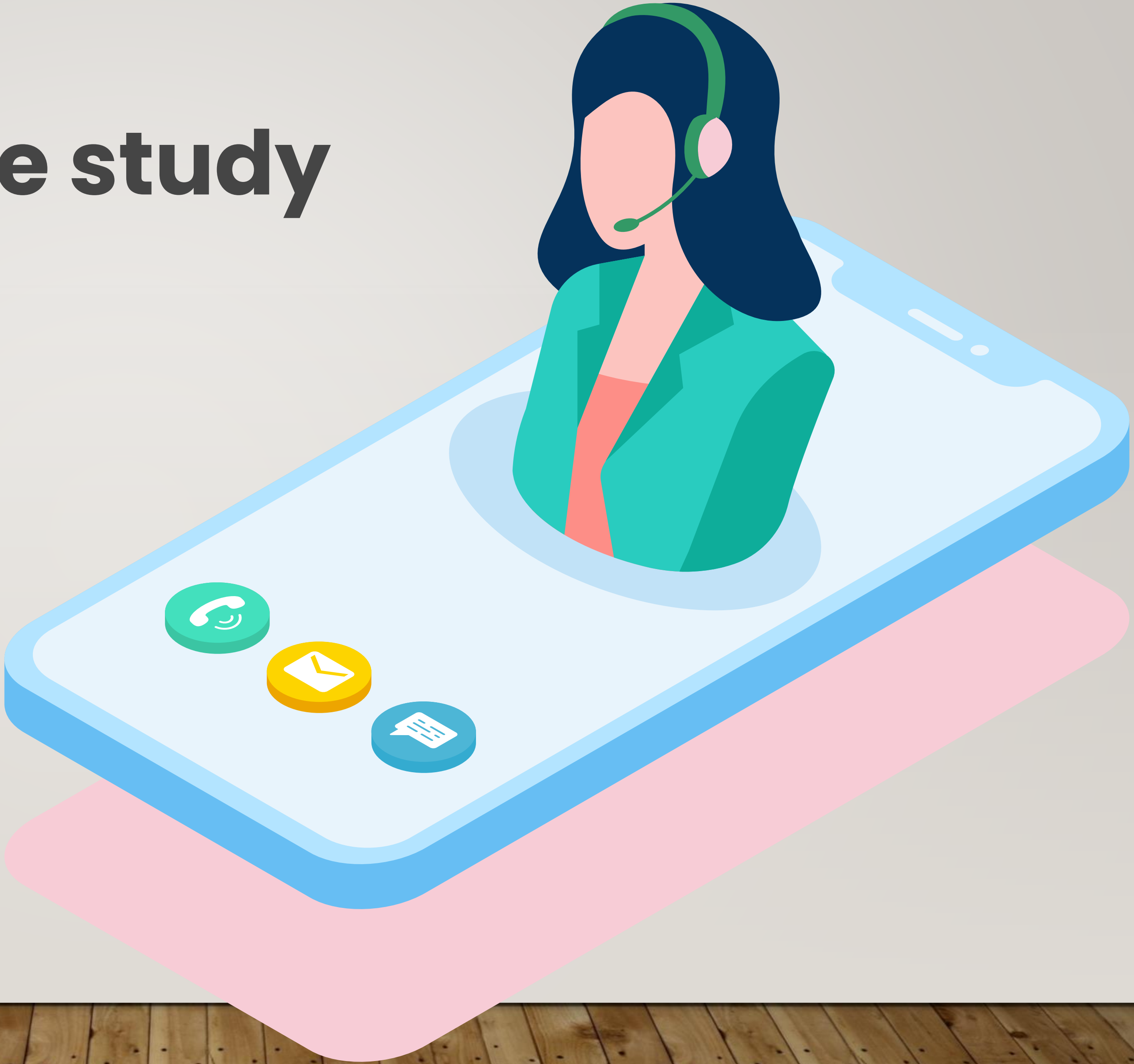
Telecom Churn Case study



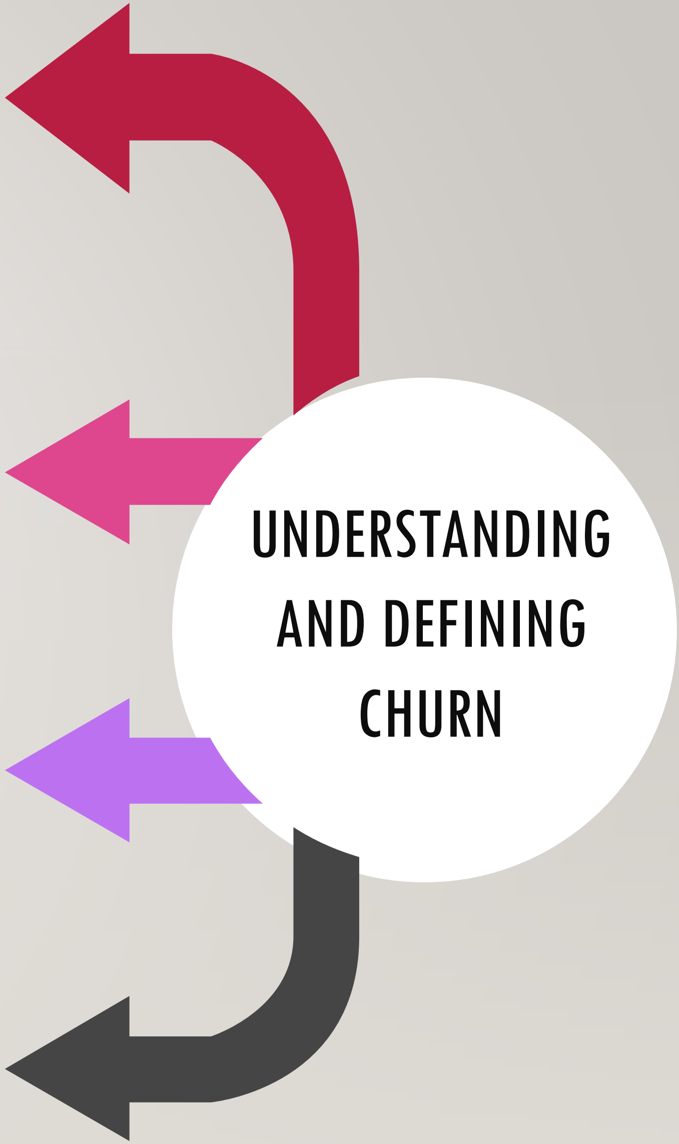
SHIVAKUMAR

DEEPTHY T BABU

R N HARIHARAN

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.



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).

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. The modelling based on Usage Based Churn

OBJECTIVES

1. TO DERIVE THE CHARACTERISTICS OF THE DATA BASED ON ATTRIBUTES PRESENT ON THE DATA

2. FILTER HIGH-VALUE CUSTOMERS TAG CHURNERS AND REMOVE ATTRIBUTES OF THE CHURN PHASE

3. RECOMMEND STRATEGIES TO MANAGE CUSTOMER CHURN BASED ON YOUR OBSERVATIONS..

Contents

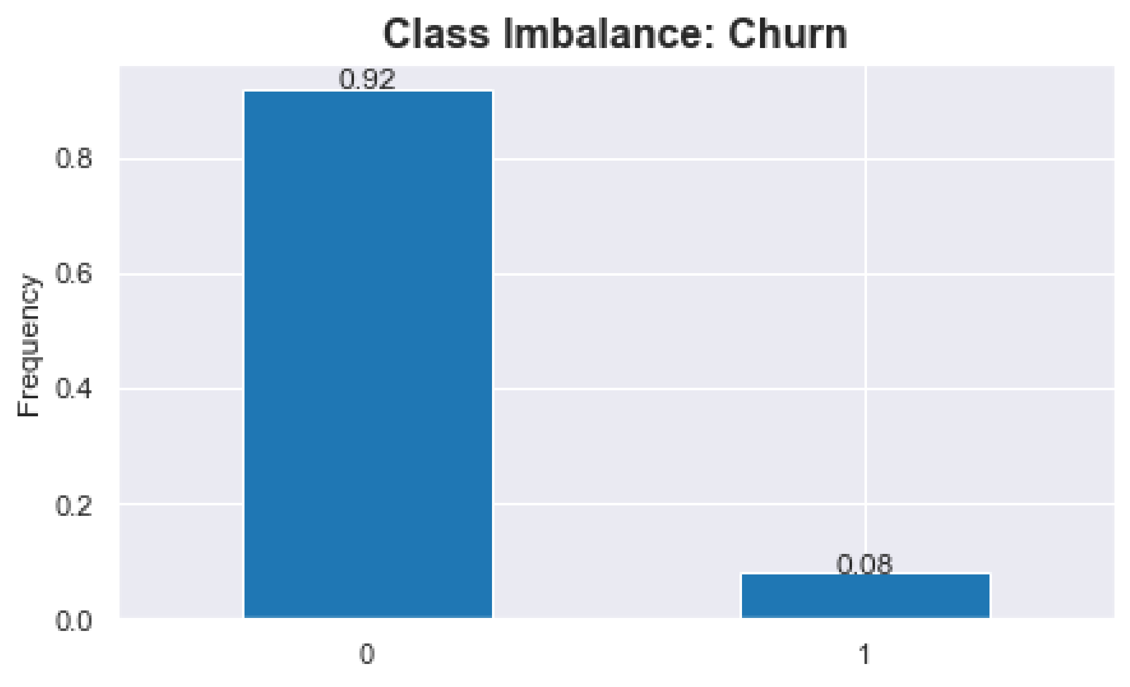
- Data Overview
- Class Imbalance
- Univariate Analysis of Categorical Variables
- Univariate Analysis for Numerical Variables
- Bivariate Analysis
- Multivariate Analysis
- Handling Data Imbalancing

Data overview

- There are 99999 rows and 226 columns in the churn data-set.
- Some columns representing volume based users have month specified in their name will update these to nos.
- Multiple columns have null values and will need to be treated.
- 15+ columns have only one unique value, and hence can be dropped as these are redundant from modelling perspective.
- The date feature is classified as object data-type, will convert these to data time format.
- Many of the numerical features potentially have outliers given the difference in the mean and median values.

CLASS IMBALANCE

- There is a clear class imbalance in the data given, 92% of the customers do not churn vs. only ~8% that churn.
- We will use appropriate methods to try and resolve this issue before proceeding with modelling.



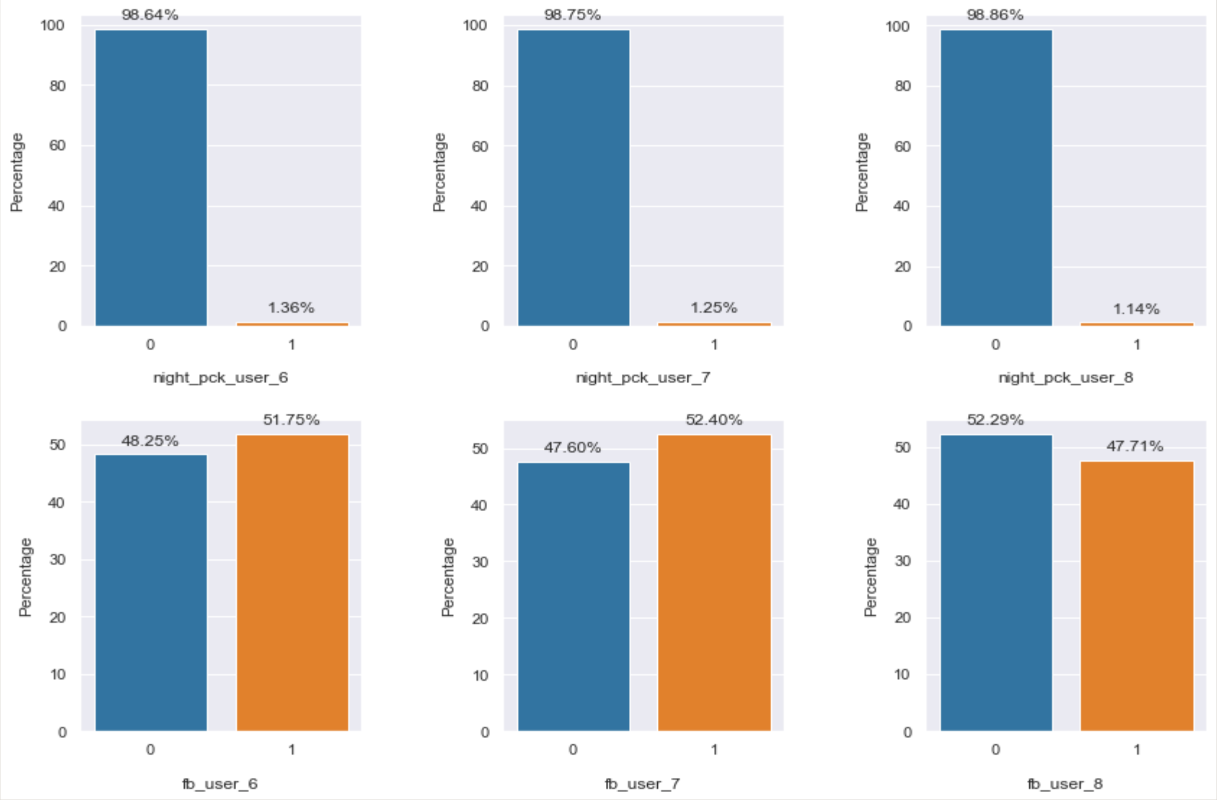
Handling Data Imbalancing

- Standard classifier algorithms like Decision Tree and Logistic Regression have a bias towards predicting majority class data. The features of the minority class are treated as noise and are often ignored. Thus, there is a high probability of misclassification of the minority class as compared to the majority class.
- Synthetic Minority Over-sampling Technique (SMOTE):

Given the class imbalance in our target variable churn, we will use SMOTE to resolve the issue of overfitting. SMOTE mitigates the problem of overfitting caused by random oversampling as synthetic examples are generated rather than replication of instances. Also, there is no loss of useful information as would have been the case with random under sampling.

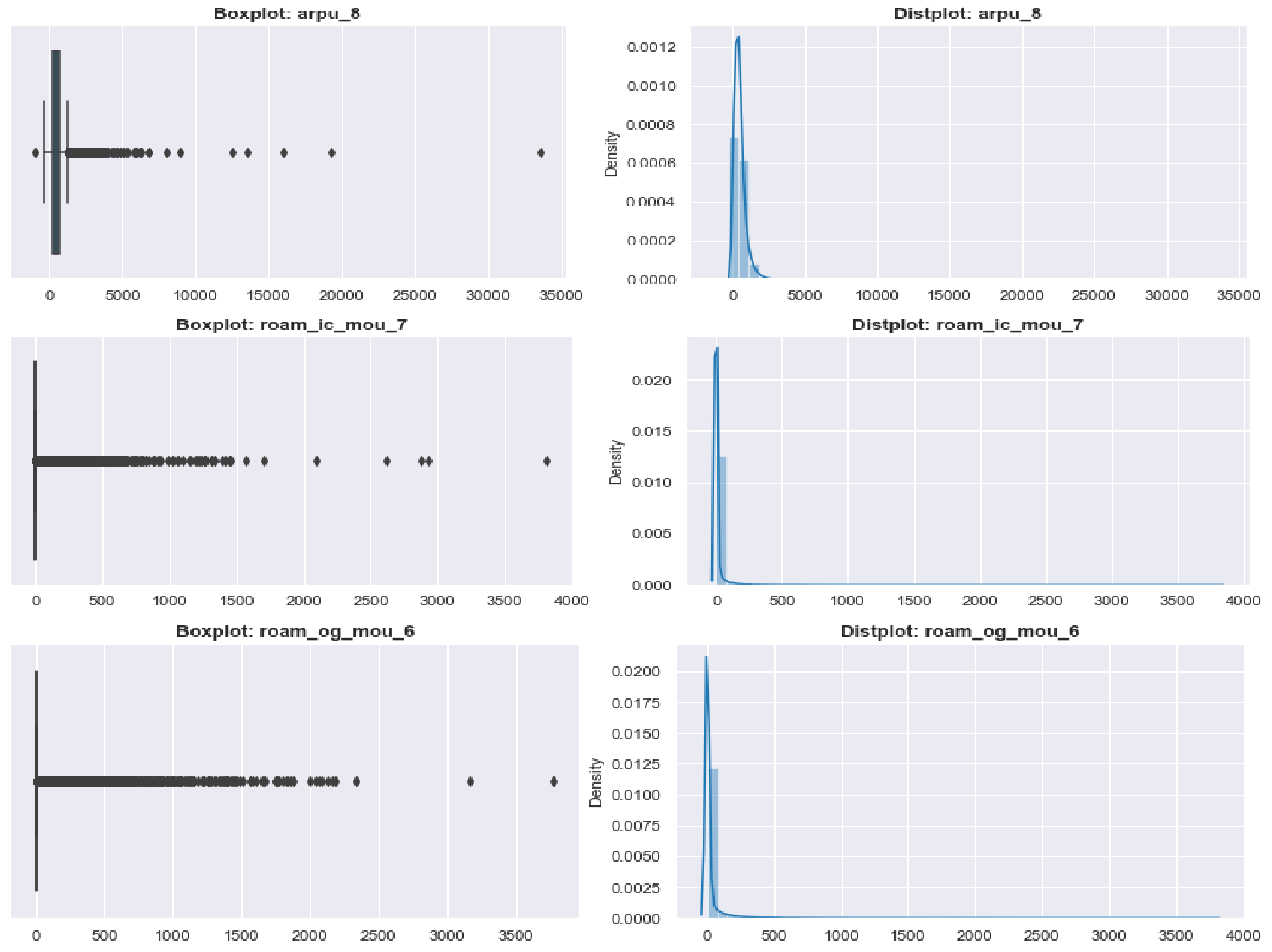
Univariate Analysis of Categorical Variables

- night_pck_user: ~99% of users do not use the nightly pack and this is consistent across all the 3 months.
- fb_user: There is an almost equal split between fb_users among churners and non-churners. However, by the 8th month the churners seems to decline their usage to some extent.

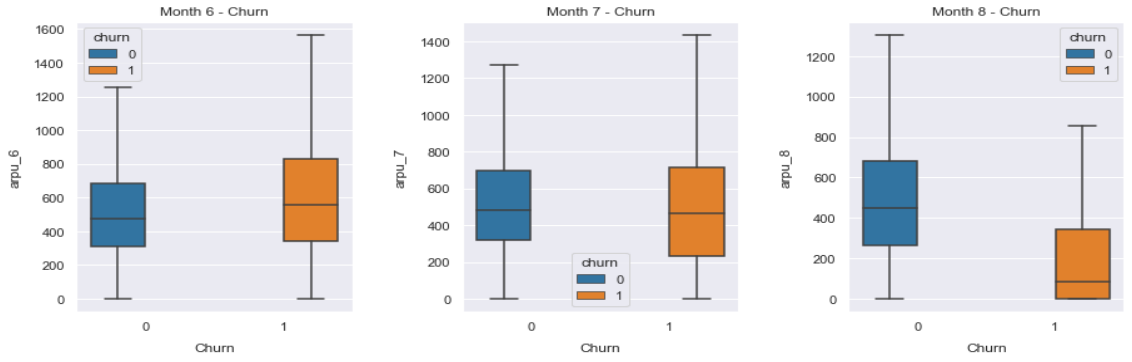


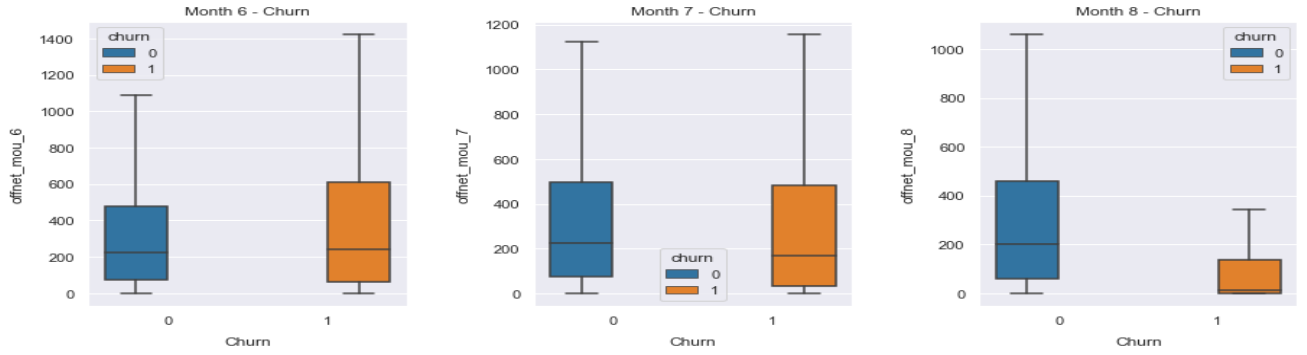
Univariate Analysis for Numerical Variables

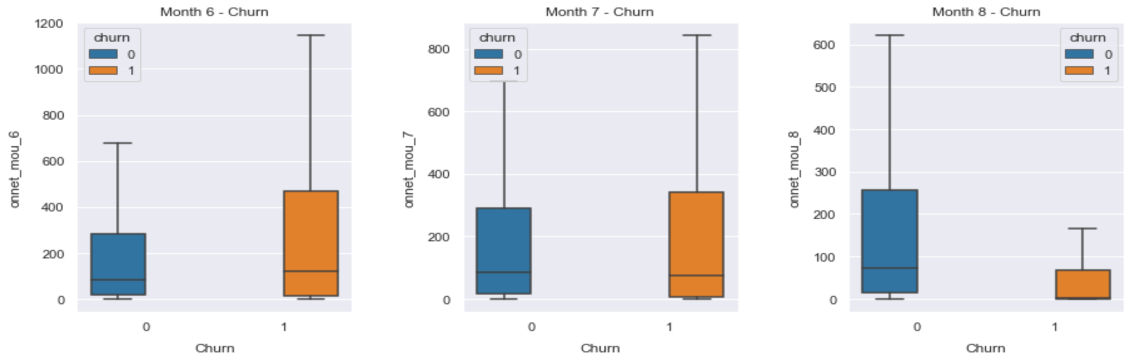
- Most of the columns have outliers
- Capping of outliers is done by Standard Deviation method

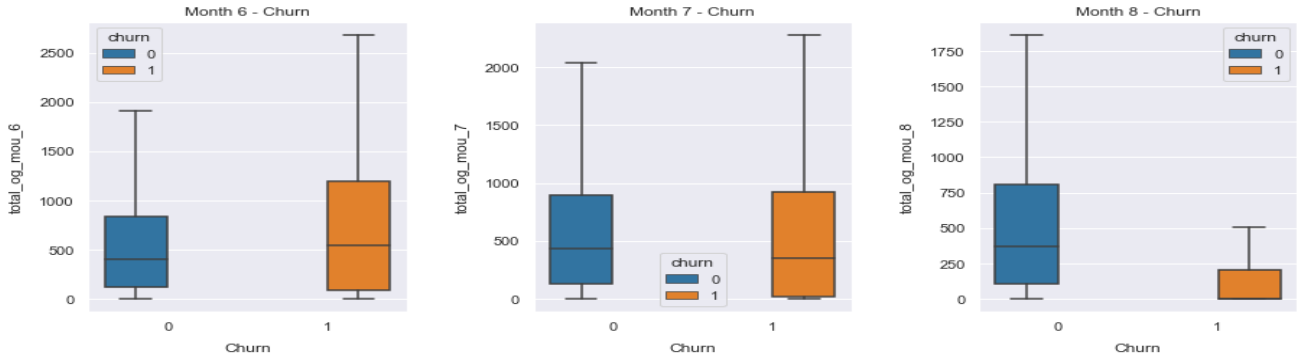


Bivariate Analysis





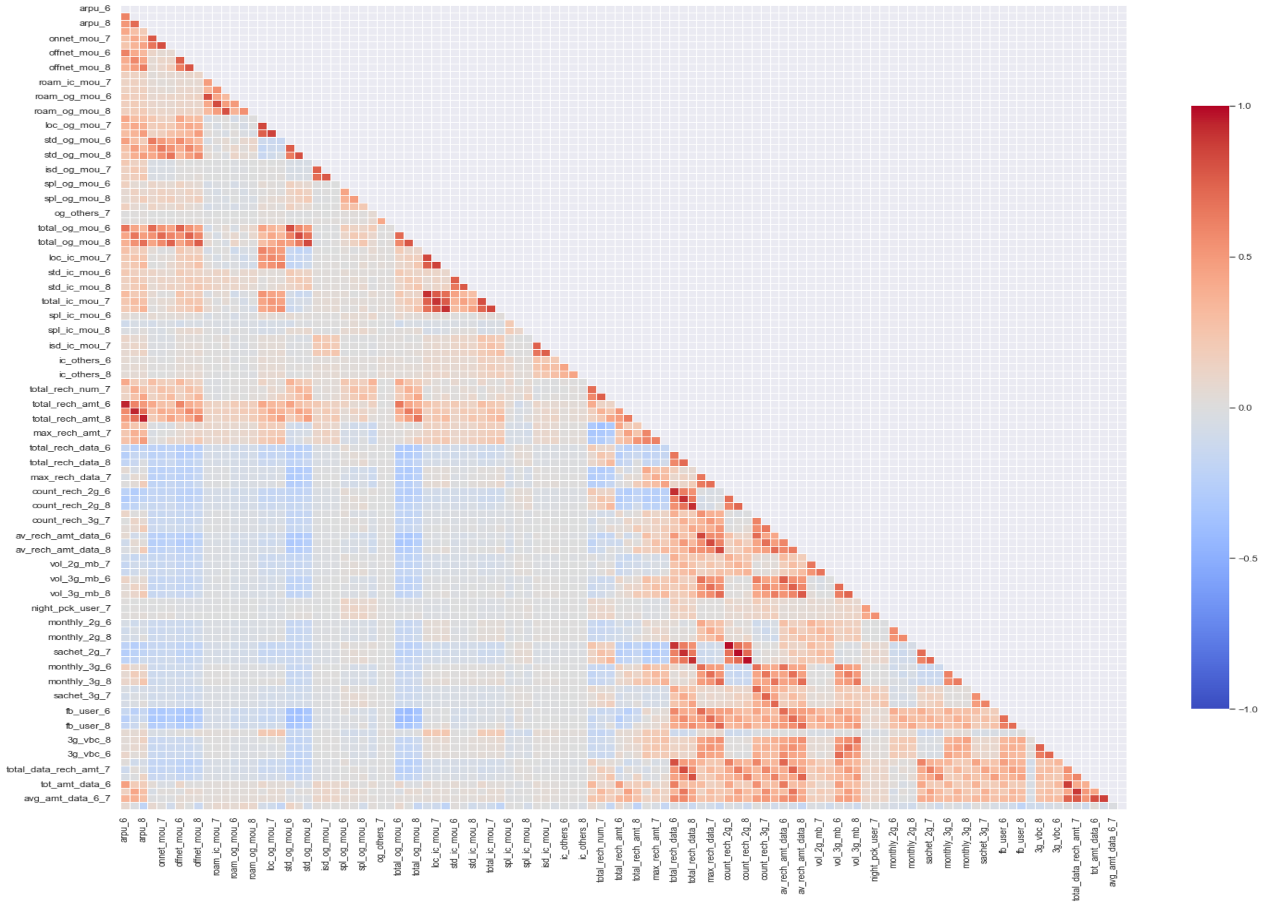




- Across most features we analyzed, the mean values for month 8 drop drastically for the churn customers vs. more or less consistent usage pattern across non-churners.

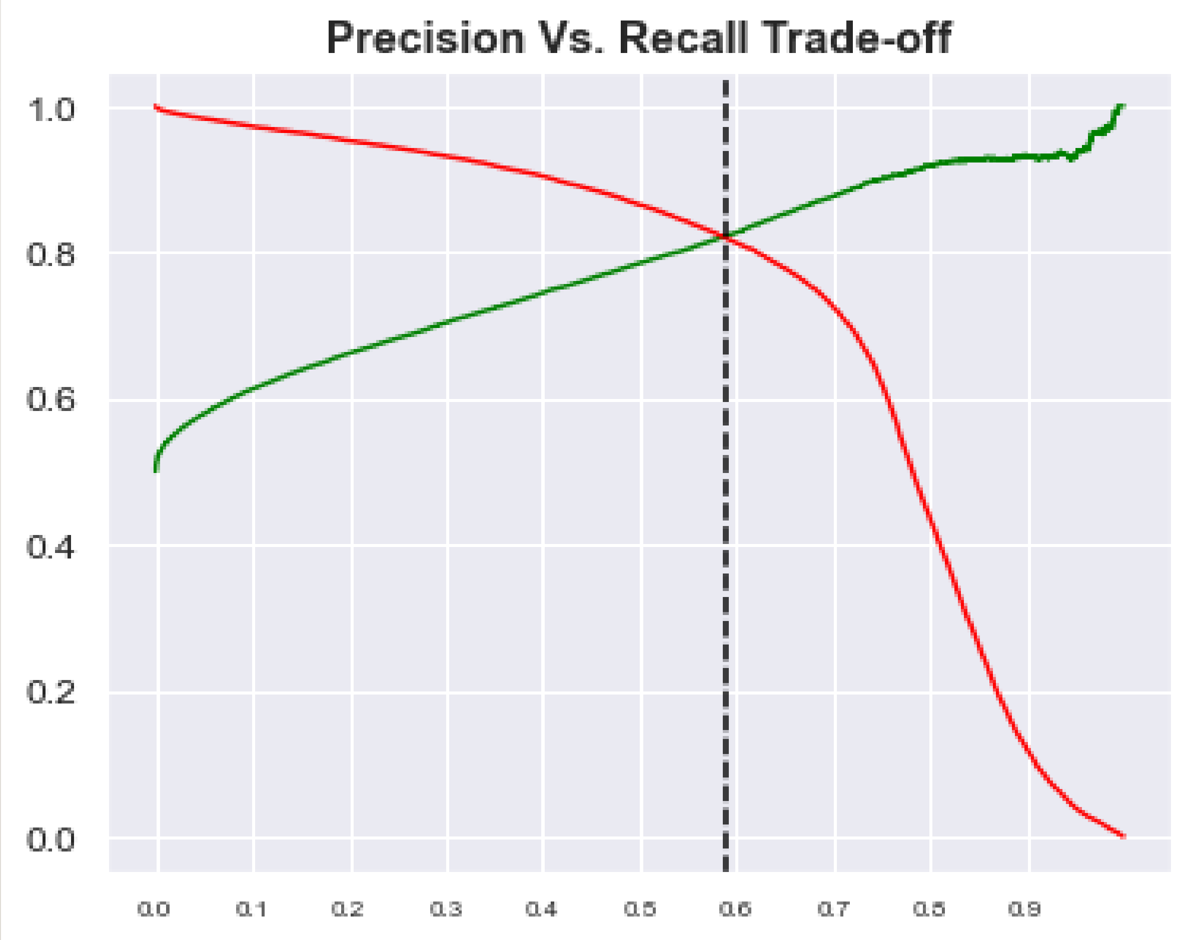
Multivariate Analysis

- There are ~30 features have correlation >=80%.



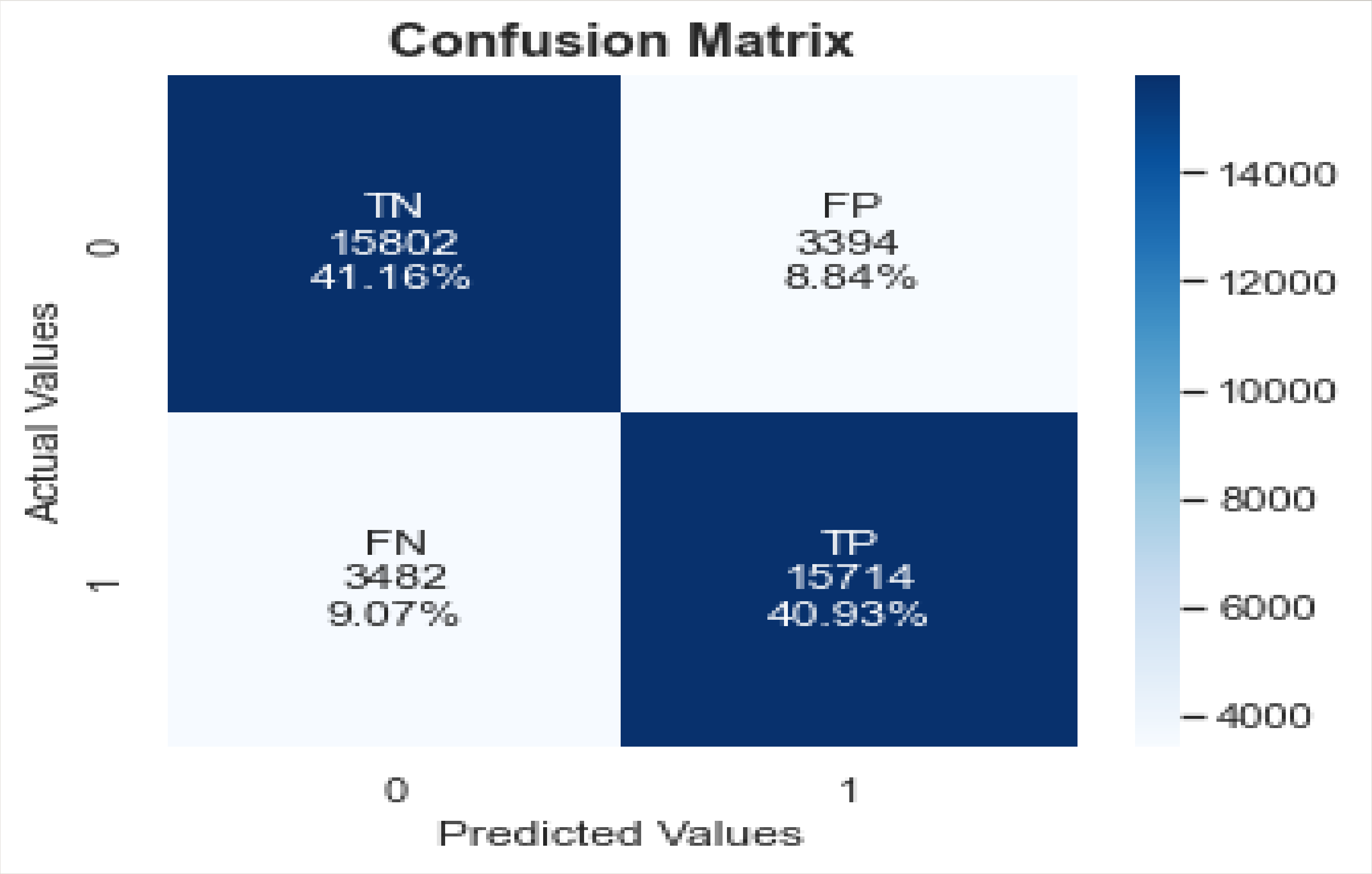
Precision and recall tradeoff

- As seen from above, there is tradeoff between Precision and Recall.
- Precision and Recall are inversely related means if one increases other will genuinely decrease
- Based on Precision- Recall Trade off curve, the cutoff point seems to 0.59.



Confusion matrix

- ➢True Negative : 15802- 41.16%
- ➢False Negative : 3394- 8.84%
- ➢False Negative : 3482- 9.07%
- ➢True Positive: 15714- 40.93%



Plotting the ROC Curve

- While the accuracy and recall were more or less consistent across both train and test data-set, the precision and F-1 score took a major hit in the test set
- Given the our business problem, AUC is likely the most appropriate metric to use, since the Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.
- Given teh AUC remained consistent at ~0.88-0.89 in the train and test data set, the logistic regression model is performing well.

Final Model Selection

- Models approach:
- To predict behavior of a customers over a time period of 4 months split into good phase (first 2 months) and the action phase (3rd month) we created about 3 predictive models as mentioned below:
- 1) Logistic Regression Model
- 2) Decision Tree Classifier
- 3) Random Forest Classifier
- The above models were initially created with default parameters which were then hypertuned and then used to predict using the best estimators. The hyper tuned model showed an increase in the classification scores though marginally. These scores were still not good enough to say that any of these model are decent predictors of churn.

Final Model Selection

➢ Evaluation Methods:

- While, we calculated the accuary, precision, recall, classification reports, AUC-score etc for predictions from each of these models. We think AUC-Score is most reliable given its capability to distinguish between classes which is the business objective for us.

➢ Best Model:

- Based on the AUC scores, our hyper-parameter tuned Random Forest Classifier model performed the best with an AUC-Score of 0.85 vs. the other models the AUC-scores for which are below:

➢ AUC scores:

- Logistic regression (optimal cut-off) : 0.814
- Decision Tree Classifier (Base Model) : 0.831
- Decision Tree Classifier (Hyper-Parameter Tuned Model): 0.832 Random Forest Classifier (Base Model) : 0.836
- Random Forest Classifier (Hyper-Parameter Tuned Model): 0.849

Strategies to manage customer churn

- Monitoring drop in usage seems to be a strong predictor of churn and hence should be monitored on a regular basis.
- Usage of roaming services among churners is high, suggesting network quality and service issues on roaming can be a possible reason for churn.
- The Network operators must further monitor and provide competitive roaming tariffs, better quality of both network and services.
- Given the extreme competition among competitor networks, it is extremly crucial to monitor competitor marketing campaigns.
- Marketing team should run more campaigns which target customers who are high value users of roaming services. Like,
  - Discounted roaming rates during particular hours of the day.
  - Free monthly roaming mou's depending on the users past roaming mou usage.

THANK YOU