# Domain Oriented Case Study: Telecom Churn

Divyansh Sharma

# Business Objective

In this project, we will 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.

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.

We assume that there are three phases of the 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.
- The '**churn**' phase: In this phase, the customer is said to have churned. We define churn based on this phase.

# Data Preparation

The following crucial steps were performed as part of preparing the data for our analysis:

1) **Filtering High Value Customers:**

   The objective is to predict churn for High Value customers only, i.e. 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

2) **Tag Churners and create the target variable:**

   The churned customers (churn=1, else 0) are tagged based on their usage in the fourth month. Those who have not made any calls (either incoming or outgoing) AND have not used mobile internet even once in the churn phase are tagged as churners .

The attributes used to tag churners are:

- total_ic_mou_9
- total_og_mou_9
- vol_2g_mb_9
- vol_3g_mb_9

# Dropping columns

Columns that fit the following criteria were dropped from the dataset:

1) All columns for month of September as mentioned in the business objective.

2) All columns having more than 30% null values.

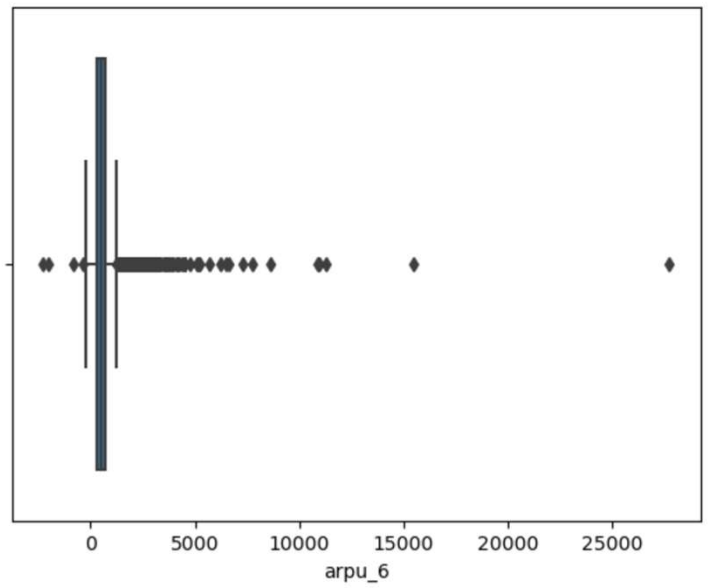
3) All columns having only 1 unique value as they cannot provide any insight in our analysis.

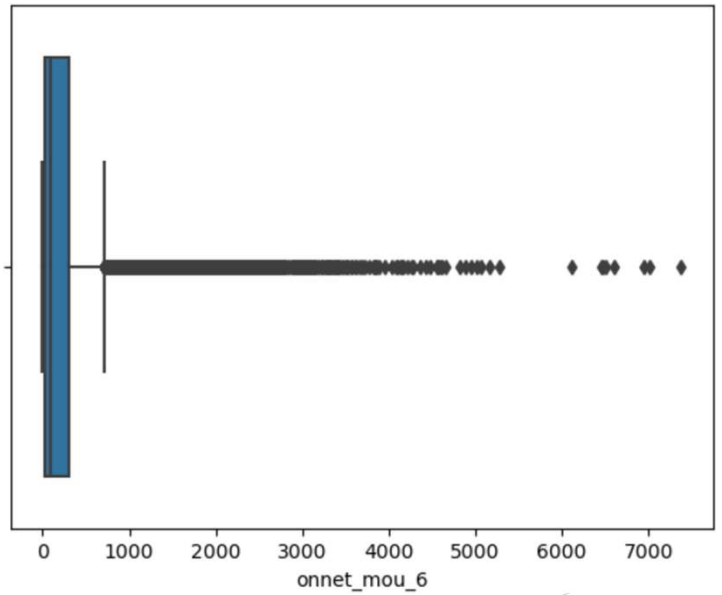
4) All date columns as we would need only numeric columns for our analysis.

5) Some columns had less than 4% missing values. We chose to delete all such rows as it would not have much impact on the number of records.

# Outlier Treatment

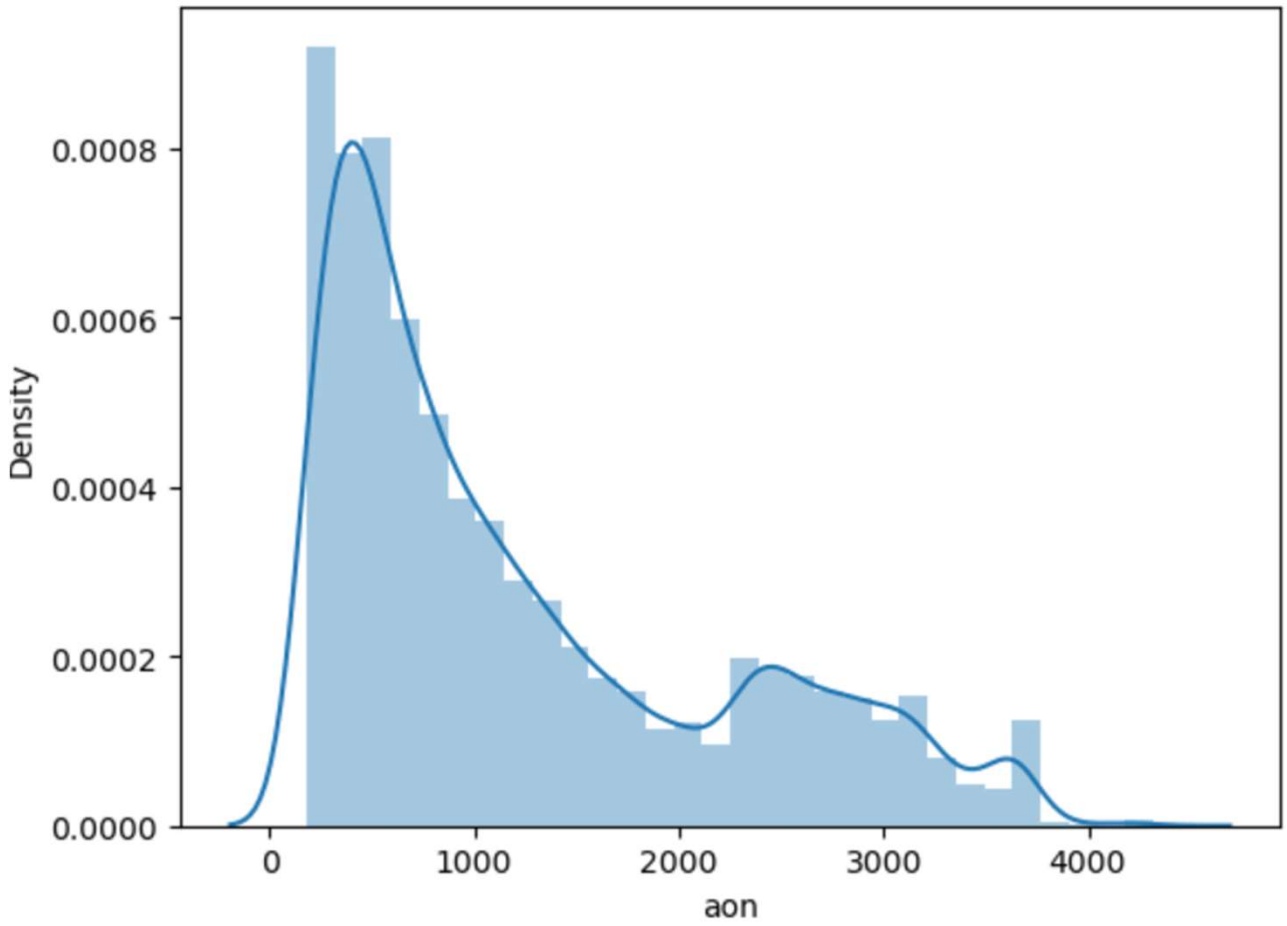
- We could see a lot of variance between 99 percentile and 100 percentile values for most columns.
- Plotting the BOX plot for some columns helped us affirm the same.
- These outliers were handled by increasing the IQR range by 15 percentile on both sides.



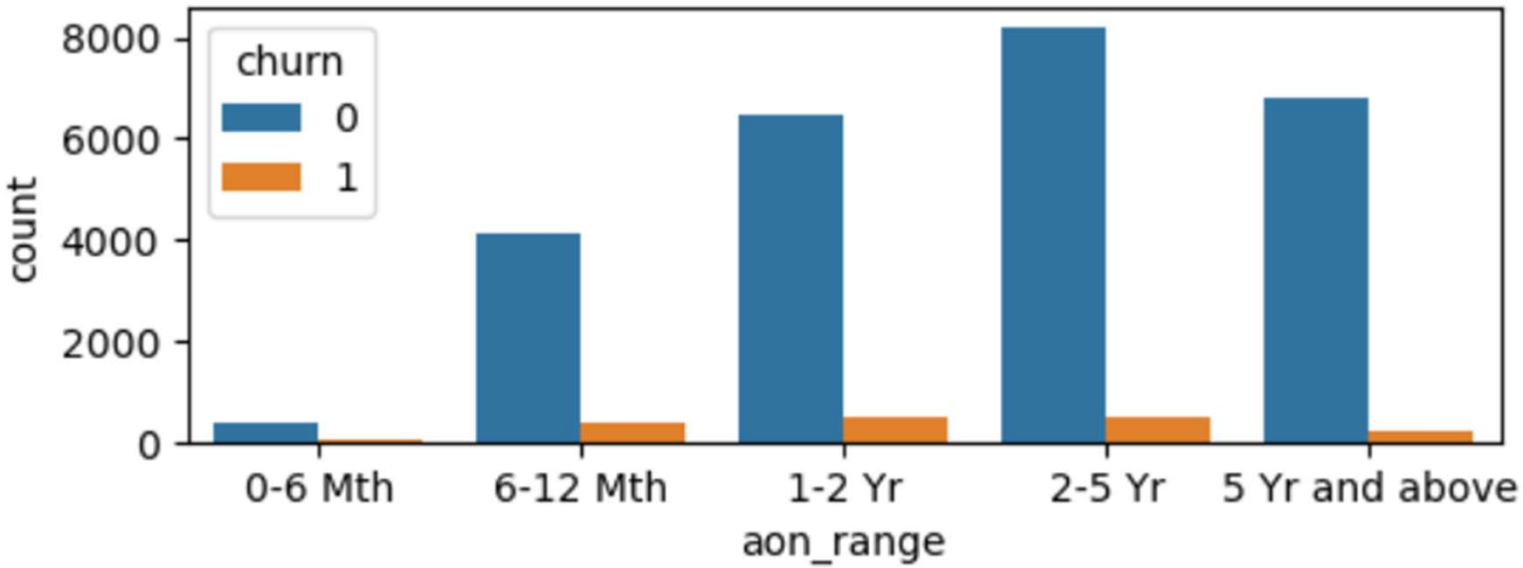
# Exploratory Data Analysis (EDA)

Visualizing the distribution of the *Age on network* values showed us skewed data, which shows that people tend to churn after using the telecom services for a few years.



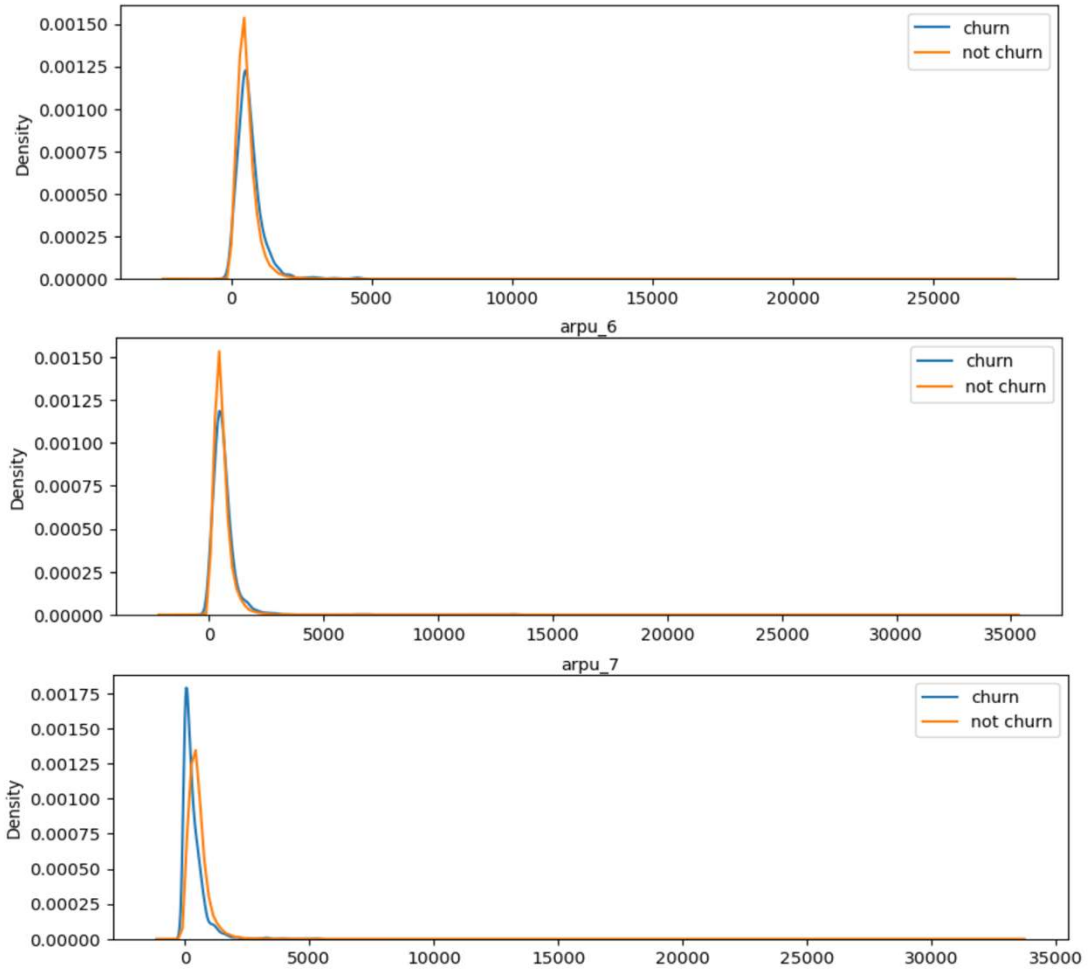
# Exploratory Data Analysis (EDA)

Creating bins for *Age on network* values for a 6 month span, showed us that customers using the telecom services for over a year tend to churn



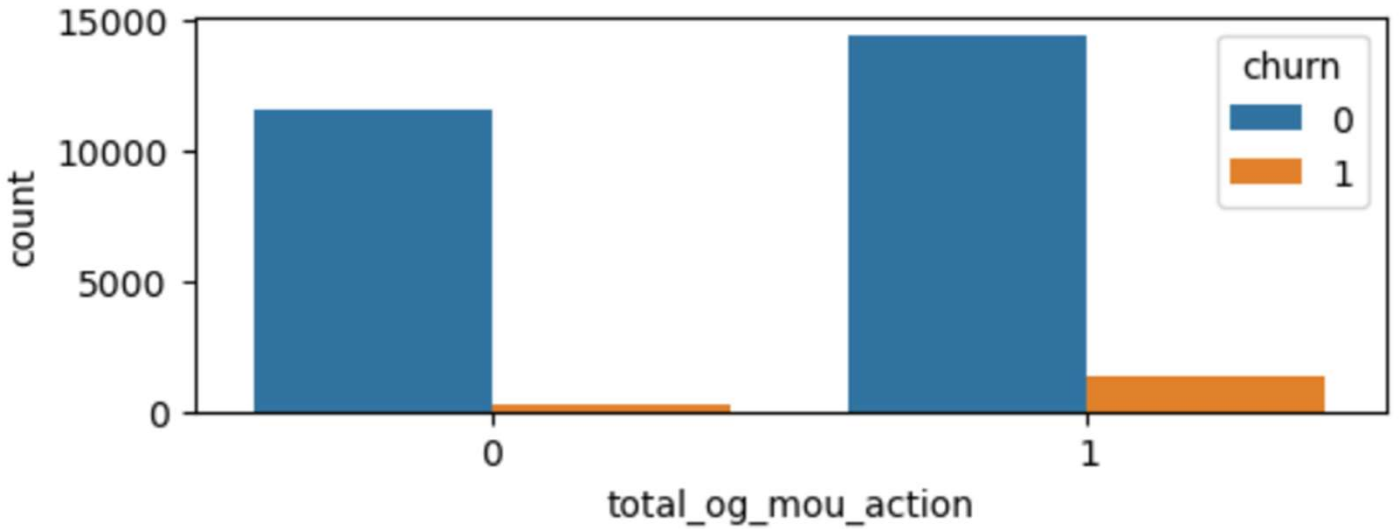
# EDA: Average Revenue Per User (arpu)

A decrease in Average revenue per user results in a higher churn rate



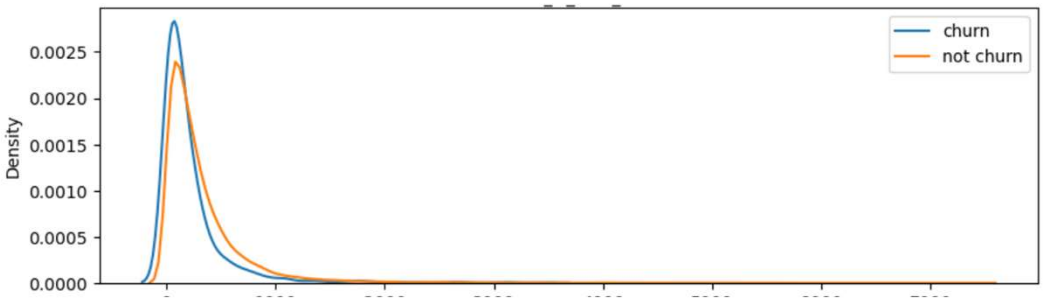
# EDA : Total Outgoing Calls Minutes Of Use (total_og_mou)

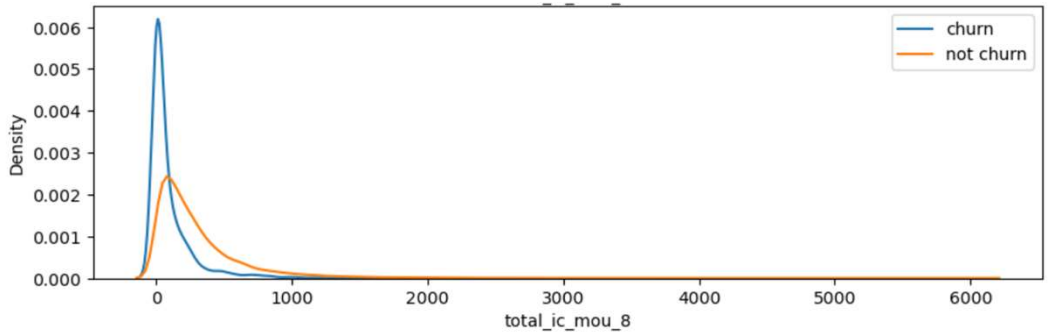
A decrease in Total outgoing per user in the action phase results in a higher churn rate

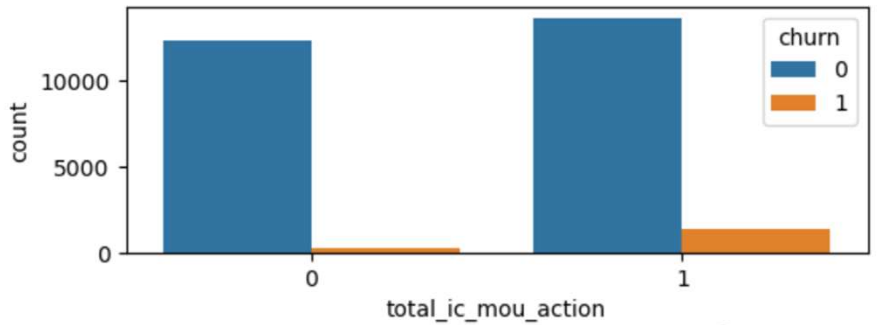


# EDA: Total Incoming Calls Minutes Of Use (total_ic_mou)

A decrease in Total incoming per user in the action phase results in a higher churn rate

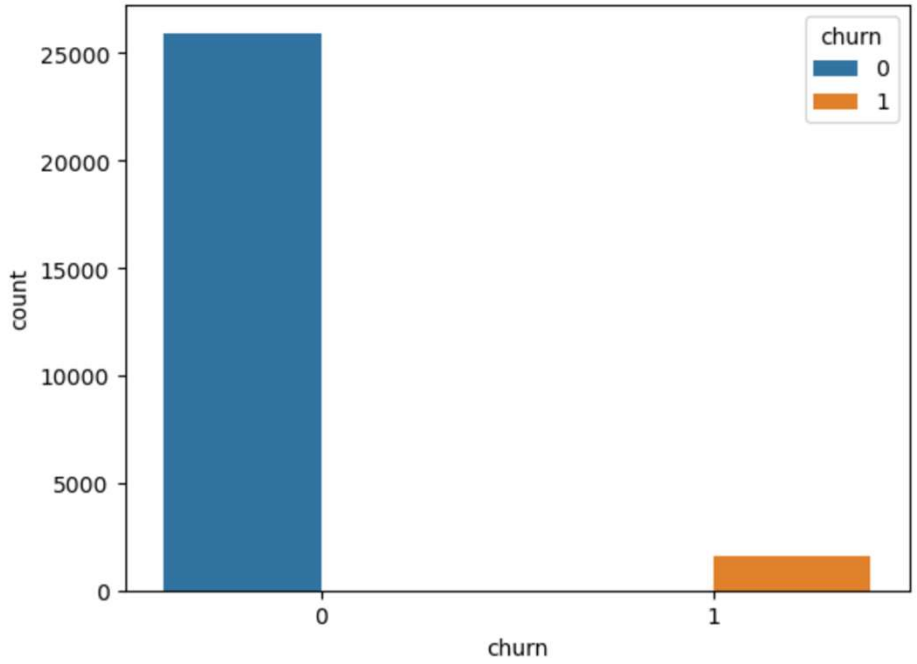


# Handling Class Imbalance

Plotting the bar graph to visualize the Churn variable showed a heavy imbalance of data




To account for class imbalance, Synthetic Minority Class Oversampling Technique (SMOTE) was used.

# Model Building

- To predict the churn, we have used Random Forest classifier as it can handle the collinearity and give better results.

- After hyper parameter tuning, we concluded that an Accuracy of 90.48% can be achieved with the following params:

    - max_depth = 7,
    - max_features = 15,
    - min_samples_leaf = 20,
    - n_estimators = 300

# Model Building

Making predictions on the train set and test set, we calculated the following metrics:

- Accuracy
- Sensitivity
- Specificity

The values were found to be quite similar, which suggests that the model built was good.

| | Score of train data in % | Score of test data in % |
|---|---|---|
| **Accuracy** | 91.37 | 88.87 |
| **Sensitivity** | 91.42 | 66.95 |
| **Specificity** | 91.33 | 90.20 |

# Feature Importance

Fitting the model in the Random Forest classifier helped us get the importance of the features in the data set.

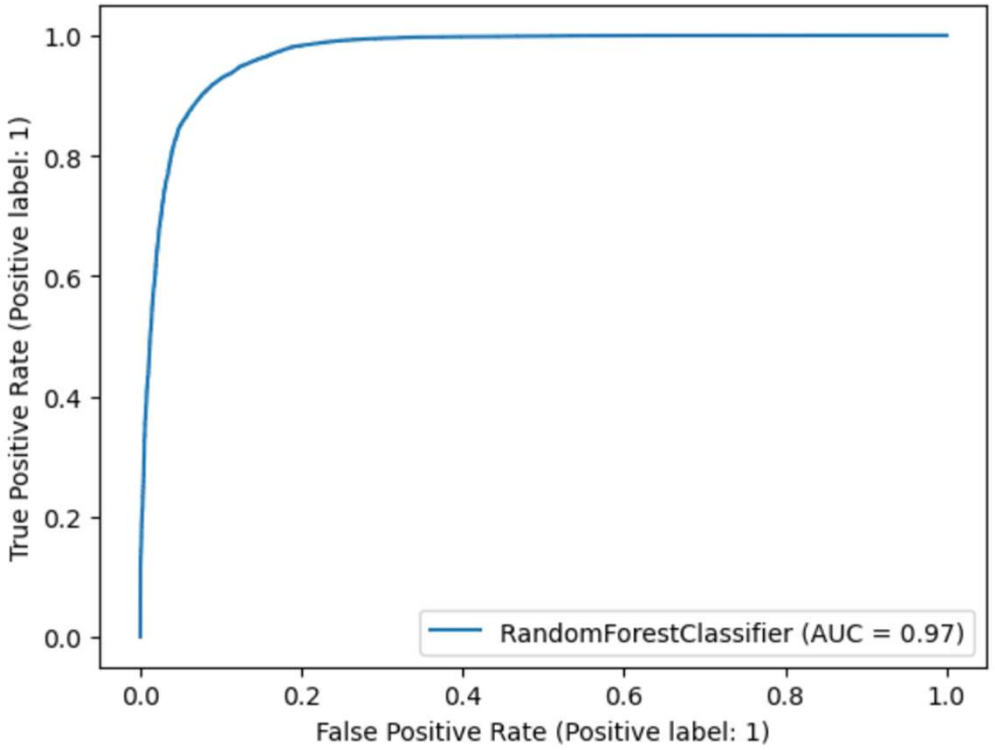
The top 10 features according to our model are as follows:

| | var_name | Importance |
|---|---|---|
| **14** | roam_og_mou_8 | 0.170950 |
| **11** | roam_ic_mou_8 | 0.139397 |
| **80** | total_ic_mou_8 | 0.064037 |
| **65** | loc_ic_mou_8 | 0.049256 |
| **59** | loc_ic_t2m_mou_8 | 0.040771 |
| **101** | last_day_rch_amt_8 | 0.038835 |
| **95** | total_rech_amt_8 | 0.037238 |
| **126** | total_rech_amt_data_8 | 0.035361 |
| **29** | loc_og_mou_8 | 0.031664 |
| **20** | loc_og_t2m_mou_8 | 0.028691 |

# ROC Curve

The ROC curve provides valuable insights into the trade-off between true positive rate and false positive rate for different threshold settings.

The Area Under Curve (AUC) shows a healthy 97% for our model.



# Inference

1. The feature importance shows that in the action phase, the roaming incoming and outgoing minutes of usage are very important factors that the telecom company should be aware of.

2. Some attractive roaming plans can help lure customers in staying with the company for longer duration.

3. Along with roaming, local and total incoming minutes are also required to be tracked by the company.

4. Recharge amount is another important factor and some attractive recharge offers can help reduce the churn.

Specificity is the proportion of True Negatives which is 90.20%, hence model is predating that the 90.20% of the customer will not churn.

Hence we will eliminate these 90.20% customer and concentrate on 9.80% customer who will likely to churn.

# Thank You