DOMAIN ORIENTED CASE STUDY

(Telecom churn case study)

Submitted by
Rekha V S

PROBLEM STATEMENT & BUSINESS OBJECTIVE

Problem statement:

- ❑ The case study aims to build a predictive model to help reduce customer churn of telecom companies by predicting which customers are at high risk of churn.
- ❑ 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. It is required to 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.

Business Objective:

- ❑ To predict the churn in the last month using the data (features) from the first three months. The dataset containing customer-level information for a span of four consecutive months - June, July, August and September is available

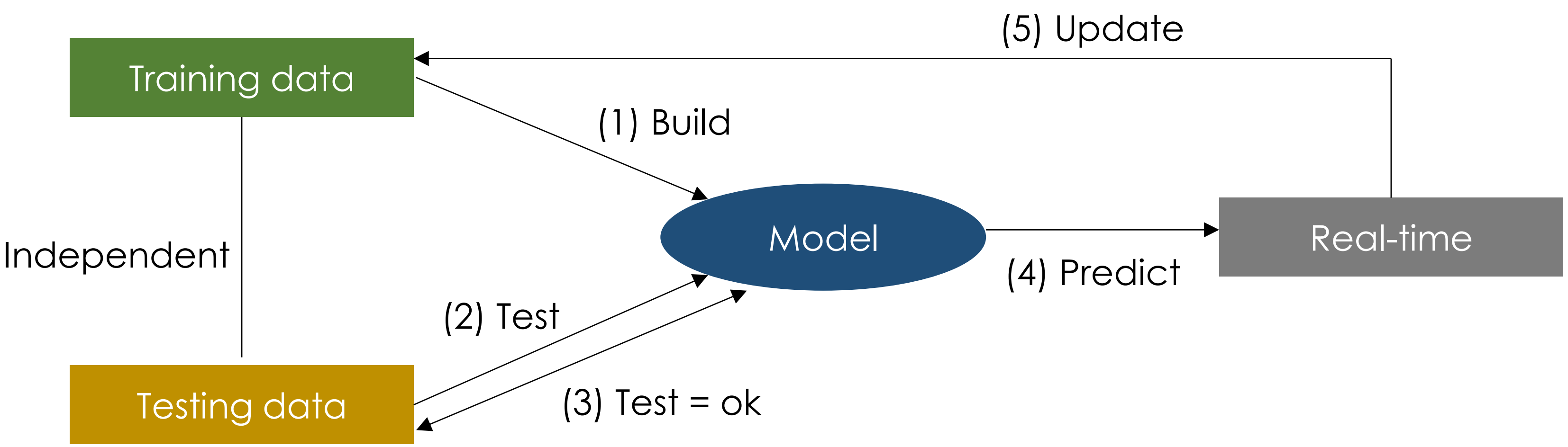
METHODOLOGY & WAY FORWARD

- ❑ Extensive experimentation with different models done to identify the best model that can predict the customer behavior which can help the company to take proactive steps to retain the customer.

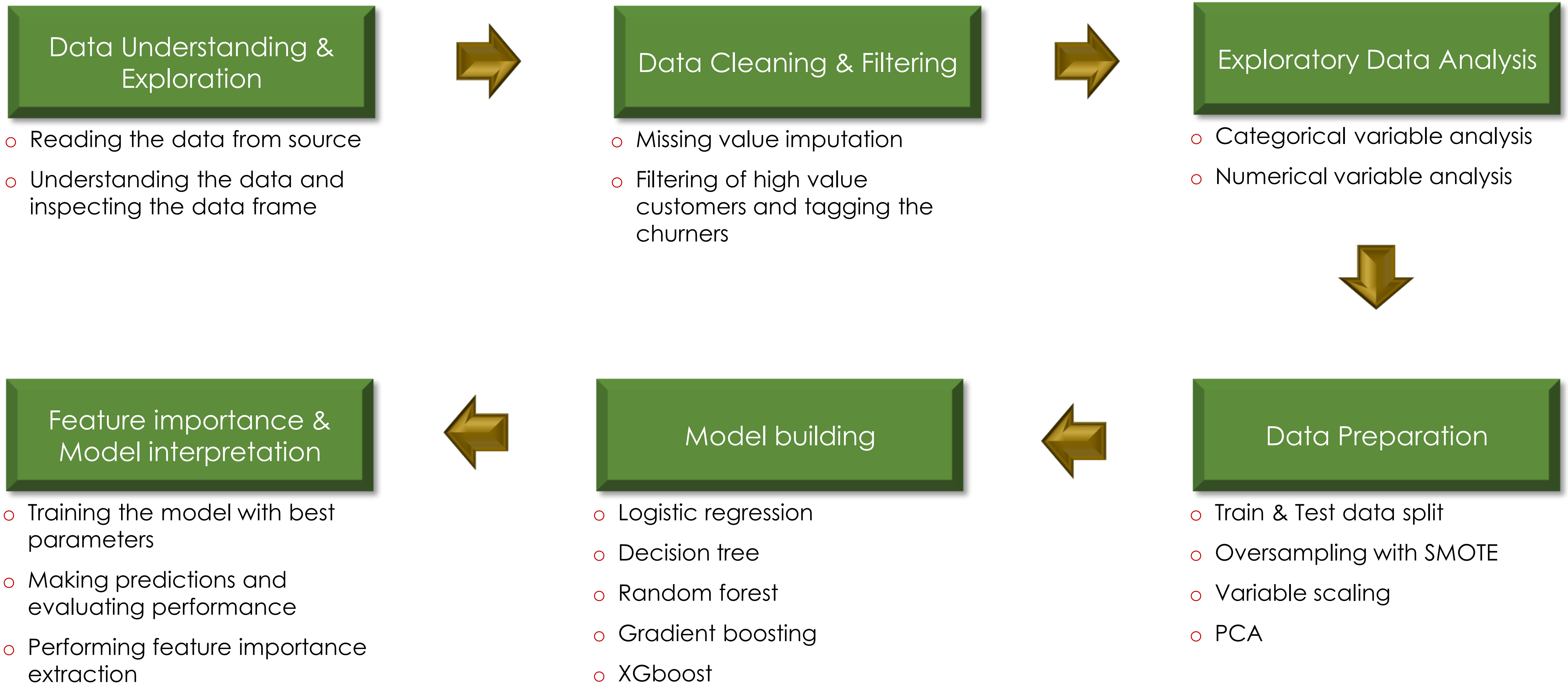
- ❑ Because of imbalance in data set, oversampling performed with SMOTE on training dataset & to remove collinearity and faster training dimensionality reduction technique PCA is performed

- ❑ For model building, below models are explored:
  - Logistic regression
  - Decision tree
  - Random forest
  - Gradient boosting
  - XGboos

Basic skeleton representation



STEPS FOLLOWED



Data Understanding & Exploration

- Reading the data from source
- Understanding the data and inspecting the data frame

Data Cleaning & Filtering

- Missing value imputation
- Filtering of high value customers and tagging the churners

Exploratory Data Analysis

- Categorical variable analysis
- Numerical variable analysis

Data Preparation

- Train & Test data split
- Oversampling with SMOTE
- Variable scaling
- PCA

Model building

- Logistic regression
- Decision tree
- Random forest
- Gradient boosting
- XGboost

Feature importance & Model interpretation

- Training the model with best parameters
- Making predictions and evaluating performance
- Performing feature importance extraction

DATA UNDERSTANDING & EXPLORATION

- Read the data from the 'telecom_churn_data' CSV file and understood the meaning of variables from 'Data Dictionary'
- The dataset contains total of 99999 rows and 226 columns
- There are missing values in the dataset, with few columns having >70% of null values. They need to be treated
- Datatypes of the column : 179 float64, 35 int64, 12 object (214 Numeric and 12 categorical columns)
- The variables circle_id, loc_og_t2o_mou,std_og_t2o_mou and loc_ic_t2o_mou have same minimum and maximum value with 0 standard deviation. This implies that they have a single unique value and hence such columns can be dropped
- There are no duplicate records as the number of rows are same when cross verified through mobile number (unique identifier)

Note: Only key points/insights from the analysis is added in the presentation in the best possible '**concise**' way for all the steps.

Other analysis are detailed in the notebook

DATA CLEANING AND FILTERING

- Dropped the columns not useful for analysis. E.g.,

  - Following columns have only one value, i.e., last day of corresponding month.

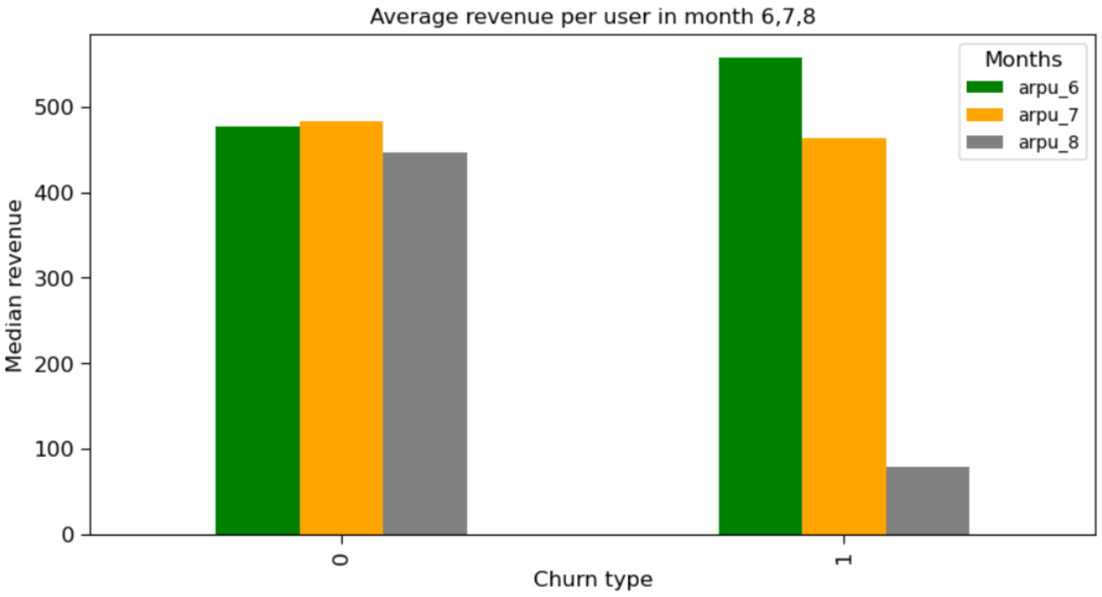
    Hence, we can drop them. "last_date_of_month_6", "last_date_of_month_7", "last_date_of_month_8", "last_date_of_month_9“

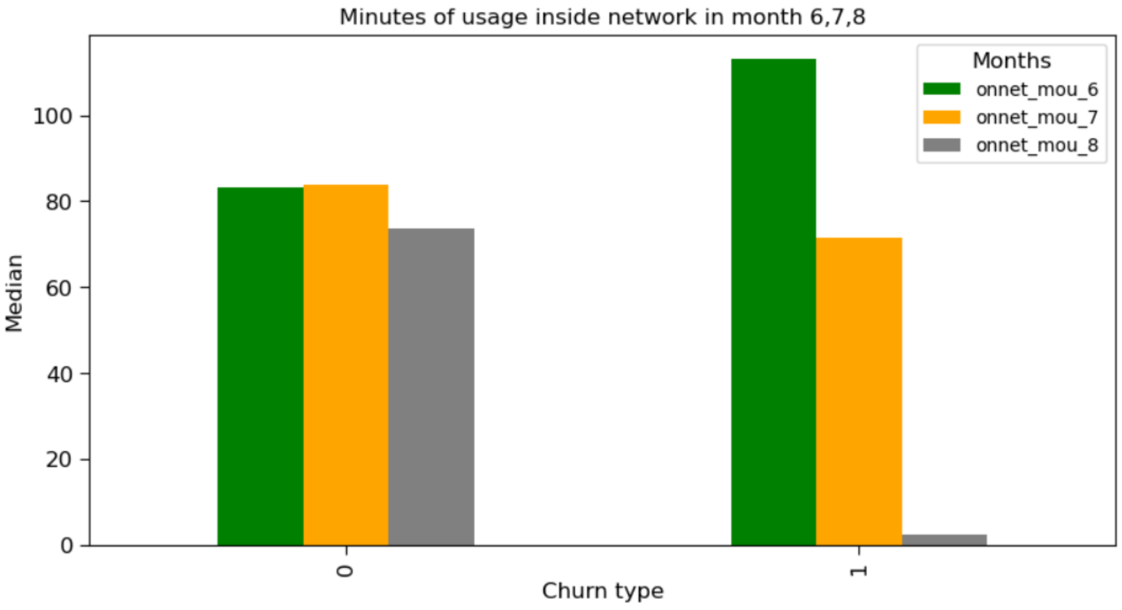
  - Also, date of Last recharge of corresponding month can be dropped, as we will not get useful insights from them

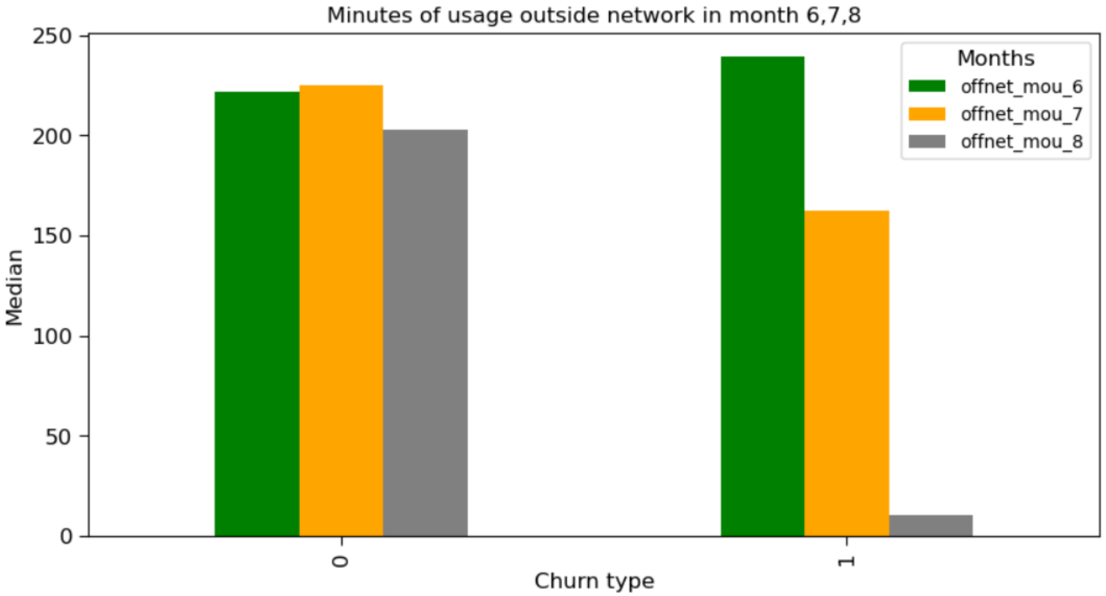
  - Dropping ID columns as well as they will not contribute anything for analysis
- Missing value treatment done through column-by-column analysis and imputation done with apt means
- Defined the high value customers and filtered them using recharge amount in the first two months. There were 30001 rows of high value customers with 185 columns
- Tagged the churners and removed the attributes of the churn phase
- As the churn percentage is close to 8% and non-churners are approx 92%, the data is imbalanced

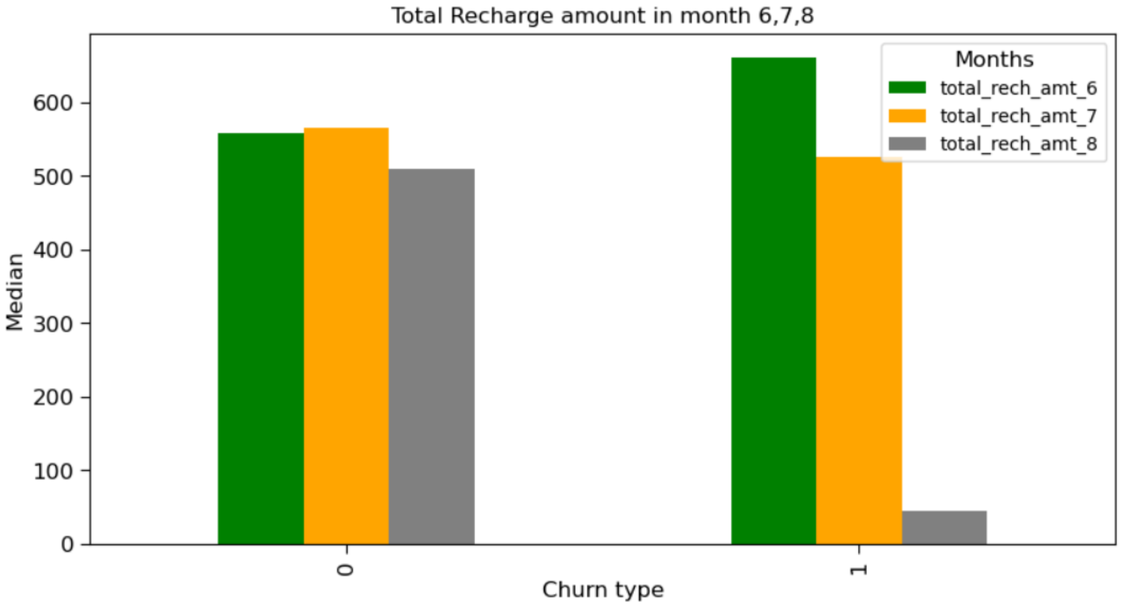
EXPLORATORY DATA ANALYSIS

- Deriving new features by comparing 8th month features vs average of 7th and 6th month features

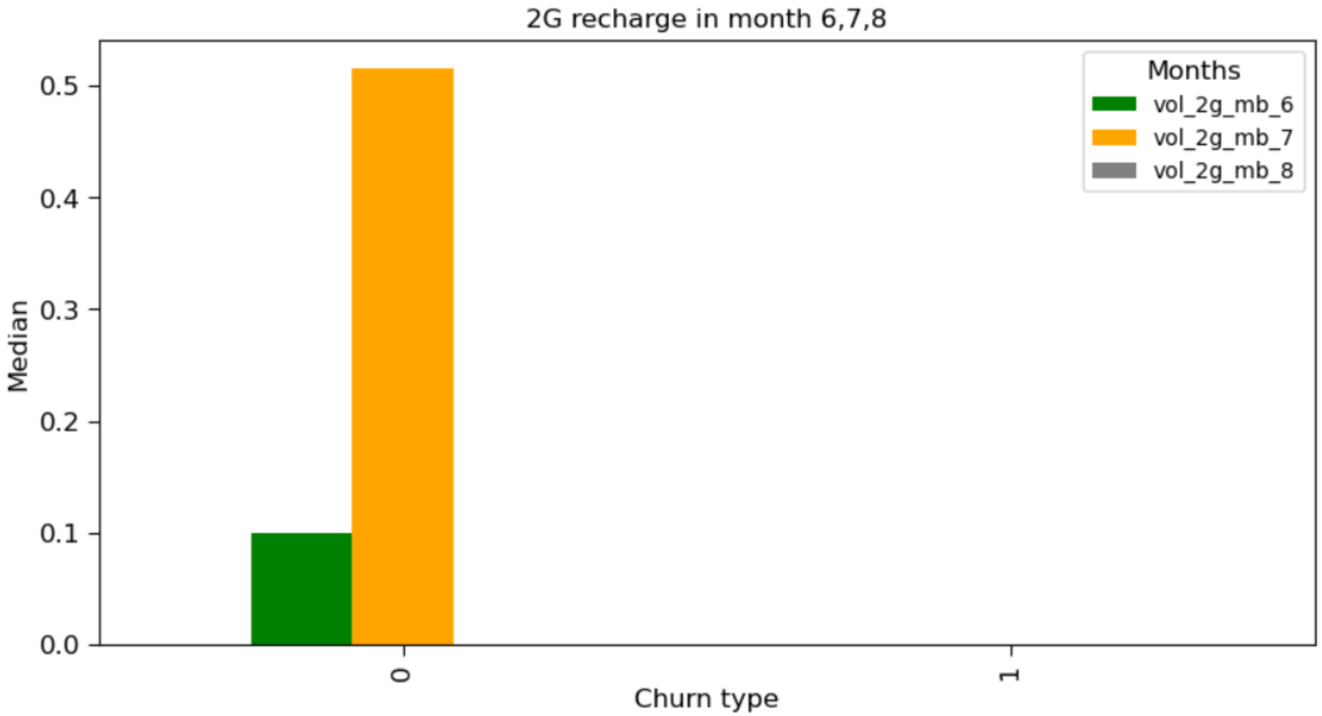

Average revenue per user in month 6,7,8


Minutes of usage inside network in month 6,7,8


Minutes of usage outside network in month 6,7,8

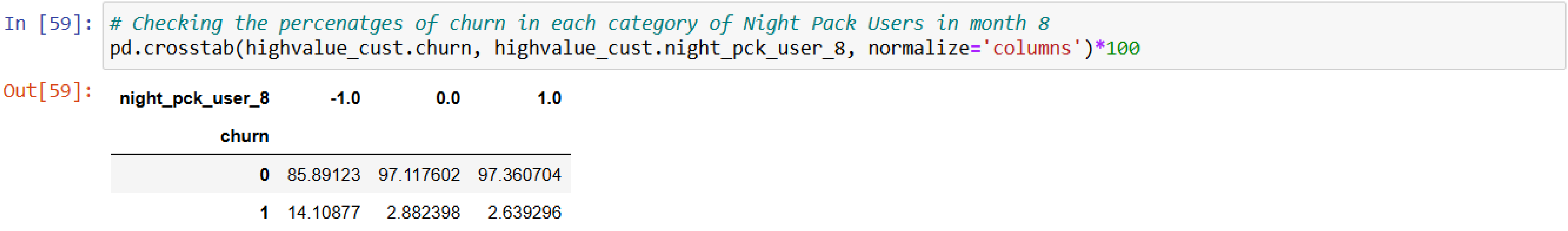

Total Recharge amount in month 6,7,8

EXPLORATORY DATA ANALYSIS



- If Average revenue per user is more in month 6 means, if they are unsatisfied, they are more likely to churn
- Users whose minutes of usage are more in month 6, they are more likely to churn.
- Users with big difference of minutes of call duration to other network between month 6 and month 7 are likely to churn.
- Also, when the difference of total recharge amount is more, those users are more likely to churn.
- 2g recharge who have not done may or may not churn as there is no concrete evidence from data

EXPLORATORY DATA ANALYSIS

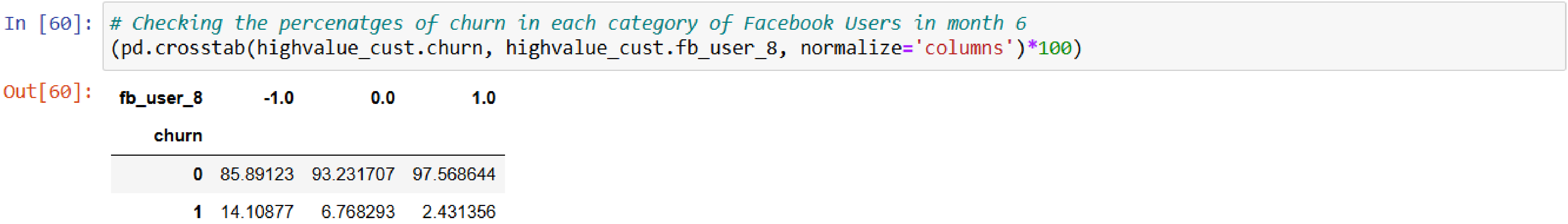


In [59]:

```python
# Checking the percenatges of churn in each category of Night Pack Users in month 8
pd.crosstab(highvalue_cust.churn, highvalue_cust.night_pck_user_8, normalize='columns')*100
```

Out[59]:

night_pck_user_8	-1.0	0.0	1.0
churn			
0	85.89123	97.117602	97.360704
1	14.10877	2.882398	2.639296



In [60]:

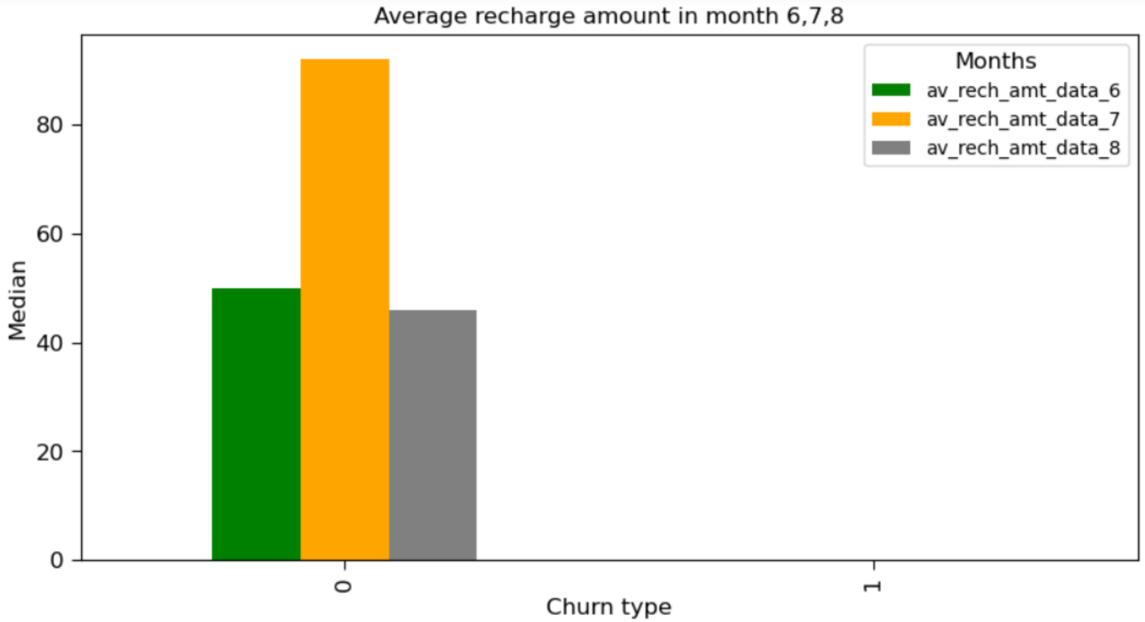
```python
# Checking the percenatges of churn in each category of Facebook Users in month 6
(pd.crosstab(highvalue_cust.churn, highvalue_cust.fb_user_8, normalize='columns')*100)
```

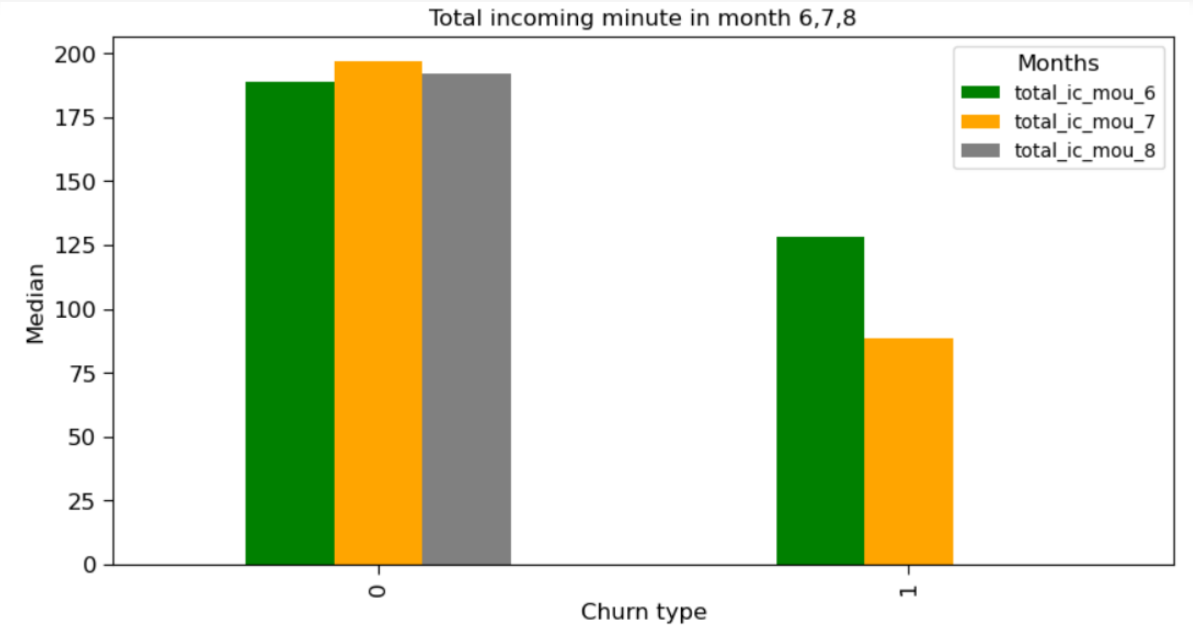
Out[60]:

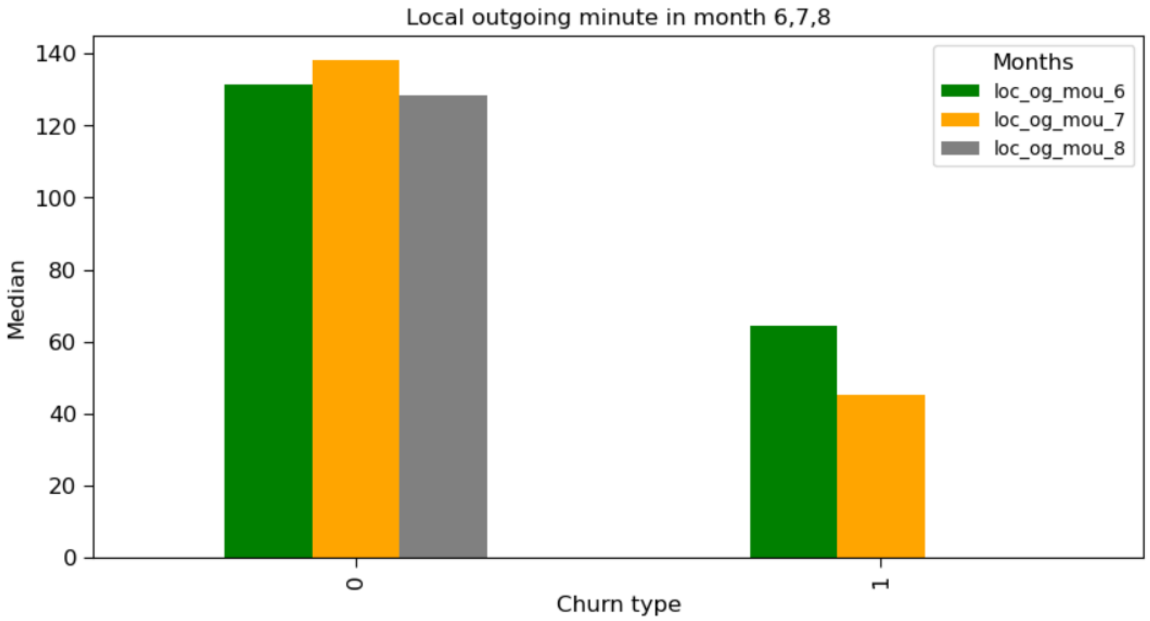
fb_user_8	-1.0	0.0	1.0
churn			
0	85.89123	93.231707	97.568644
1	14.10877	6.768293	2.431356

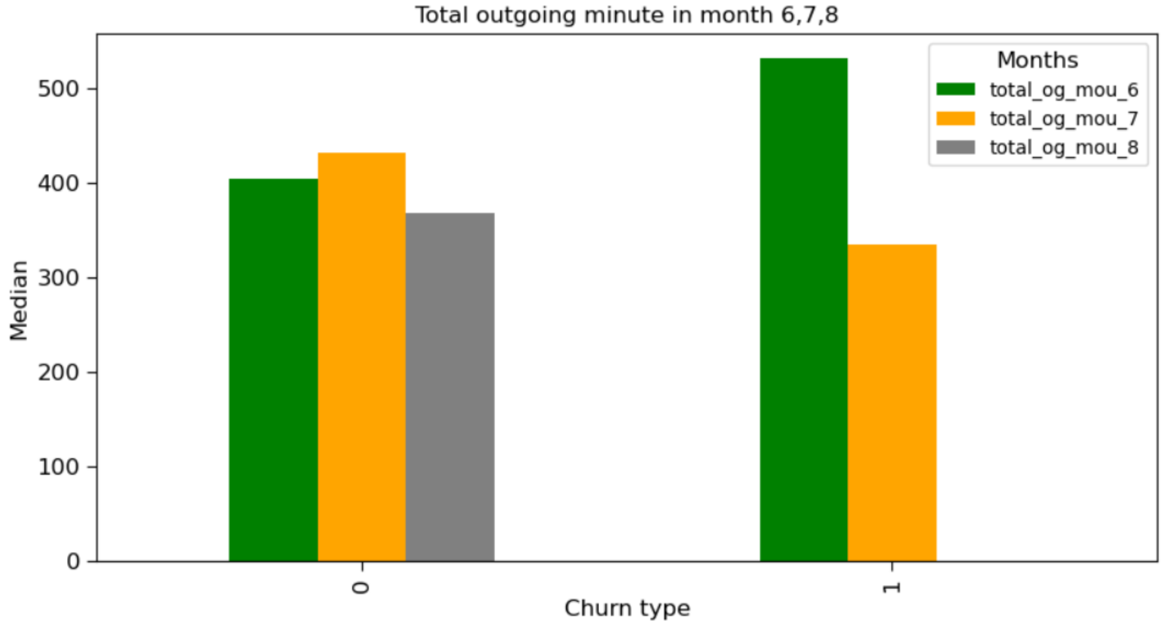
- For Night pack users in month 8, the churn rate is high. i.e., close to 14%
- Close to 2% churn among Facebook users in month 8
- Customers not using facebook, close to 7% churn in month 8

EXPLORATORY DATA ANALYSIS

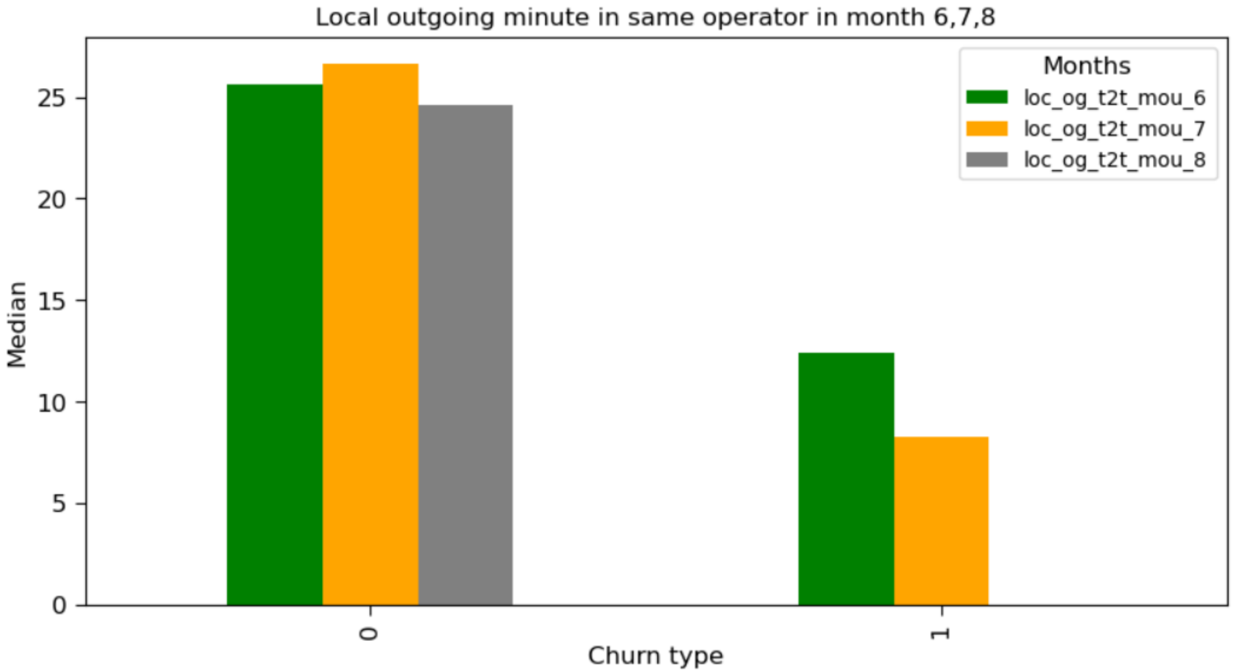


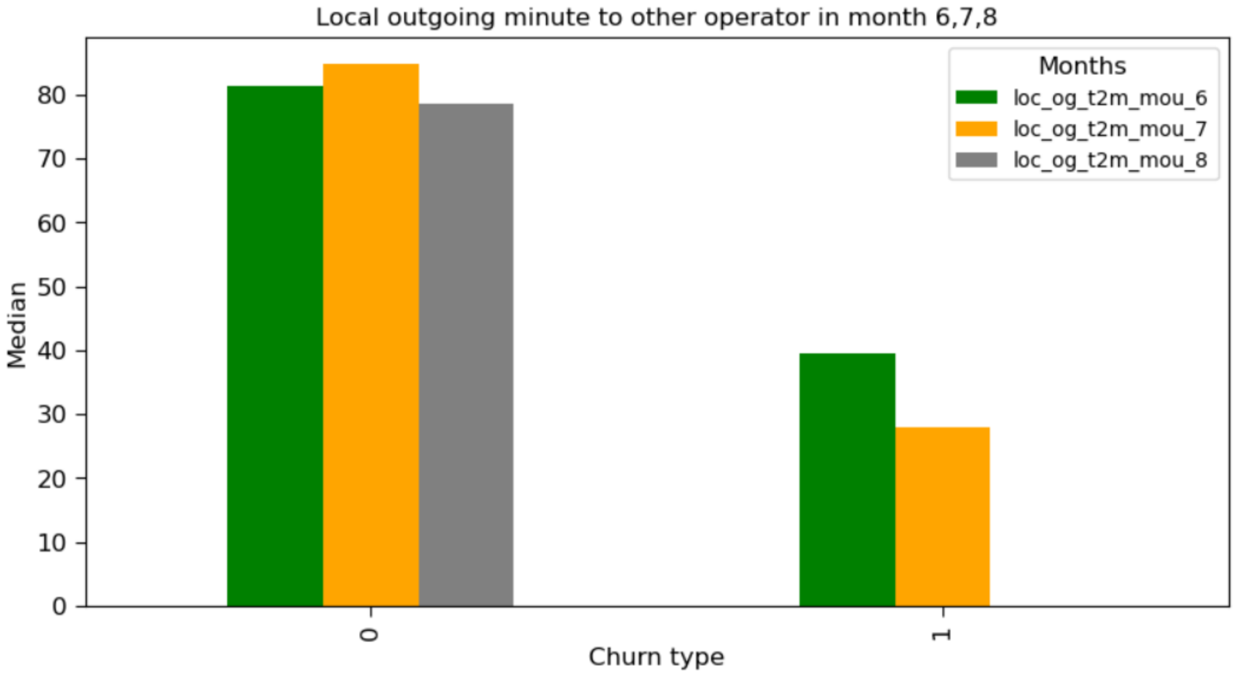






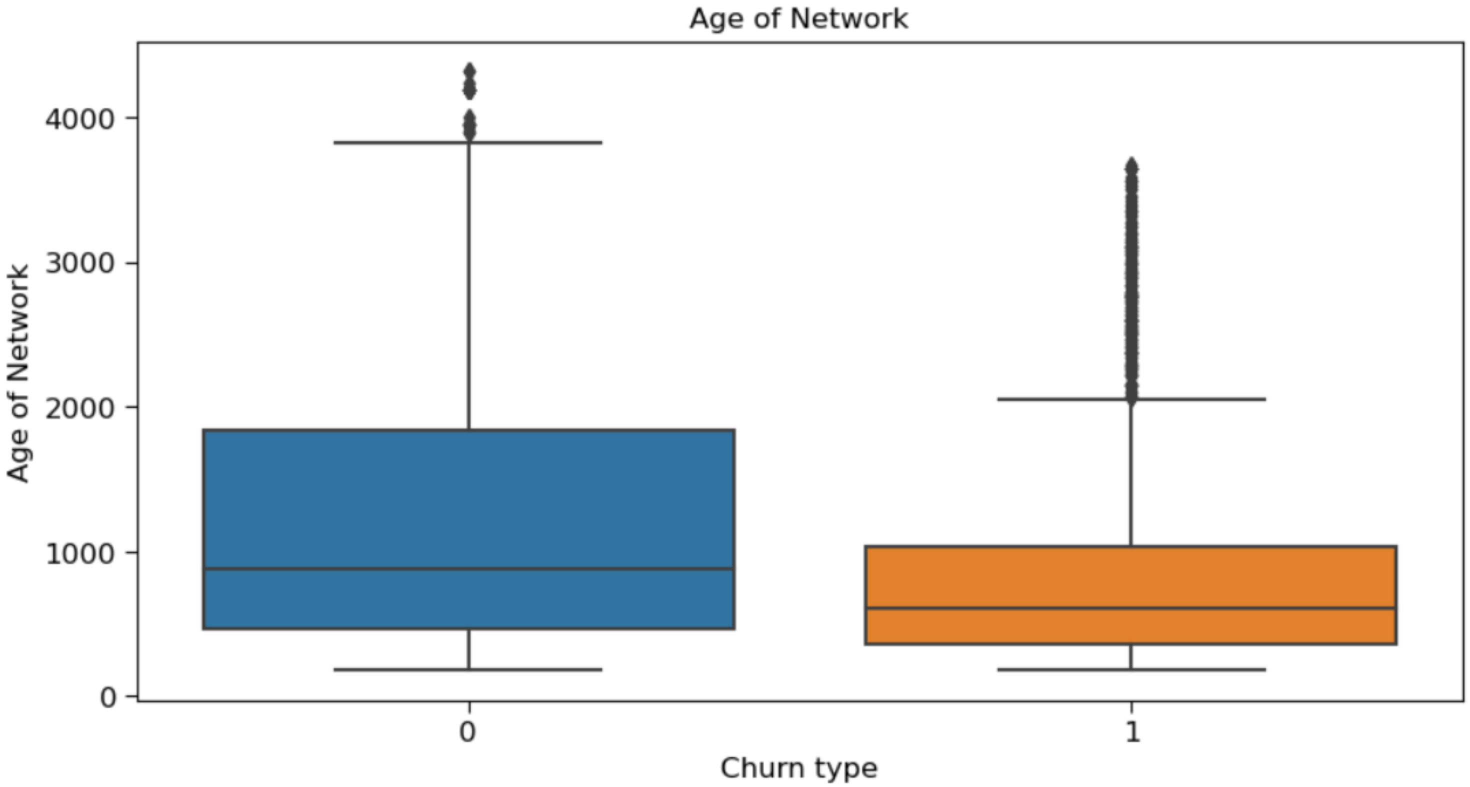
EXPLORATORY DATA ANALYSIS





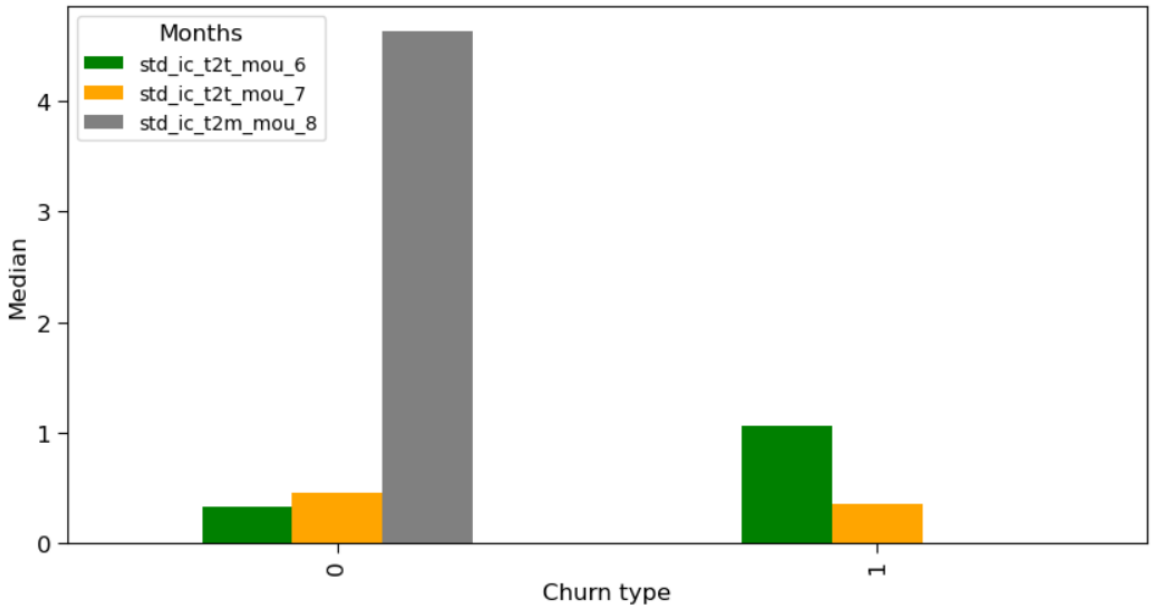
- When the average recharge amount in month 6,7,8 is none, they are more likely to churn
- Users with more difference in Total incoming minutes in month 6 & 7 are more likely to churn
- Users are likely to churn when local outgoing minutes are less
- When Total outgoing minute usage difference is more between months 6 and 7, users are expected to churn
- Users with less Local outgoing minute in same operator in months 6,7,8 are more likely to churn
- Also, users with less Local outgoing minute to other operator are expected to churn

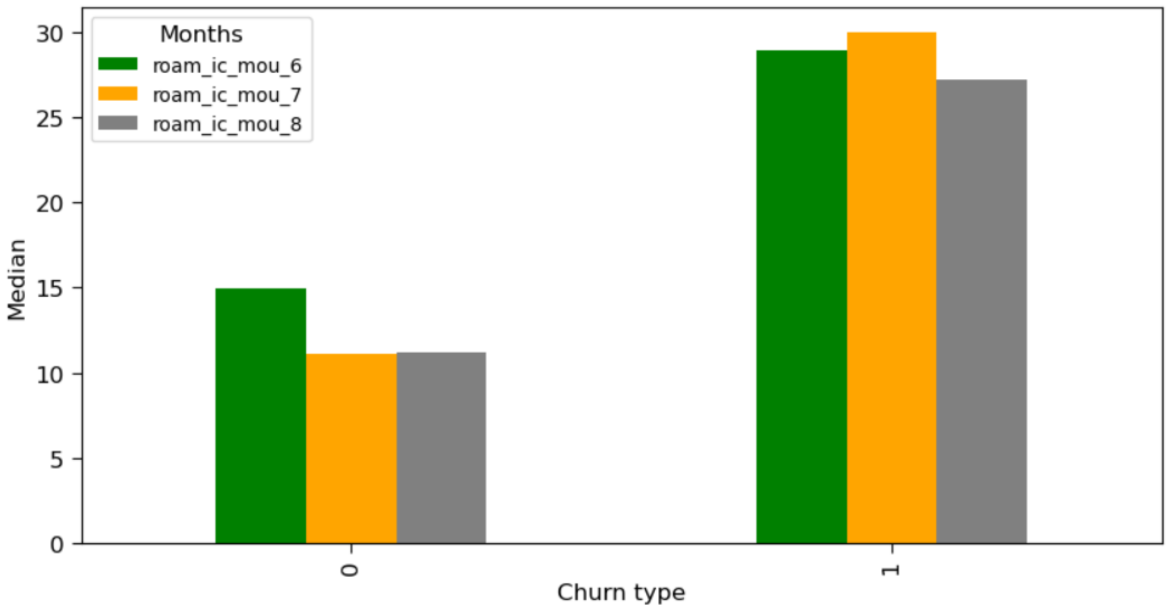
EXPLORATORY DATA ANALYSIS

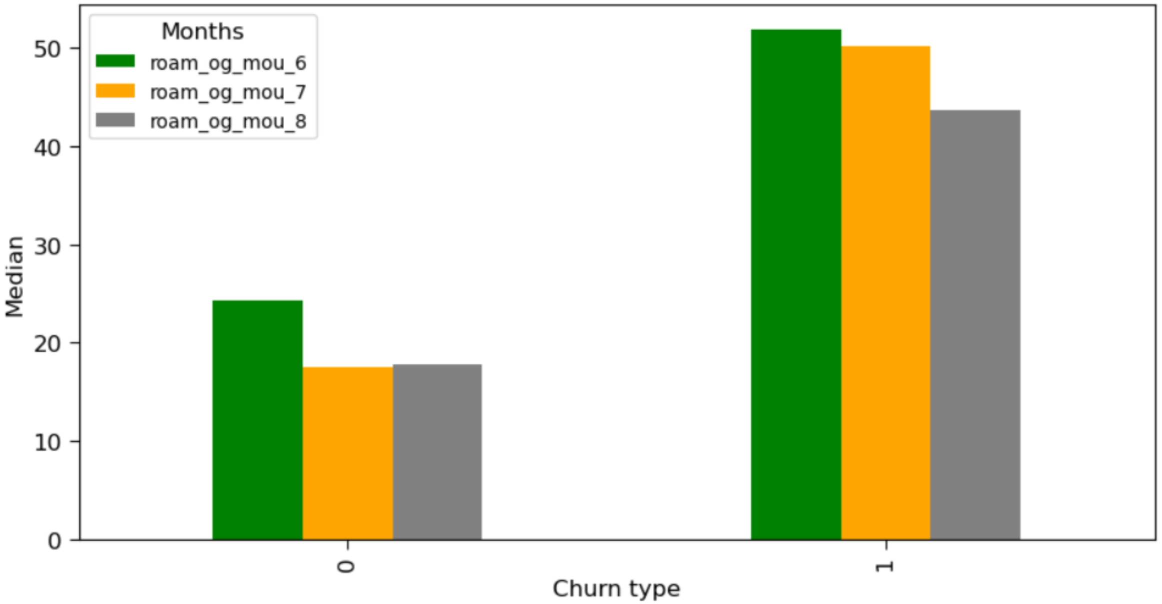


Users churning have lower Median Age of Network

EXPLORATORY DATA ANALYSIS







Below types of users are more likely to churn

- Users who are using more STD calls
- Users with more Roaming in Incoming minutes
- Users with more Roaming in Outgoing minutes

DATA PREPARATION

- The data is split into train and test set [70: 30 respectively]
- Converted categorical data to numeric columns by aggregation.
- Performed oversampling with SMOTE considering imbalance in the dataset
- Used MinMax scaler for rescaling the features
- To remove collinearity and faster training, performed dimensionality reduction technique PCA.

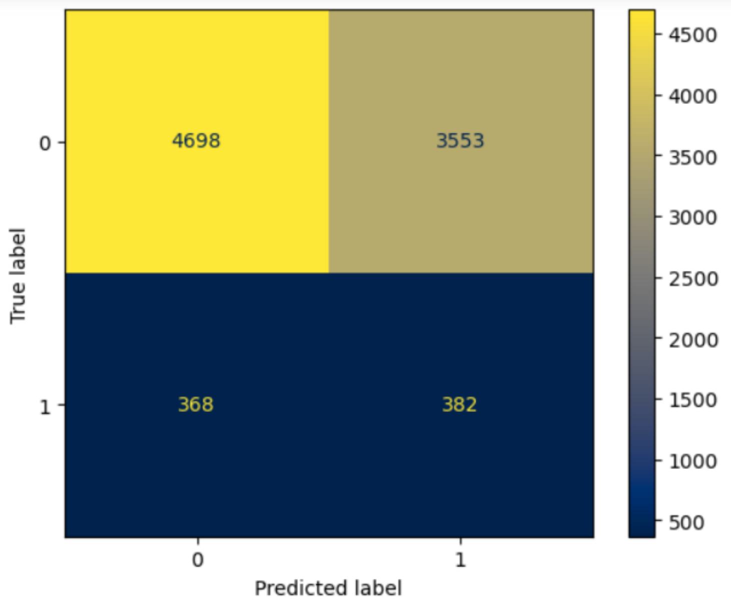
MODEL BUILDING

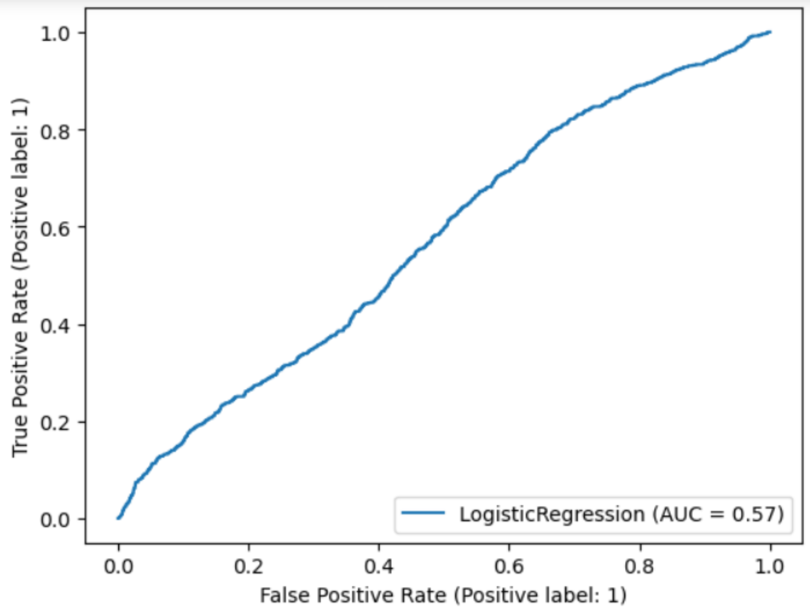
- Extensive experimentation with different models done to identify the best model that can predict the customer behavior which can help the company to take proactive steps to retain the customer.

- For model building, below models are explored:
  - Logistic regression
  - Decision tree
  - Random forest
  - Gradient boosting
  - XGboos

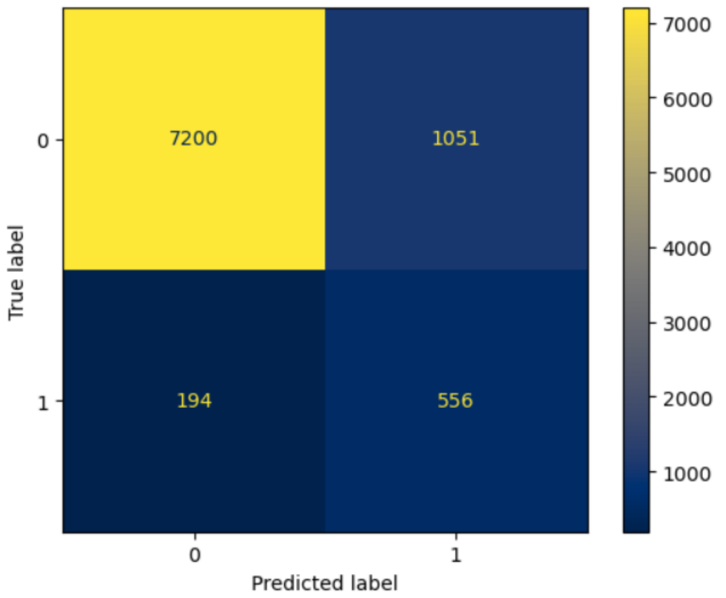
MODEL BUILDING

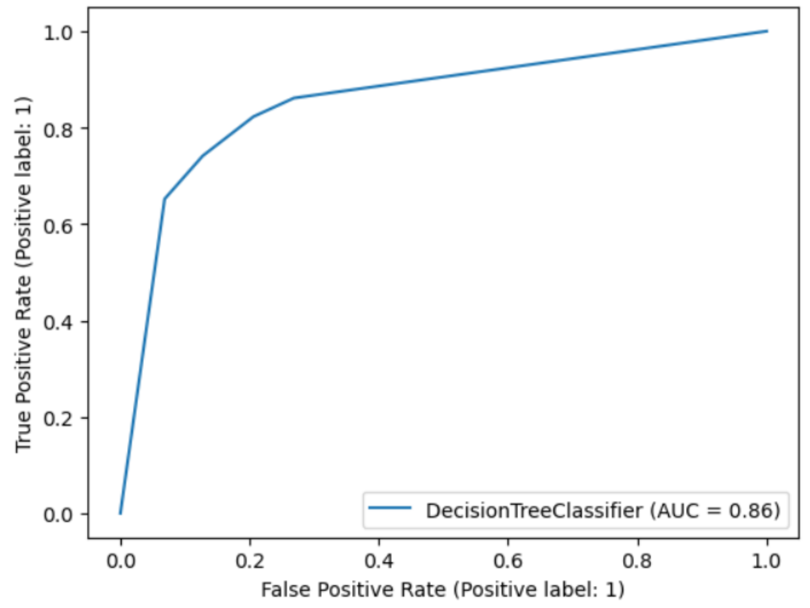
LOGISTICS REGRESSION





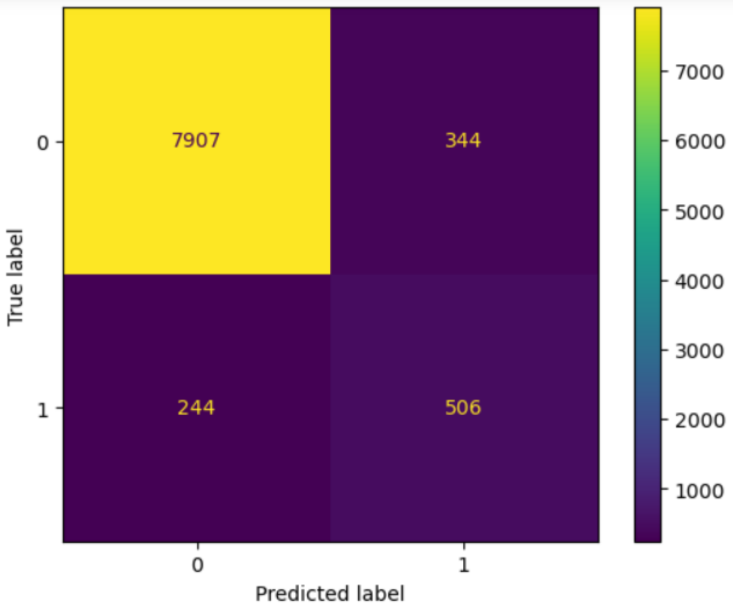
DECISION TREE

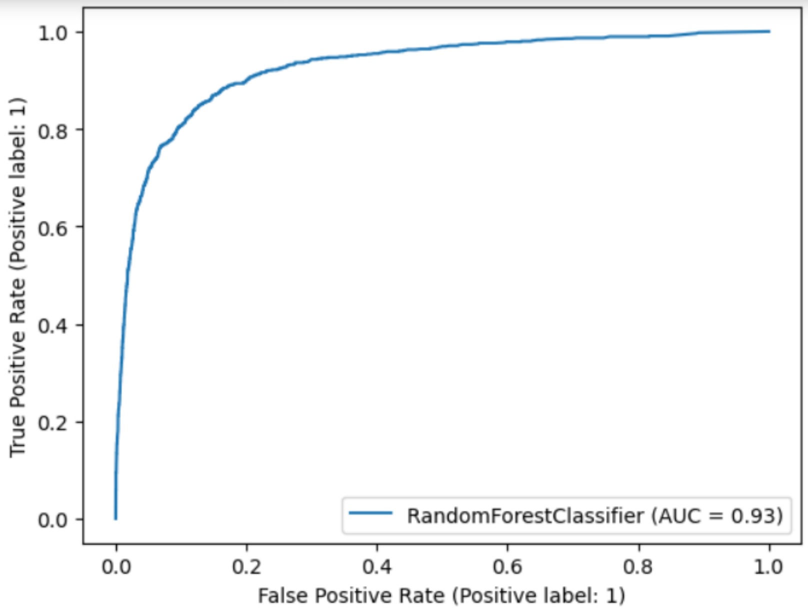




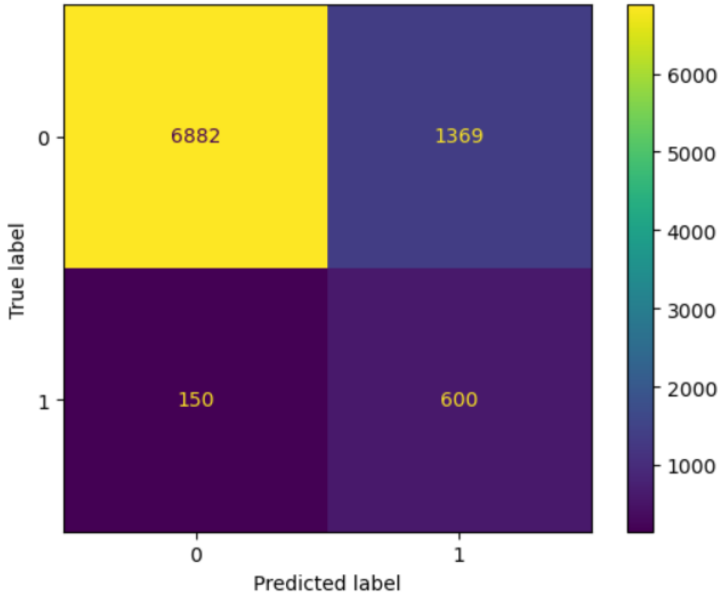
MODEL BUILDING

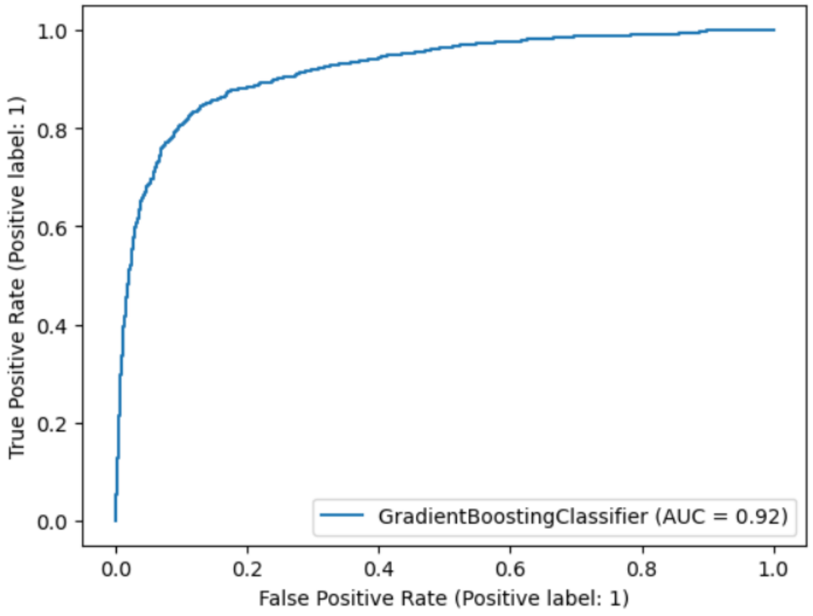
RANDOM FOREST





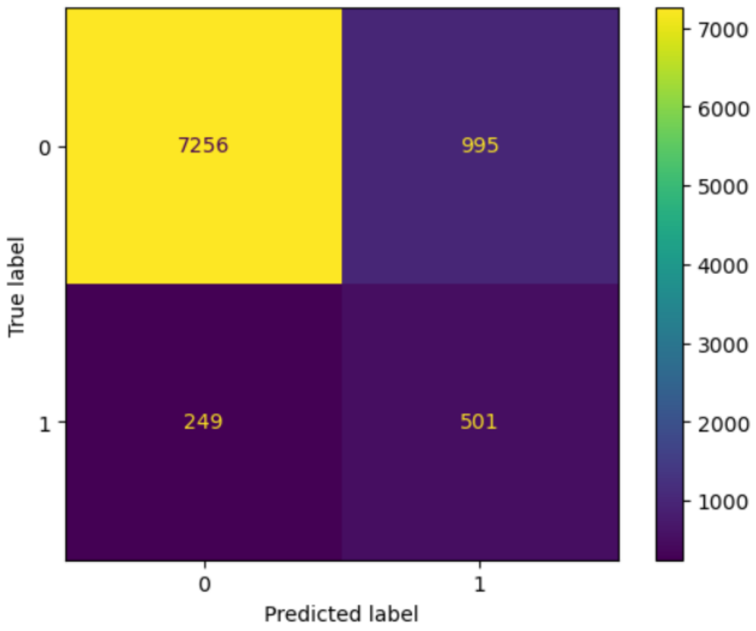
GRADIENT BOOSTING

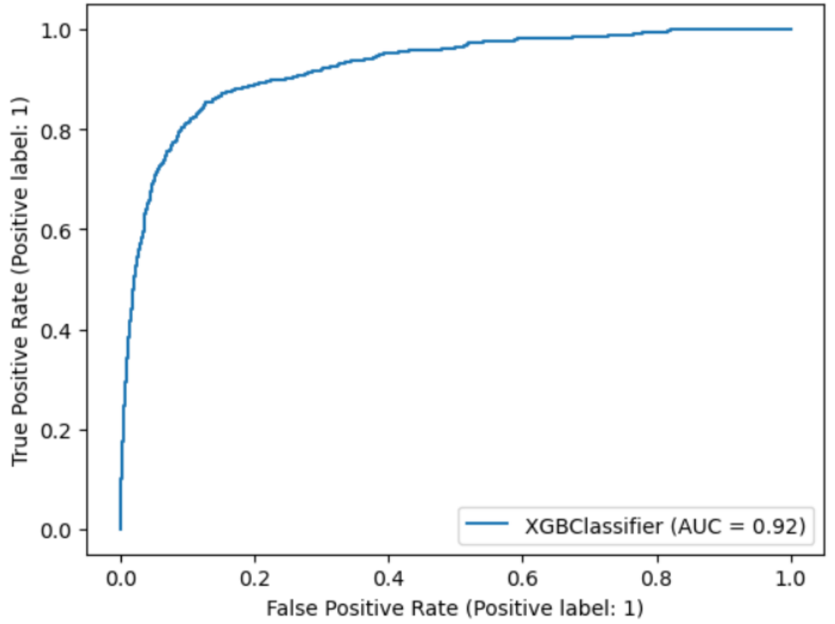




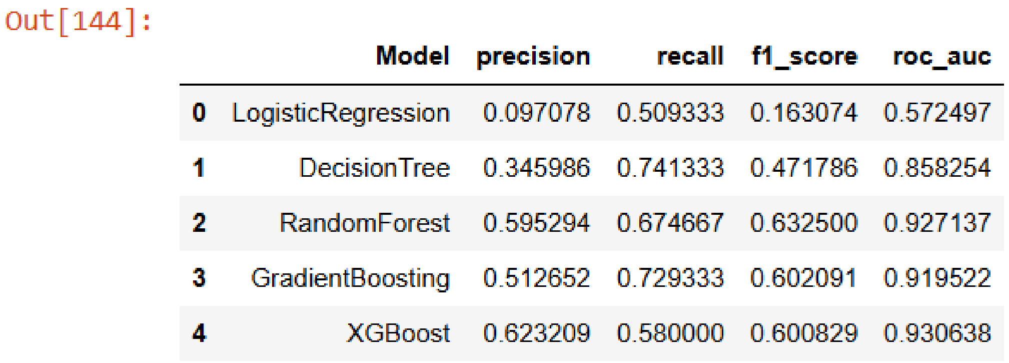
MODEL BUILDING

XGBOOST





MODEL COMPARISON



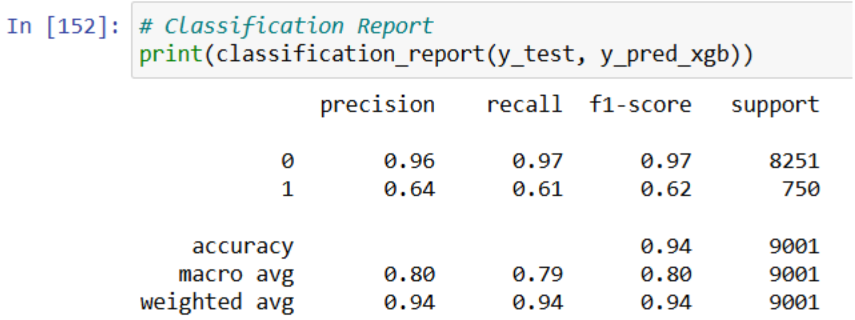
Out[144]:

	Model	precision	recall	f1_score	roc_auc
0	LogisticRegression	0.097078	0.509333	0.163074	0.572497
1	DecisionTree	0.345986	0.741333	0.471786	0.858254
2	RandomForest	0.595294	0.674667	0.632500	0.927137
3	GradientBoosting	0.512652	0.729333	0.602091	0.919522
4	XGBoost	0.623209	0.580000	0.600829	0.930638

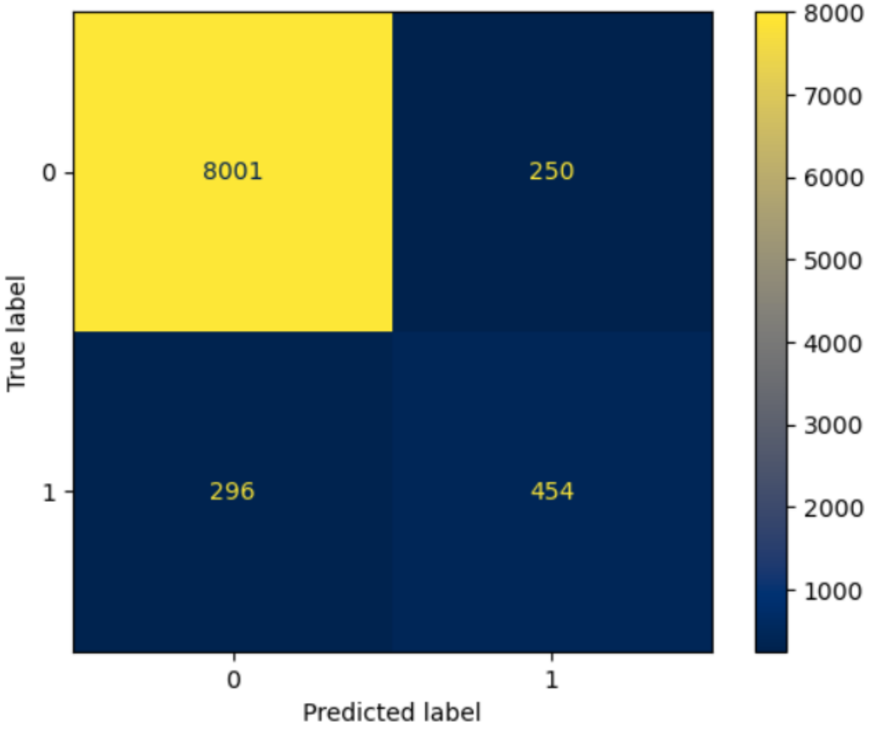
- XGBoost is the best overall, with the highest ROC_AUC and strong balance of precision & recall.
- Random Forest performs slightly better in recall & F1-score, making it good for capturing positive cases.
- Gradient Boosting is close to Random Forest, but slightly weaker in F1-score.
- Decision Tree is decent, but not as effective as ensemble models.
- Logistic Regression performs poorly due to the complexity of the dataset.

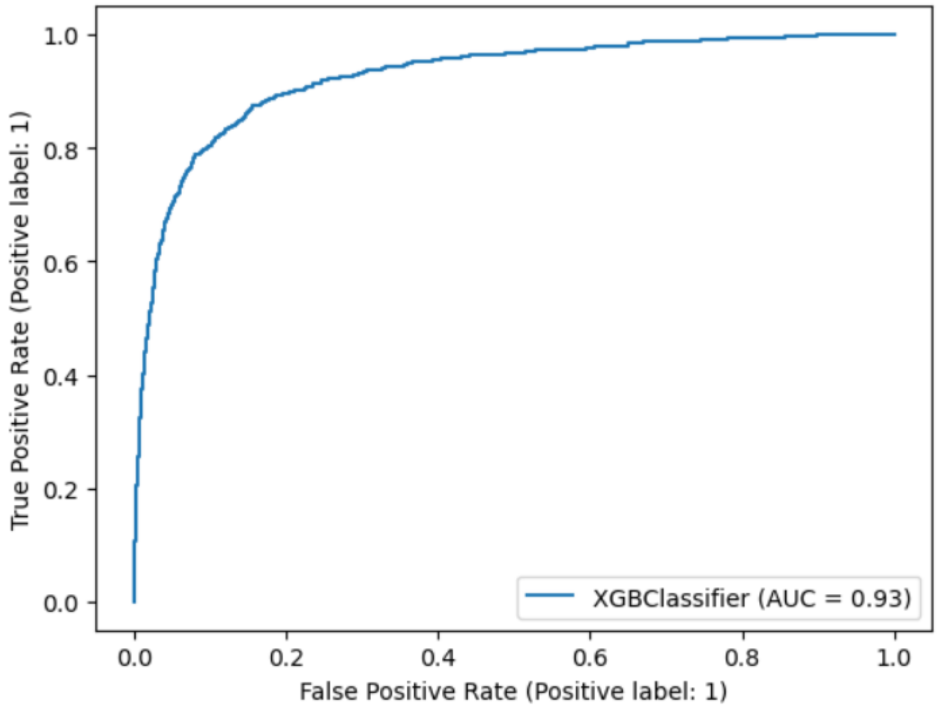
FEATURE IMPORTANCE & MODEL INTERPRETATION

- ❑ Defined the parameter grid, performed RandomisedsearchCV and found the best parameters
- ❑ Trained the XGBoost with best parameters
- ❑ Made predictions and evaluated performance
- ❑ Performed feature importance extraction to select top 40 features

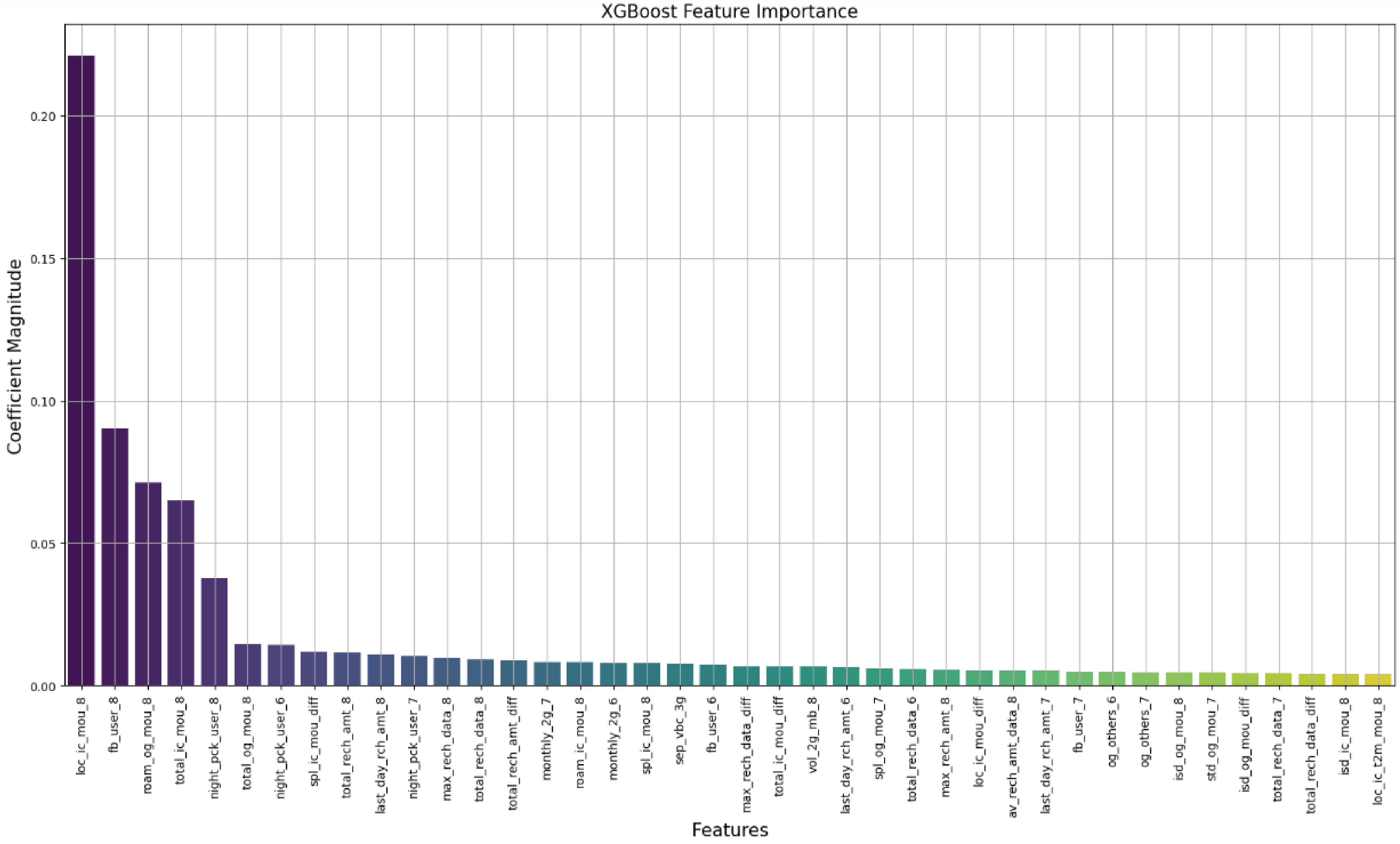


```
In [152]: # Classification Report
          print(classification_report(y_test, y_pred_xgb))

                  precision    recall  f1-score   support

               0       0.96      0.97      0.97      8251
               1       0.64      0.61      0.62       750

        accuracy                           0.94      9001
       macro avg       0.80      0.79      0.80      9001
    weighted avg       0.94      0.94      0.94      9001
```





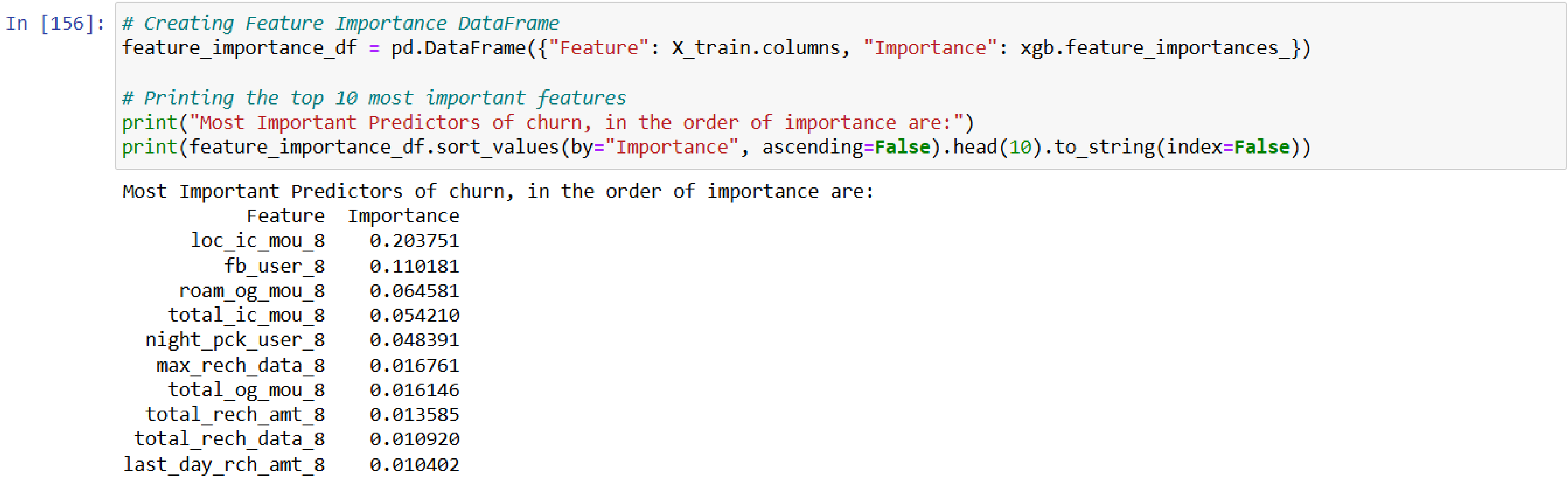
TOP 40 FEATURES



TOP 10 MOST IMPORTANT FEATURES

- The most important features are as shown in previous graph (top 40 features)

- From this, we will select only the top 10 most important features for recommendation



```
In [156]: # Creating Feature Importance DataFrame
          feature_importance_df = pd.DataFrame({"Feature": X_train.columns, "Importance": xgb.feature_importances_})

          # Printing the top 10 most important features
          print("Most Important Predictors of churn, in the order of importance are:")
          print(feature_importance_df.sort_values(by="Importance", ascending=False).head(10).to_string(index=False))
```

```
Most Important Predictors of churn, in the order of importance are:
           Feature  Importance
      loc_ic_mou_8    0.203751
         fb_user_8    0.110181
     roam_og_mou_8    0.064581
    total_ic_mou_8    0.054210
  night_pck_user_8    0.048391
   max_rech_data_8    0.016761
    total_og_mou_8    0.016146
  total_rech_amt_8    0.013585
 total_rech_data_8    0.010920
last_day_rch_amt_8    0.010402
```

CONCLUSION

1) Customers with a high volume of local incoming calls, (loc_ic_mou_8 = 0.203751), may be avoiding outgoing calls due to high costs. If competitors offer better call rates, these users are more likely to churn.

2) Heavy data users (fb_user_8 = 0.110181), are highly sensitive to data speed and pricing. If they experience slow internet or find a better data plan elsewhere, they are at a higher risk of leaving.

3) Customers who make frequent roaming outgoing calls (roam_og_mou_8 = 0.064581) may switch if roaming charges are high or if they find another provider with better roaming benefits.

4) A high volume of incoming calls (total_ic_mou_8 = 0.054210) suggests that users rely more on receiving calls rather than making them, possibly due to expensive outgoing call rates. These users may churn if they find better offers.

5) Customers subscribed to night data packs (night_pck_user_8 = 0.048391) are primarily data consumers. If the network performance is poor or night packs are not competitive, they are more likely to switch to another provider.

CONCLUSION (CONTD..)

6) Customers who frequently recharge large amounts of data (max_rech_data_8 = 0.016761) expect fast and reliable internet. If they face slow speeds or expensive data plans, they might churn.

7) High outgoing call users (total_og_mou_8 = 0.016146) are at risk if they find better voice plans elsewhere. Cost-conscious customers may migrate to competitors offering unlimited calling.

8) Customers who recharge with higher amounts (total_rech_amt_8 = 0.013585) are valuable but also expect premium services. If they perceive a lack of value, they may explore other options.

9) Heavy data users (total_rech_data_8 = 0.010920) are more likely to churn if data caps, slow speeds, or expensive pricing become a problem.

10) Customers who recharge at the last moment (last_day_rch_amt_8 = 0.010402) might be hesitant to commit and could be considering other network options before deciding to continue.

KEY TAKEAWAYS

- High data and call users are at greater risk of churn if they find better plans or experience poor network quality.

- Roaming and night pack users are price-sensitive and may leave if they perceive their current plans as expensive or restrictive.

- High-recharge customers expect premium services and are more likely to churn if they feel they are not getting good value for their money.

- Facebook users and heavy data consumers are a crucial segment to retain by offering better data plans and strong network performance.

RECOMMENDATIONS

Based on the strongest indicators of churn, the following actions are recommended:

- ❑ Target users with significantly lower local incoming call usage:

Customers who have 20.3% lower local incoming call minutes compared to the average are the most likely to churn. Implement engagement campaigns or discounted call plans for these users to retain them.

- ❑ Focus on Facebook users:

Customers who actively use Facebook (11%) are at a higher risk of churning. Offering social media data bundles or exclusive content could help retain this segment.

- ❑ Identify roaming users with lower outgoing call minutes:

Users with 6.5% lower roaming outgoing call minutes are likely to churn. Consider offering travel-friendly roaming packages to keep these users engaged.

RECOMMENDATIONS (CONTD..)

❑ Monitor total incoming call usage:

A 5.4% drop in total incoming call minutes is linked to higher churn rates. Encouraging inbound call activity through better call incentives or loyalty programs can help retain these users.

❑ Prioritize night pack users with declining engagement:

Customers subscribed to night packs (4.8%) but showing lower usage should be targeted with personalized retention offers or bonuses to maintain engagement.

THANK YOU