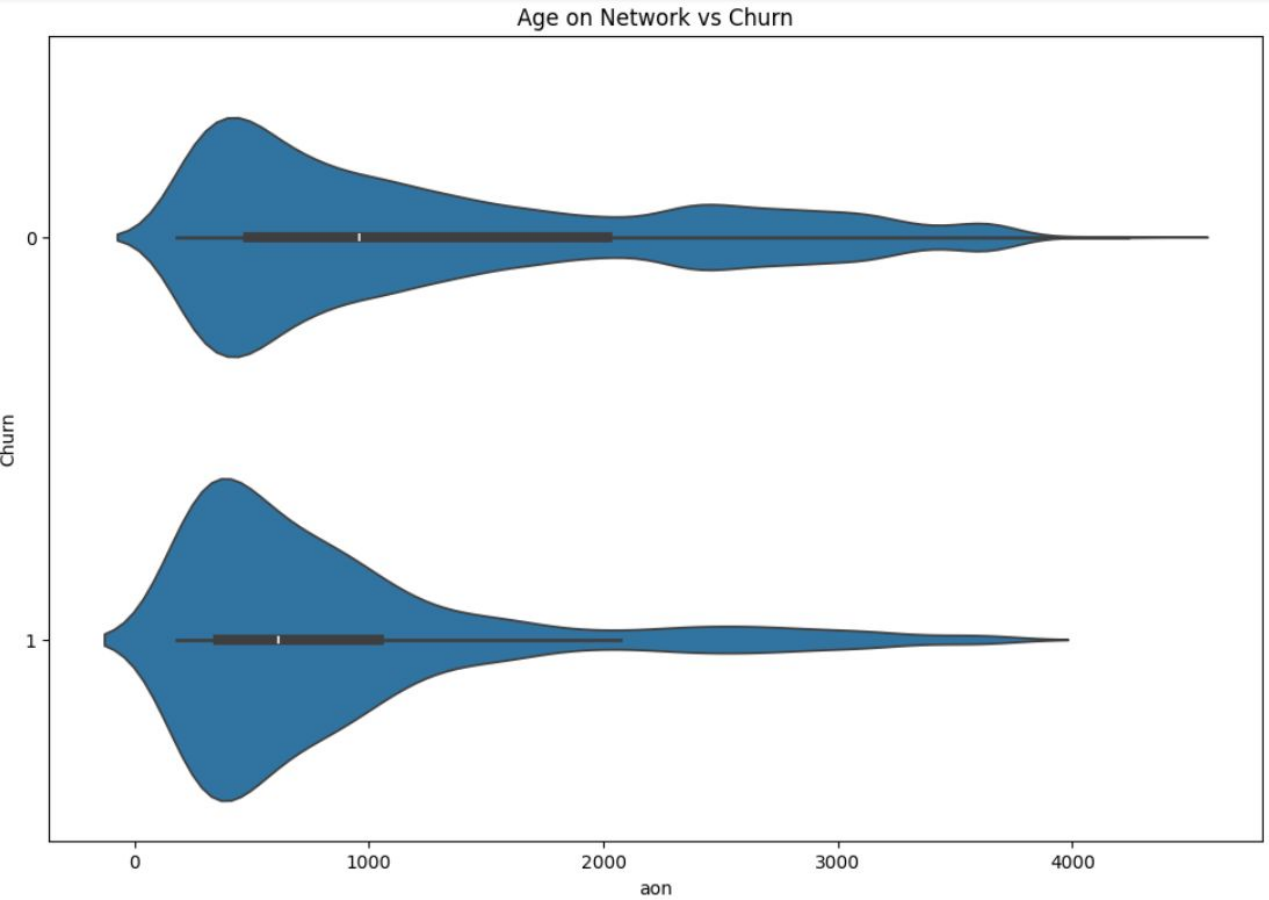
This is a classification problem, where we need to predict whether the customers is about to churn or not.

We have carried out Baseline Logistic Regression, then Logistic Regression with PCA, PCA + Random Forest, PCA + XGBoost.

Steps of the Analysis

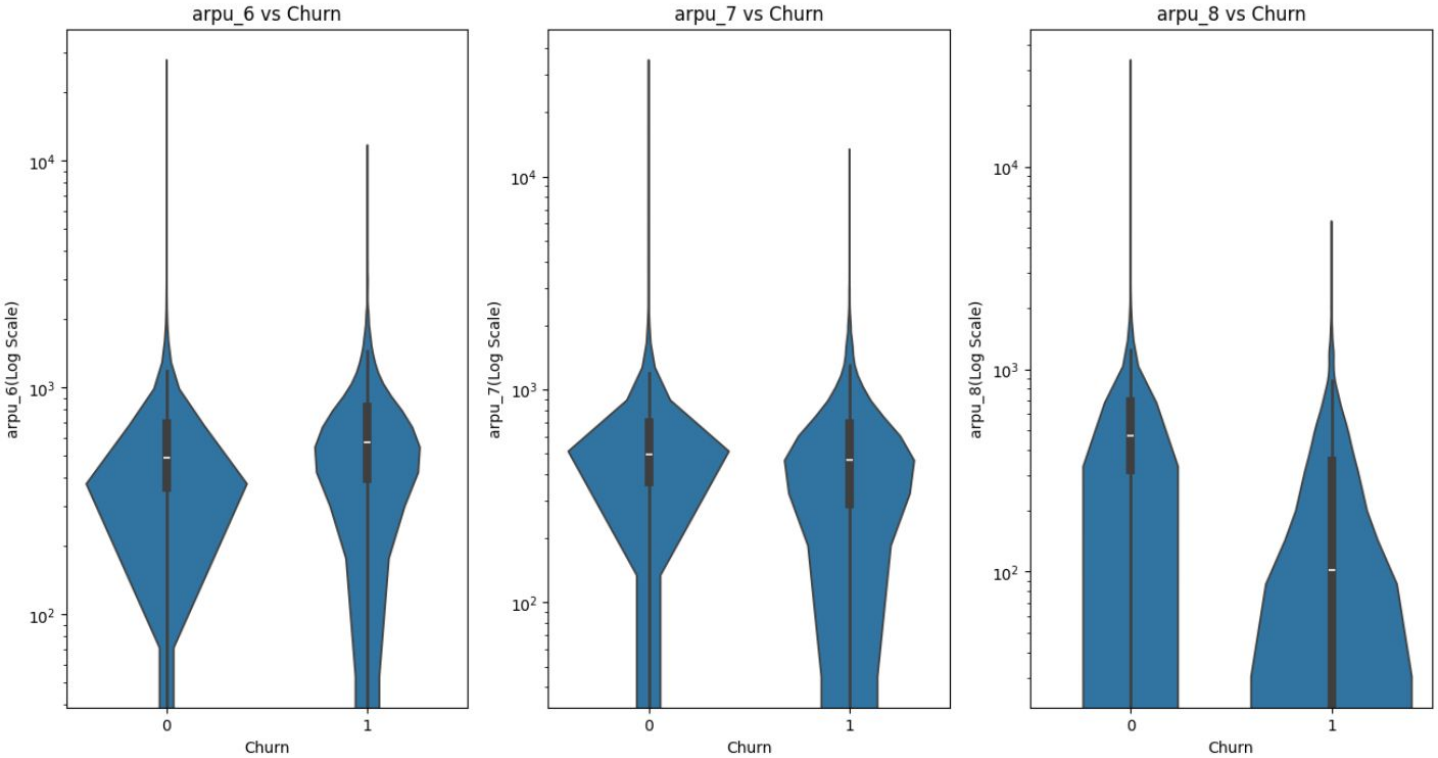
- Understanding the data
- Cleaning the data
- Identifying high value customers
  (Customers are High Valued if their average recharge amount for June and July is more than or equal to 70th percentile of average recharge amount)
- Treating the missing values
- Analysing the data

EDA - Univariate Analysis



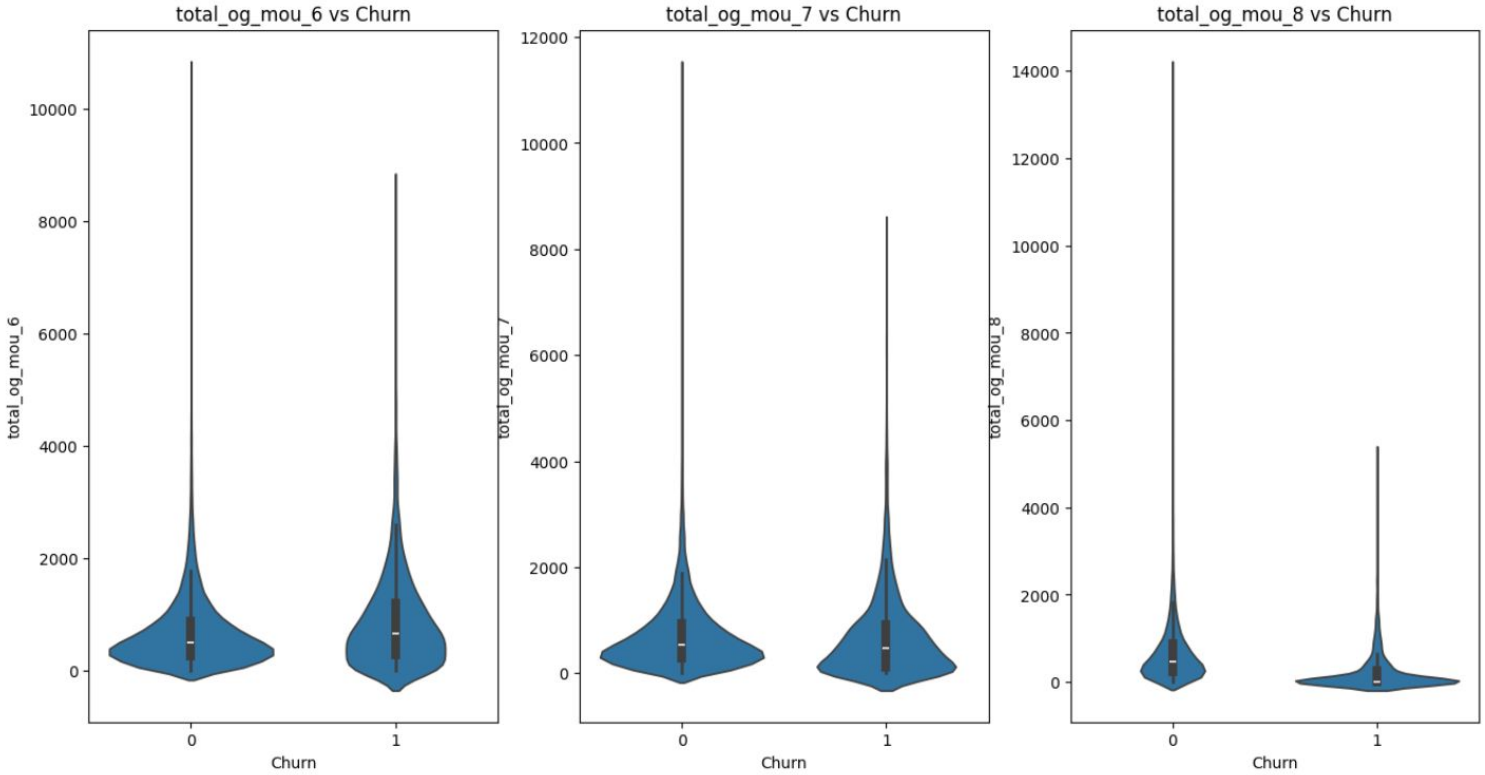
Customers with lesser 'aon' are more likely to churn as compared to the customers with higher 'aon'.

EDA - Univariate Analysis



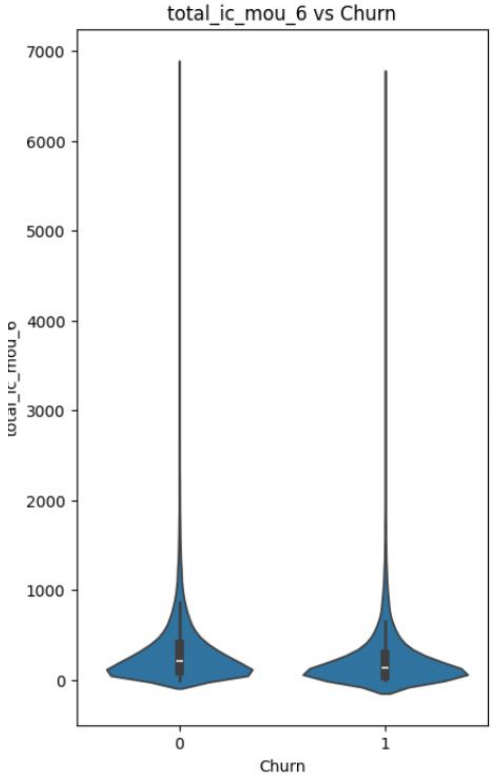
The revenue generated by the customers who are about to churn is very unstable.

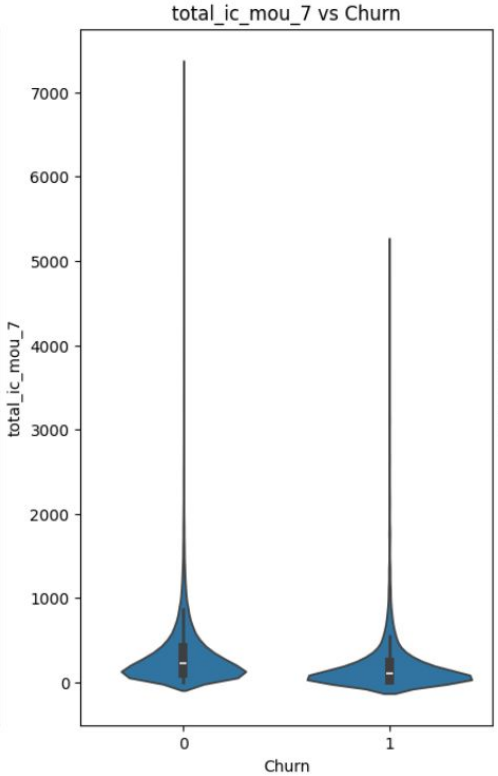
EDA - Univariate Analysis

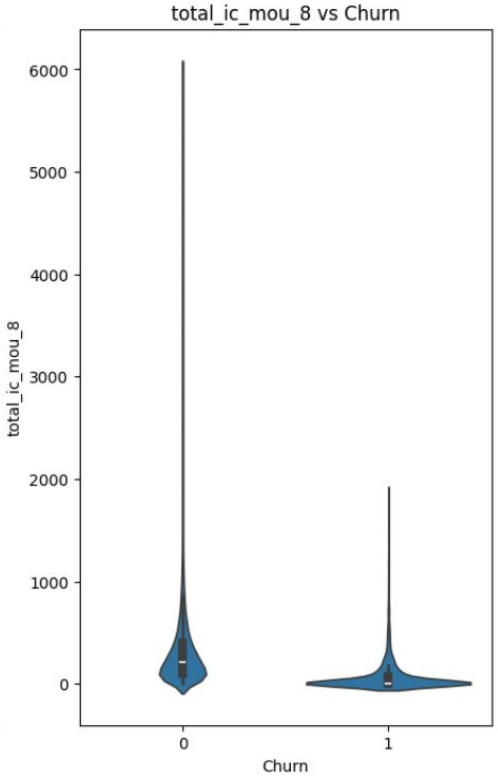


Customers with high total_og_mou in 6th month and lower total_og_mou in 7th month are more likely to churn as compared to the rest.

EDA - Univariate Analysis

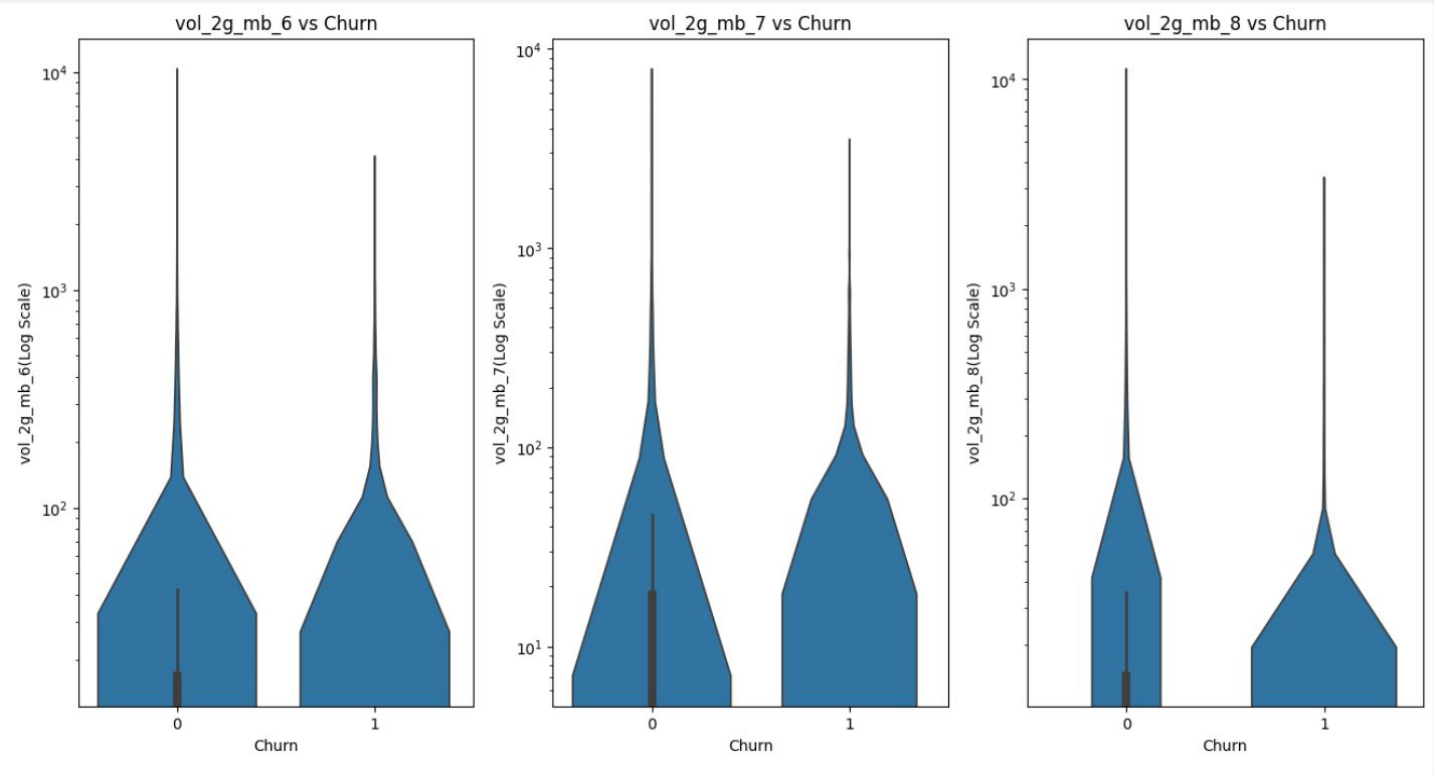






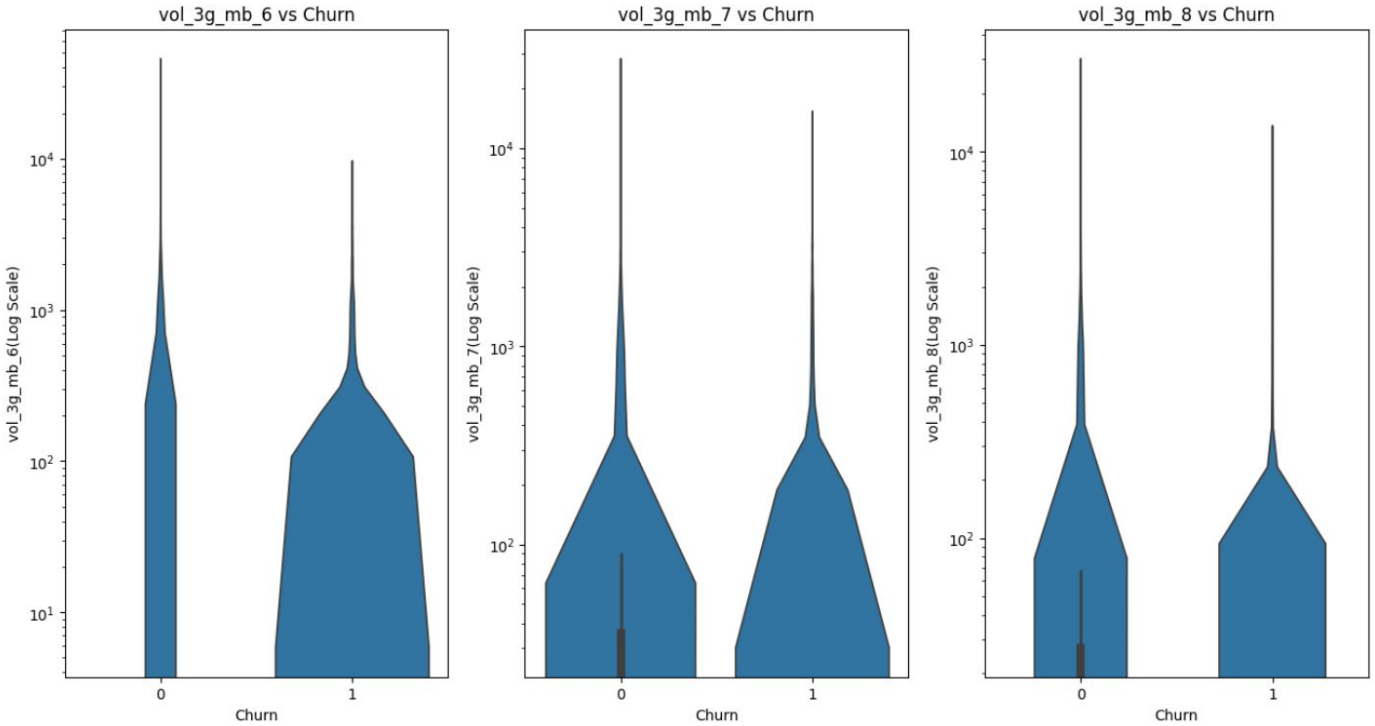
Customers with decrease in rate of total_ic_mou in 7th month are more likely to churn as compared to the rest.

EDA - Univariate Analysis



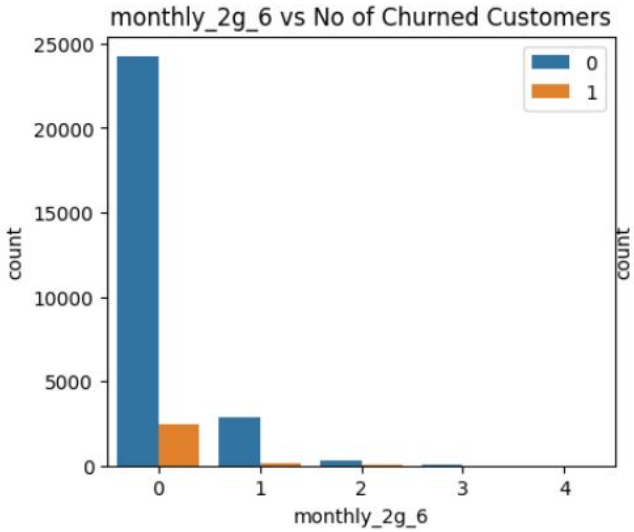
Customers with fall in consumption of 2g volumes in 7th month are more likely to churn.

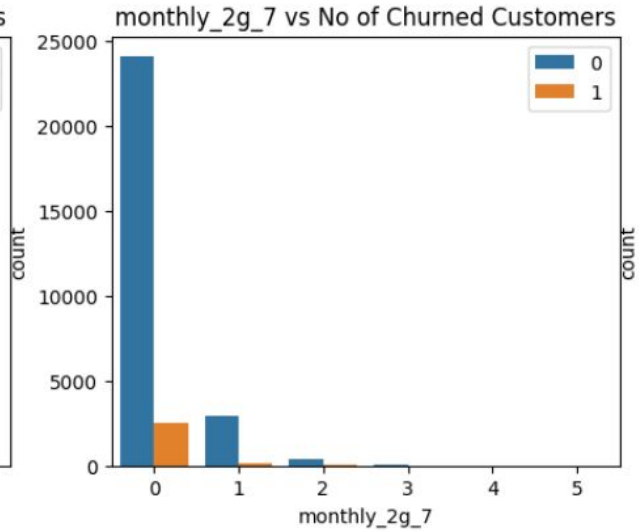
EDA - Univariate Analysis

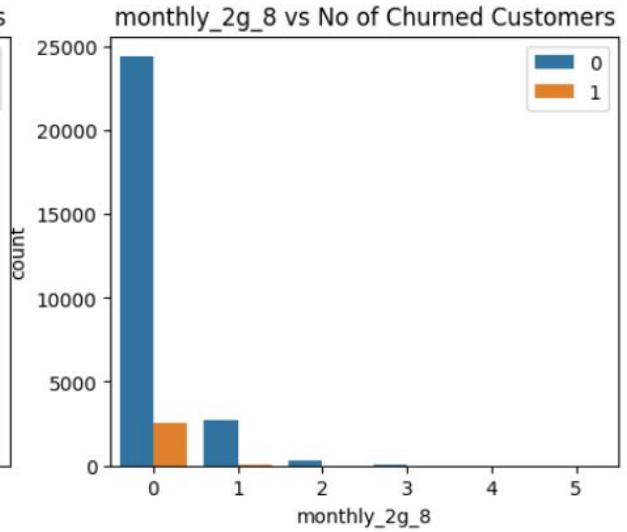


Customers with fall in consumption of 3g volumes in 7th month are more likely to churn.

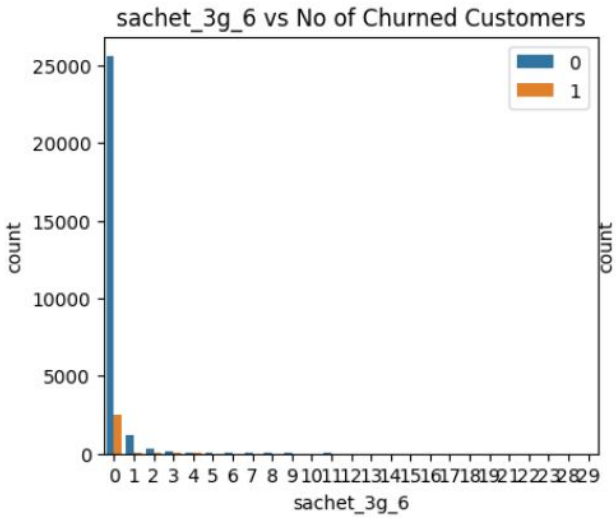
EDA - Univariate Analysis

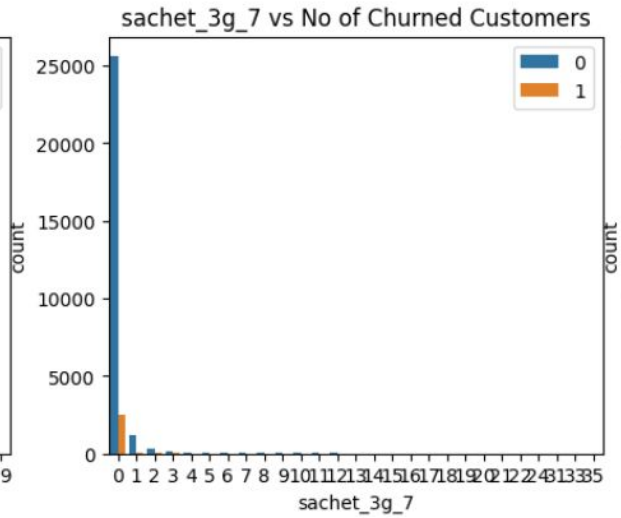


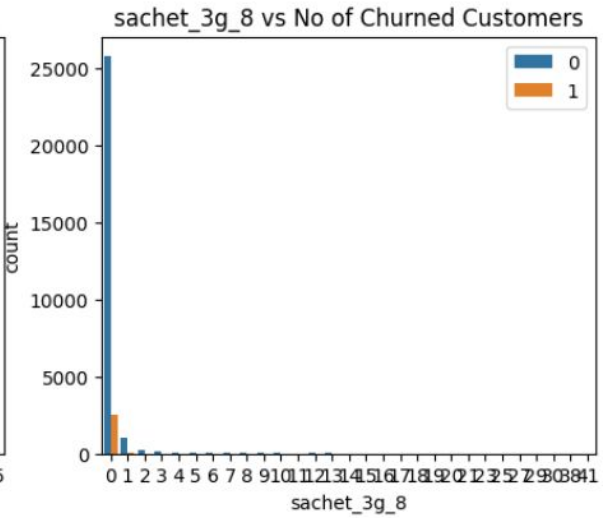




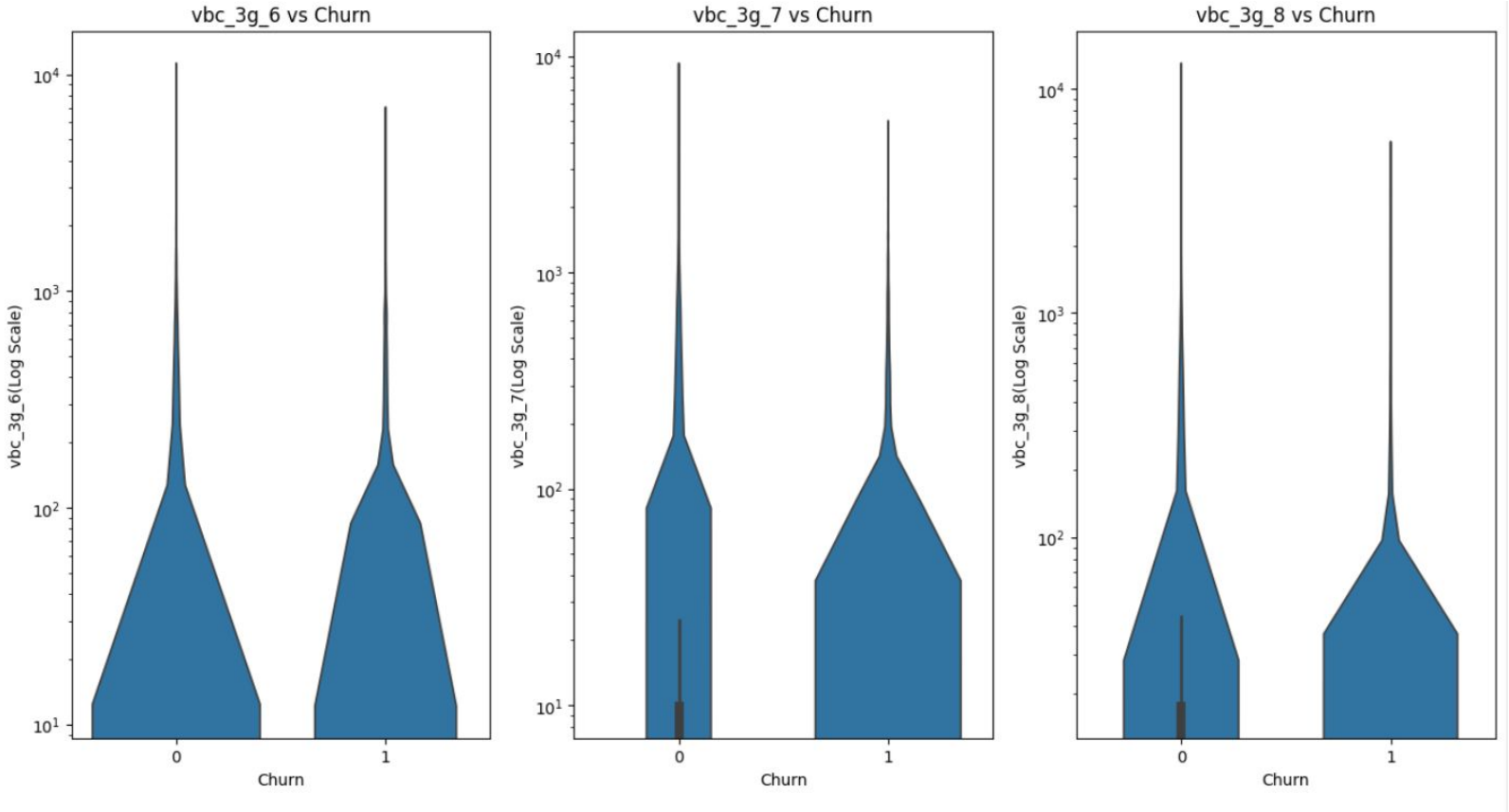
EDA - Univariate Analysis



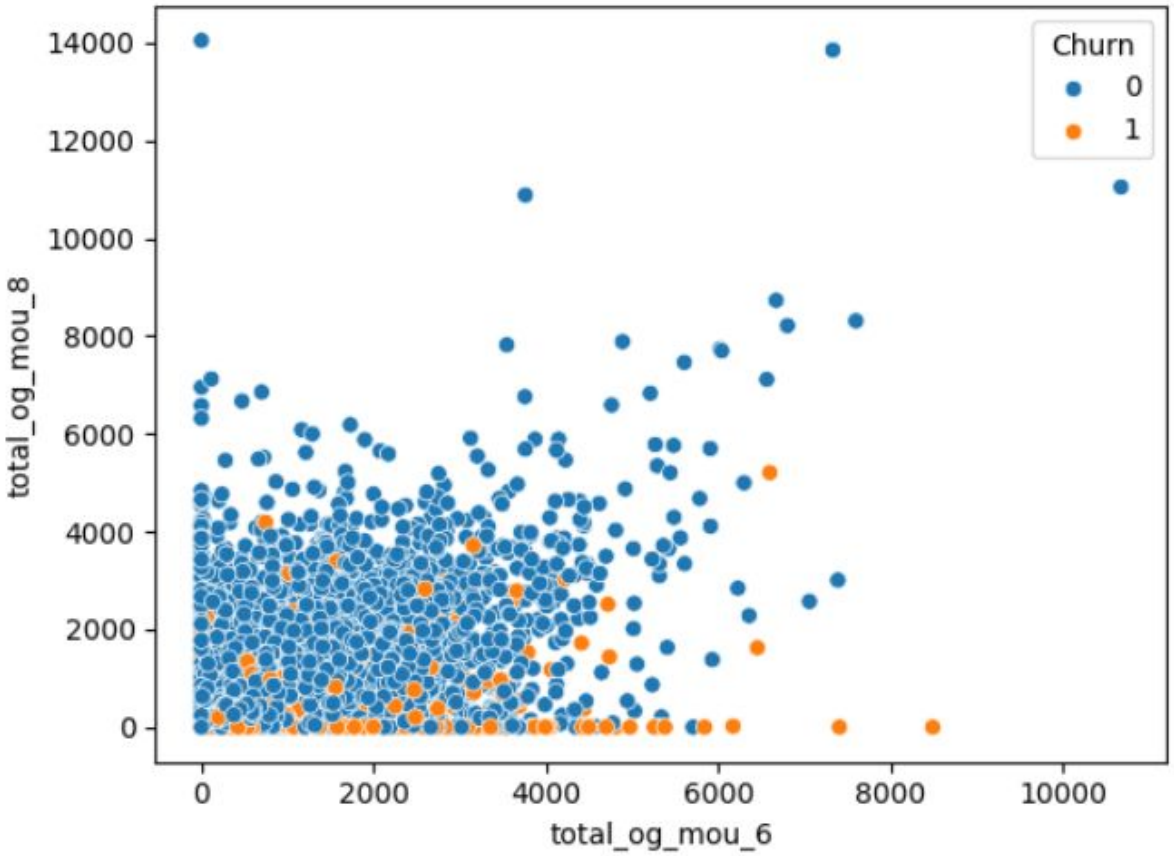




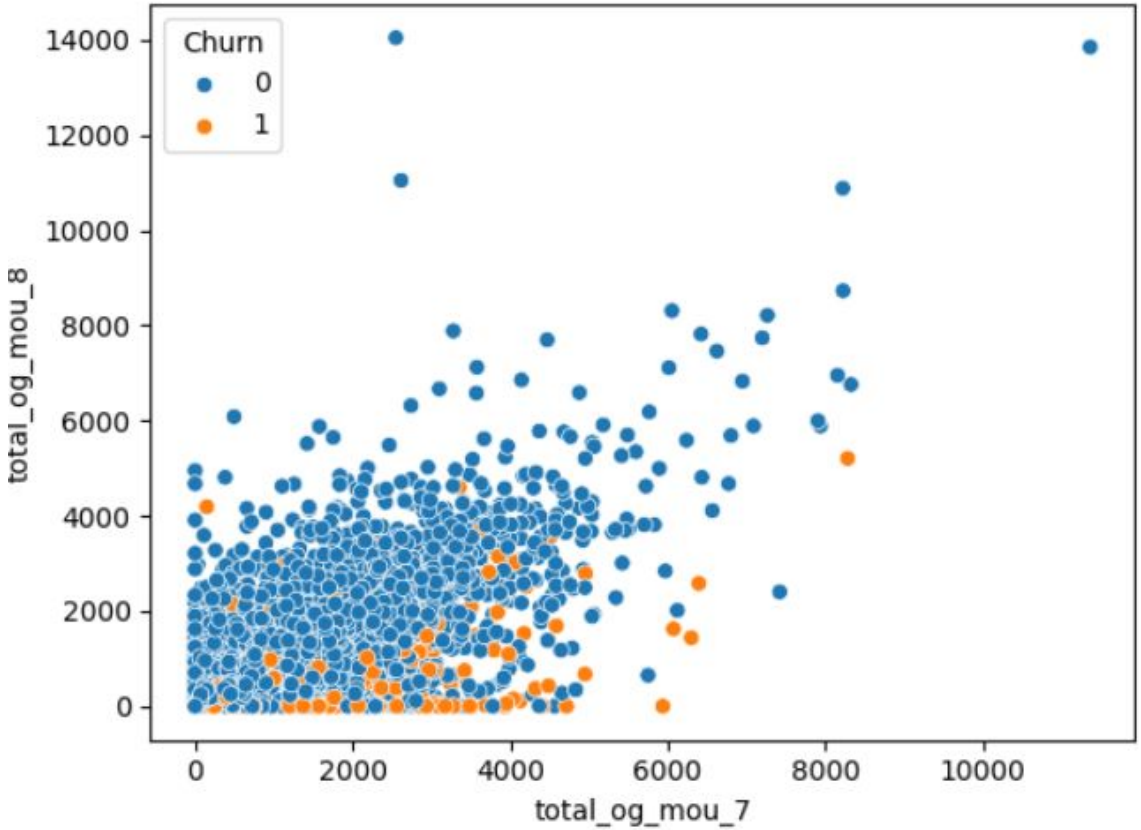
EDA - Univariate Analysis



EDA - Bivariate Analysis

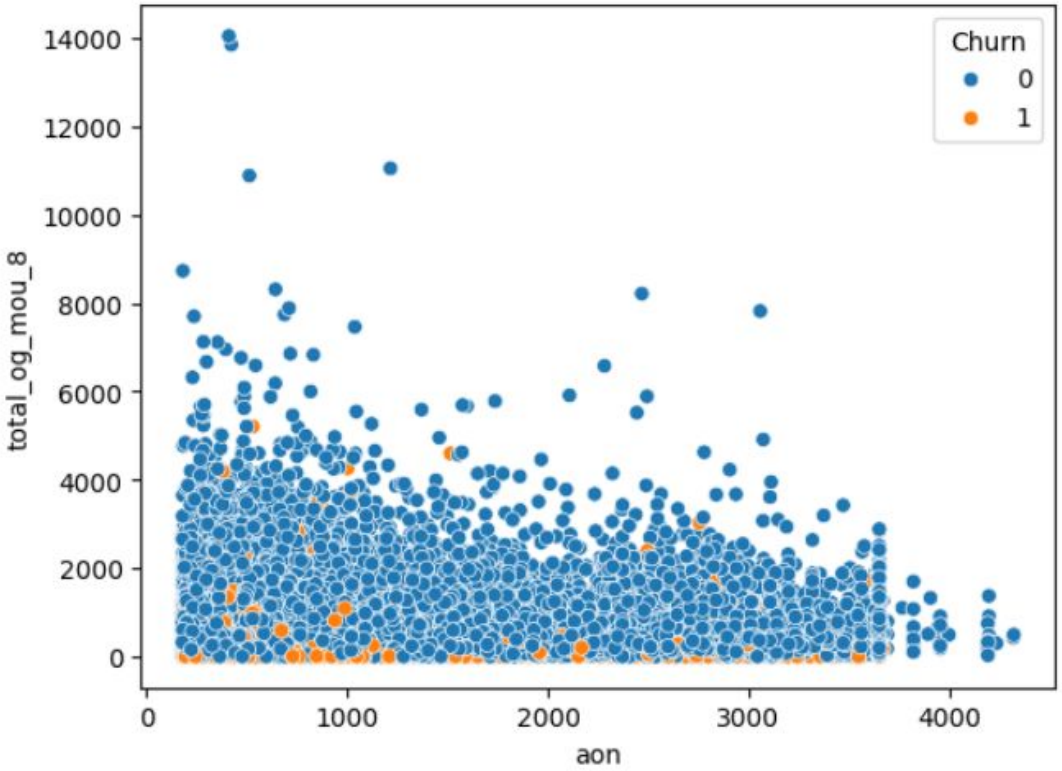


EDA - Bivariate Analysis



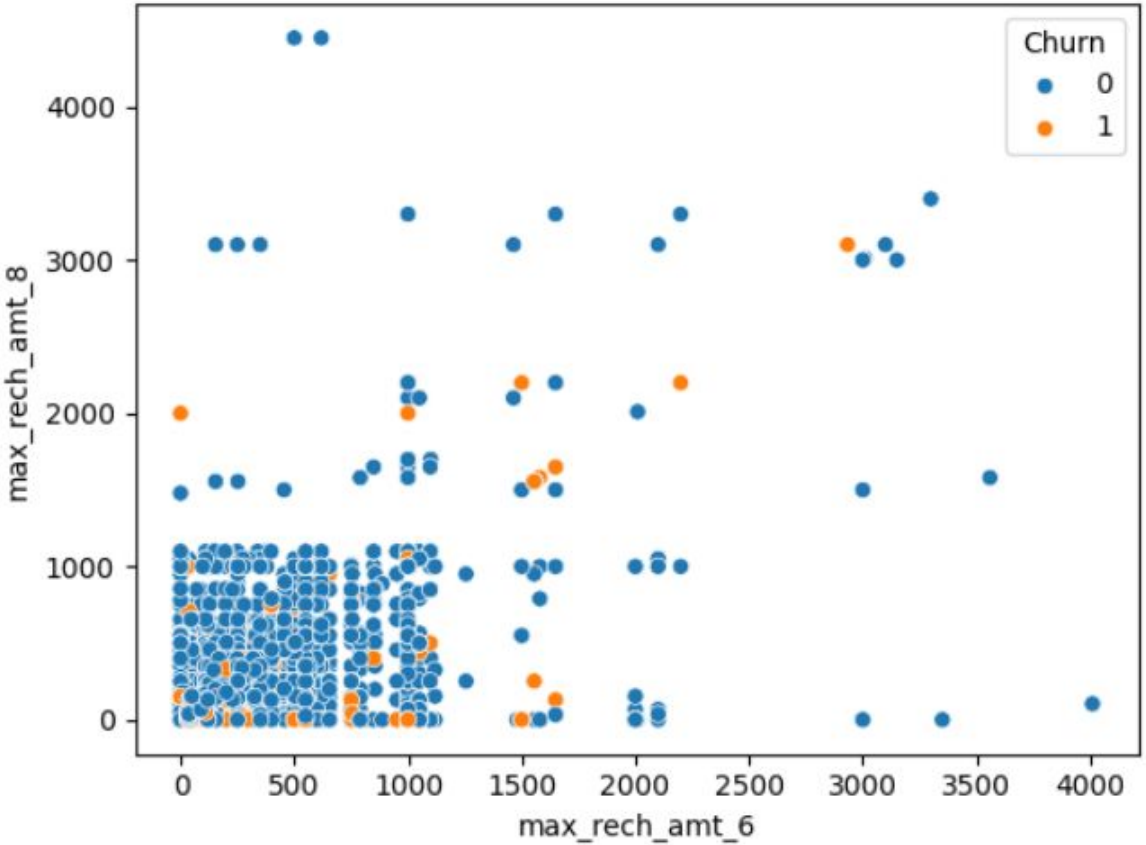
Customers with lower total_og_mou in 6th and 8th months are more likely to churn as compared to the ones with higher total_og_mou.

EDA - Bivariate Analysis

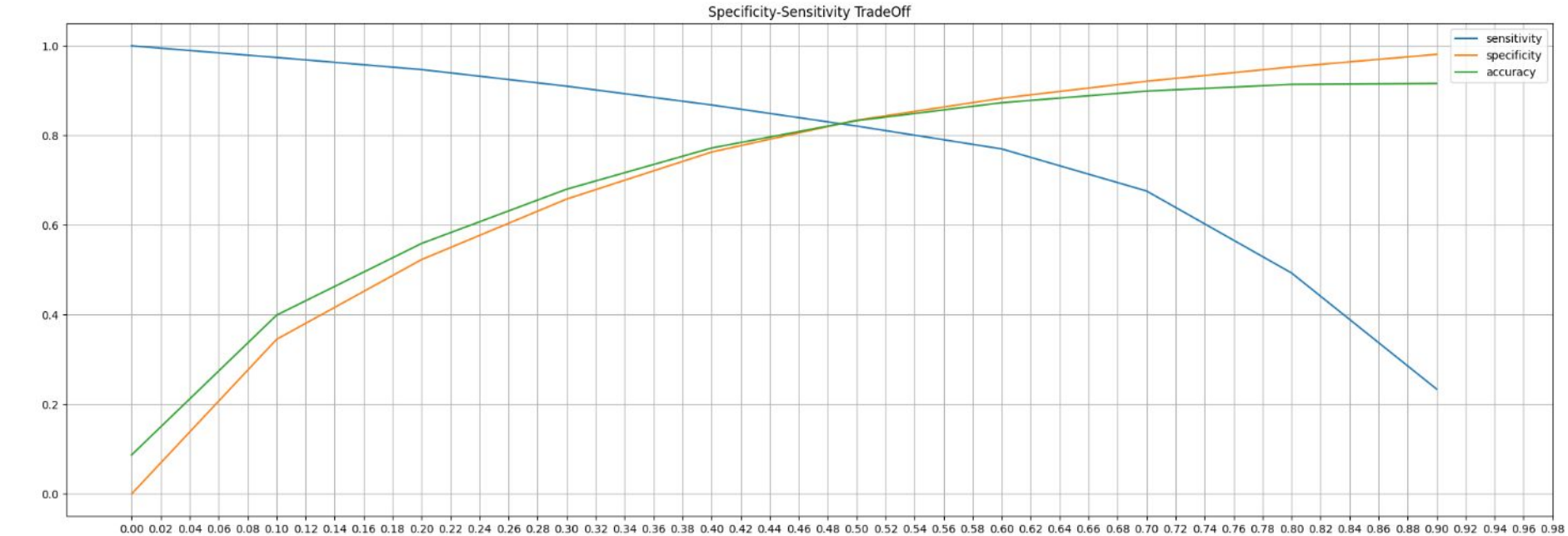


Customers with less total_ic_mou_8 are more likely to churn irrespective of aon.

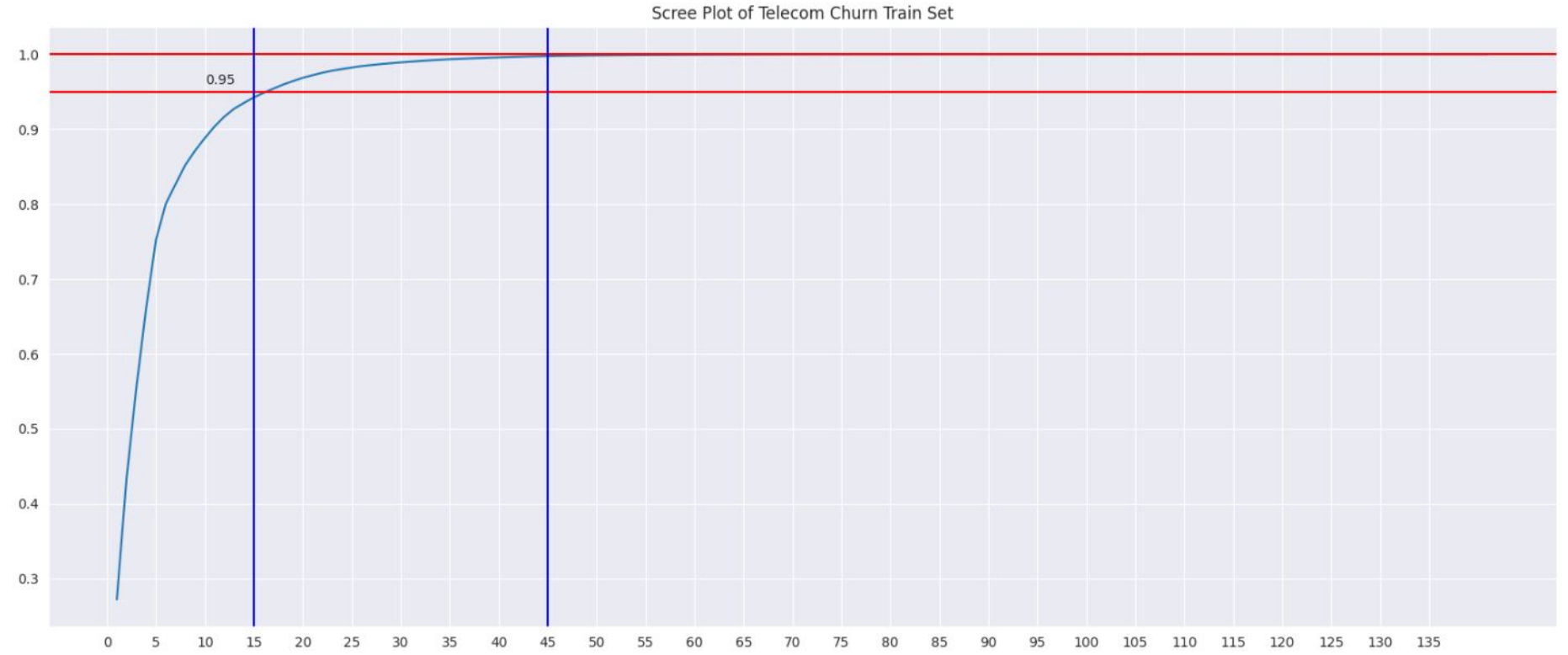
EDA - Bivariate Analysis



Modelling - Logistic Regression



PCA - Scree Plot



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.

Principal Components

Observations and Conclusions

- Customers who churn show lower average monthly local incoming calls from fixed line in the action period by 1.27 standard deviations, as 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.

Observations and Conclusions

- 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.

Recommendations

- The telecom company should concentrate on users with 1.27 std. deviations lower than average incoming calls from fixed line.
  They are most likely to churn.

- The telecom company should 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%.