Capstone Project: Telecom Churn

Prepared by *Dimple Soni*

AGENDA :

- Introduction
- Problem Definition
- Data Overview
- Data Analysis
- Model Development
- Results
- Business Recommendations

Introduction :

- **Objective:** 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 more important than customer acquisition.

For many incumbent operators, *retaining highly profitable customers is the number one business goal*.

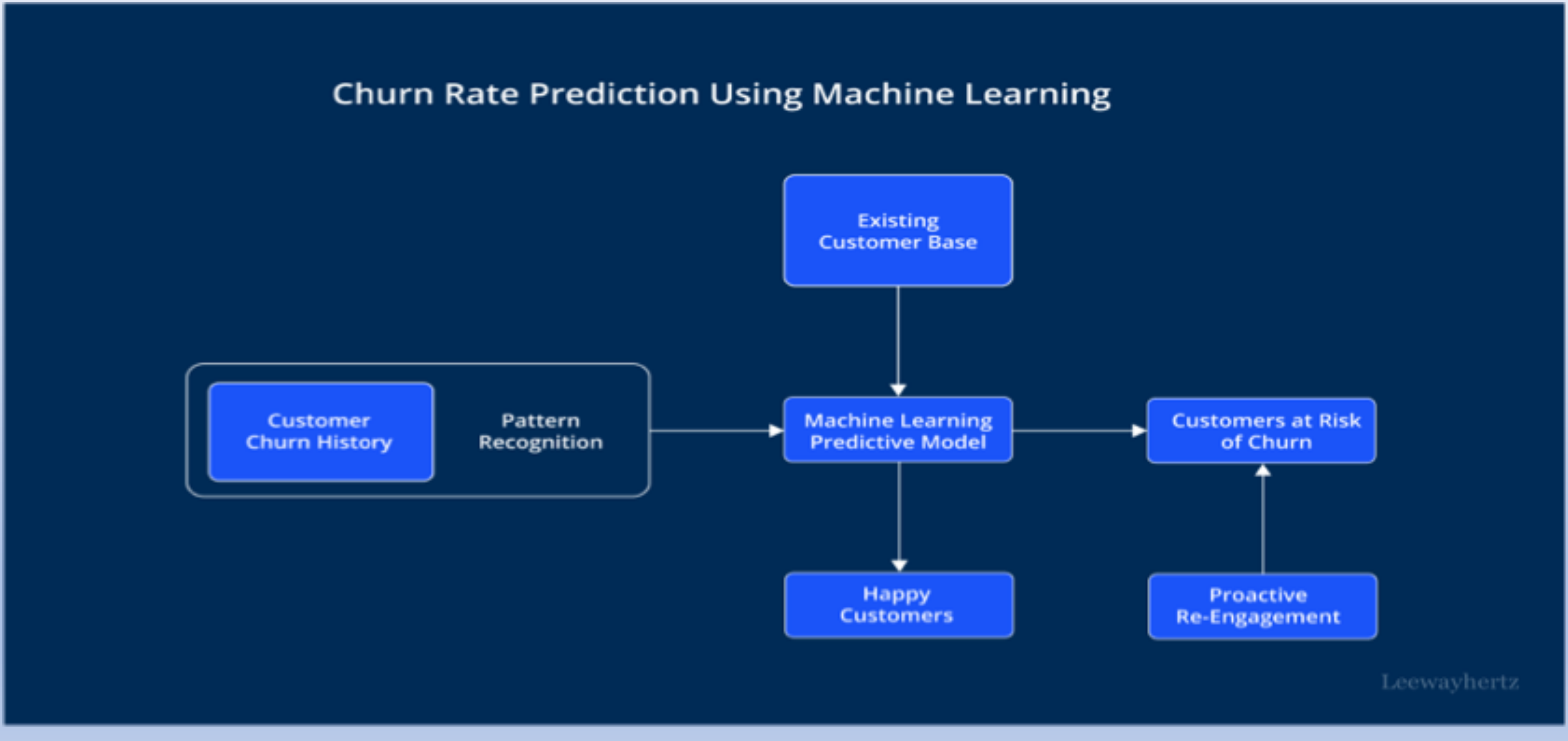
To reduce customer churn, telecom companies need to **predict which highly profitable customers are at risk of churn.**

Problem Definition :

- Churn: Churn is a problem for telecom companies because it is more expensive to acquire a new customer than to keep your existing one from leaving.

- Churn Prediction: It is the process of identifying customers who are likely to stop using a company’s products or services, based on historical data and behavioral patterns. The main goal of churn prediction is to **identify at-risk customers** in advance, allowing the business to take proactive steps to retain them.

Churn Prediction Model



Handling Missing Values

Handling missing values is a crucial step in data preprocessing, as it can significantly impact the performance of machine learning models. Different techniques are used to handle missing values based on the nature of the data and the proportion of missingness.

Common Methods for Handling Missing Values:

- Dropping Missing Values
- Mean/Median/Mode Imputation

Exploratory Data Analysis (EDA)

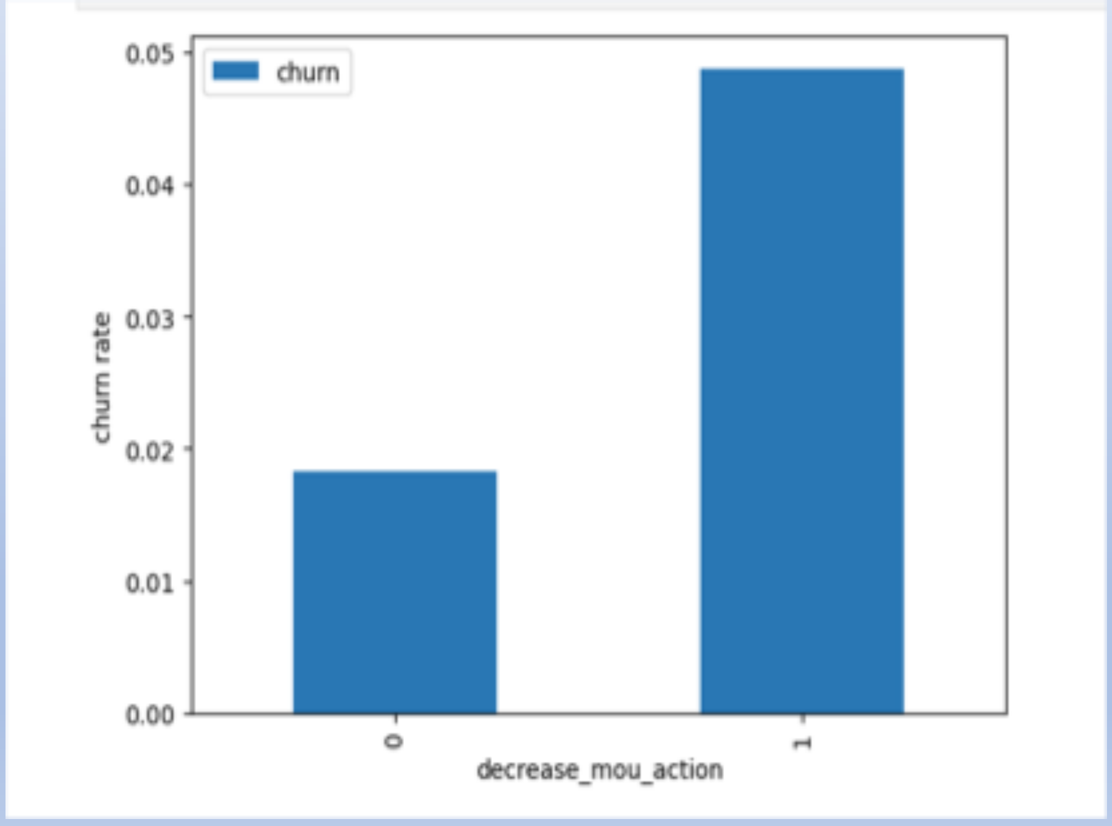
- Data Visualisation using seaborn and metplotlib.

- Data Analysis (EDA) is an approach to analyse datasets and to summarize their main characteristics , often with visual methods.

- A Statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis.

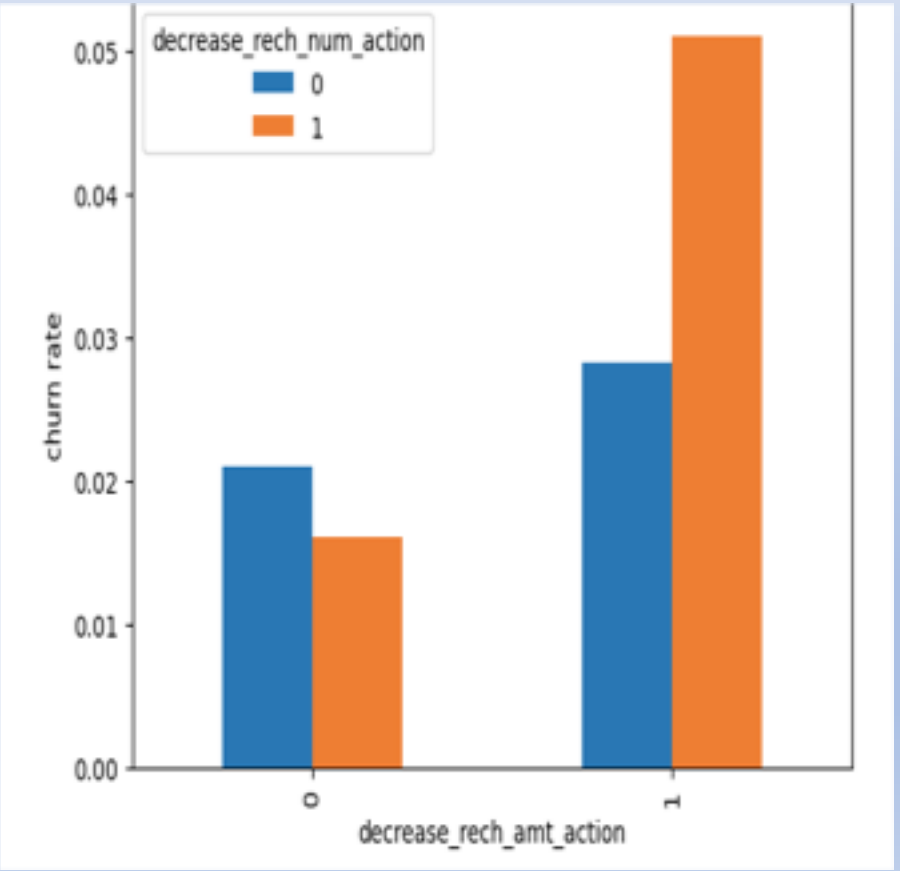
Univariate Analysis

- Univariate Analysis : We can see that the churn rate is more for the customers, whose minutes of usage(mou) decreased in the action phase than the good phase.



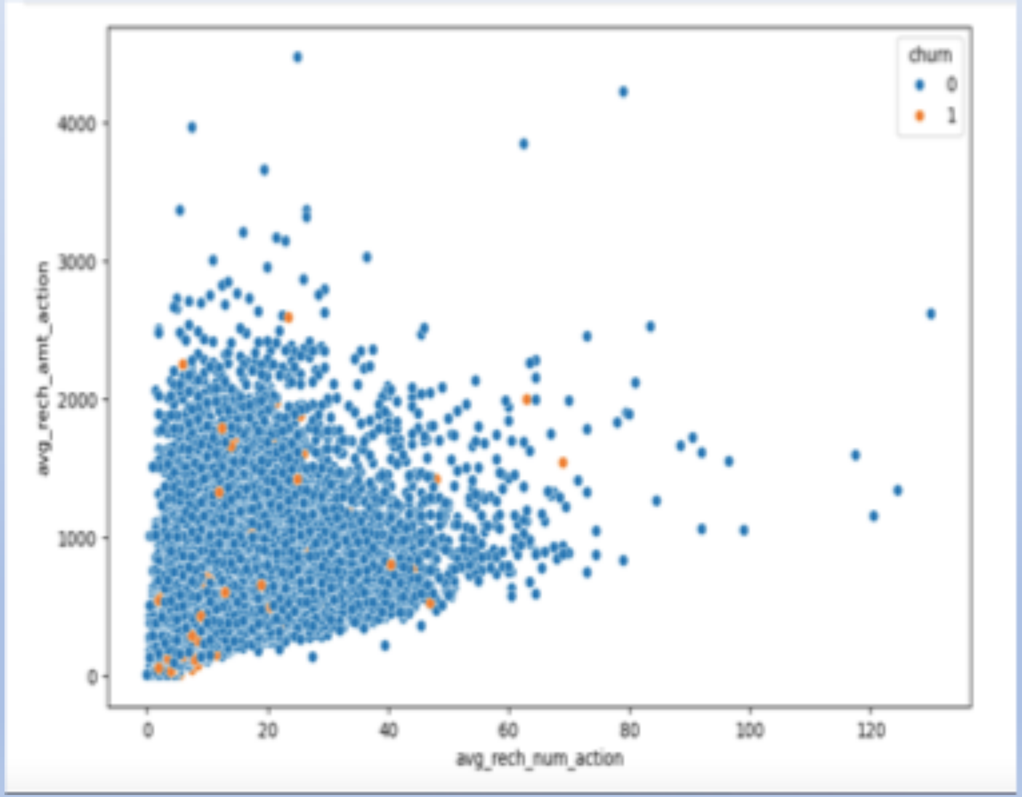
Bivariate Analysis

Bivariate Analysis : We can see from the above plot, that the churn rate is more for the customers, whose recharge amount as well as number of recharge have decreased in the action phase than the good phase.



Scatter Plot

Scatter Plot : We can see from the above pattern that the recharge number and the recharge amount are mostly propotional. More the number of recharge, more the amount of the recharge.



Feature Scaling

Feature scaling is the process of transforming the features in a dataset so that they have comparable ranges or units. This is especially important for machine learning algorithms that are sensitive to the scale of the data, such as:

- Logistic Regression
- Gradient Descent-based models like Linear Regression and Neural Networks
- Principal Component Analysis (PCA)

Training and Testing Data

Training Data: Training data is the dataset used to train the model and features. The model uses this data to adjust its parameters to minimize the prediction error.

Objective: The goal is to fit a model that captures the underlying structure of the data and generalizes well to unseen data.

Key Points:

- The model learns from training data by adjusting weights and parameters to minimize error.
- Overfitting (where the model learns the noise in the data) can occur if the model performs too well on training data but fails to generalize to unseen data.

Testing Data: Testing data is a separate portion of the dataset that the model has never seen during the training phase. It is used to evaluate the performance of the trained model to understand how well it generalizes to new, unseen data.

Objective: The goal is to assess the model's performance on new data, ensuring it hasn't overfitted the training data and can perform well in real-world scenarios.

Key Points:

- The testing data should not be used during training, as this can lead to overestimation of the model’s performance.
- Metrics like accuracy, precision, recall, F1 score, or mean squared error (depending on the problem) are used to evaluate the model on the test data.

Modelling

Modelling

- Build models to predict churn. The predictive model that we are going to build will serve two purposes:
- It will be used to predict whether a high-value customer will churn or not, in near future (i.e. churn phase). By knowing this, the company can take action steps such as providing special plans, discounts on recharge etc.
- It will be used to identify important variables that are strong predictors of churn. These variables may also indicate why customers choose to switch to other networks.

In some cases, both of the above-stated goals can be achieved by a single machine learning model. But here, we have a large number of attributes, and thus we should try using a dimensionality reduction technique such as PCA and then build a predictive model. After PCA, we can use any classification model.

Model Development

Models performed in the project are:

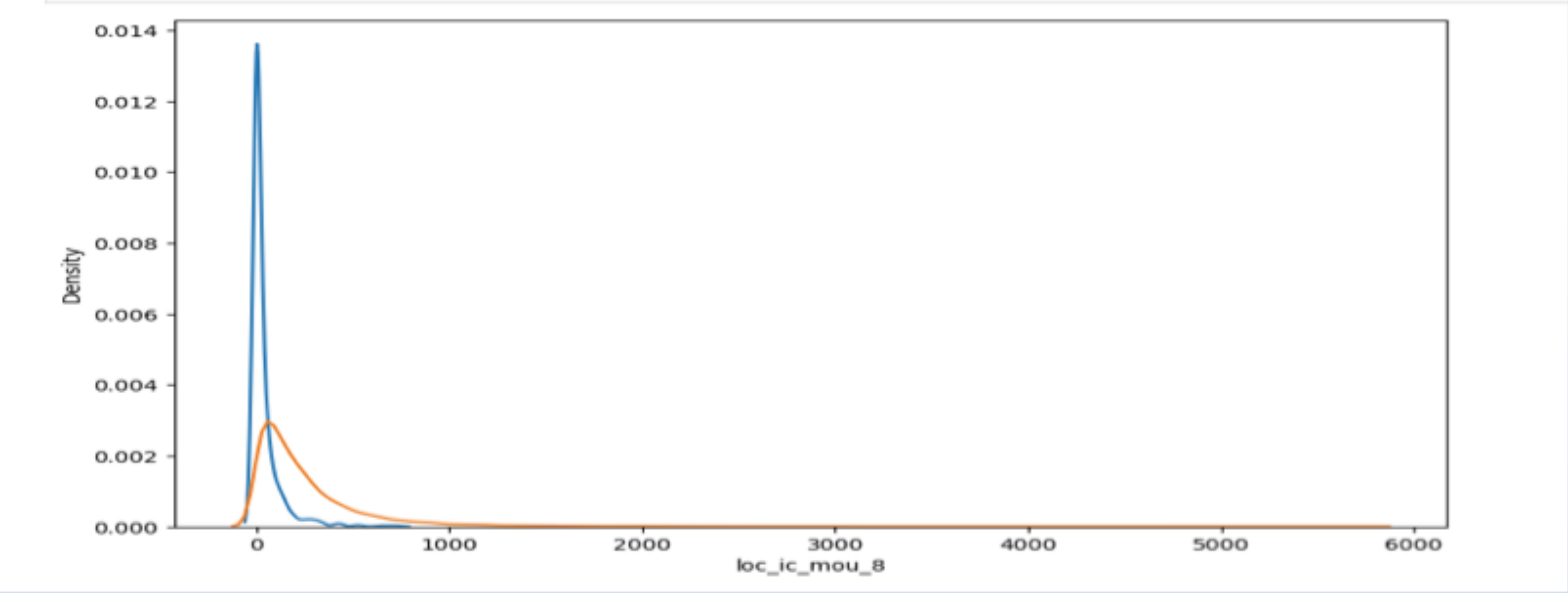
- Principle component Analysis (PCA)
- Logistic Regression
- Decision Tree
- Random Forest Classifier

Model Comparison

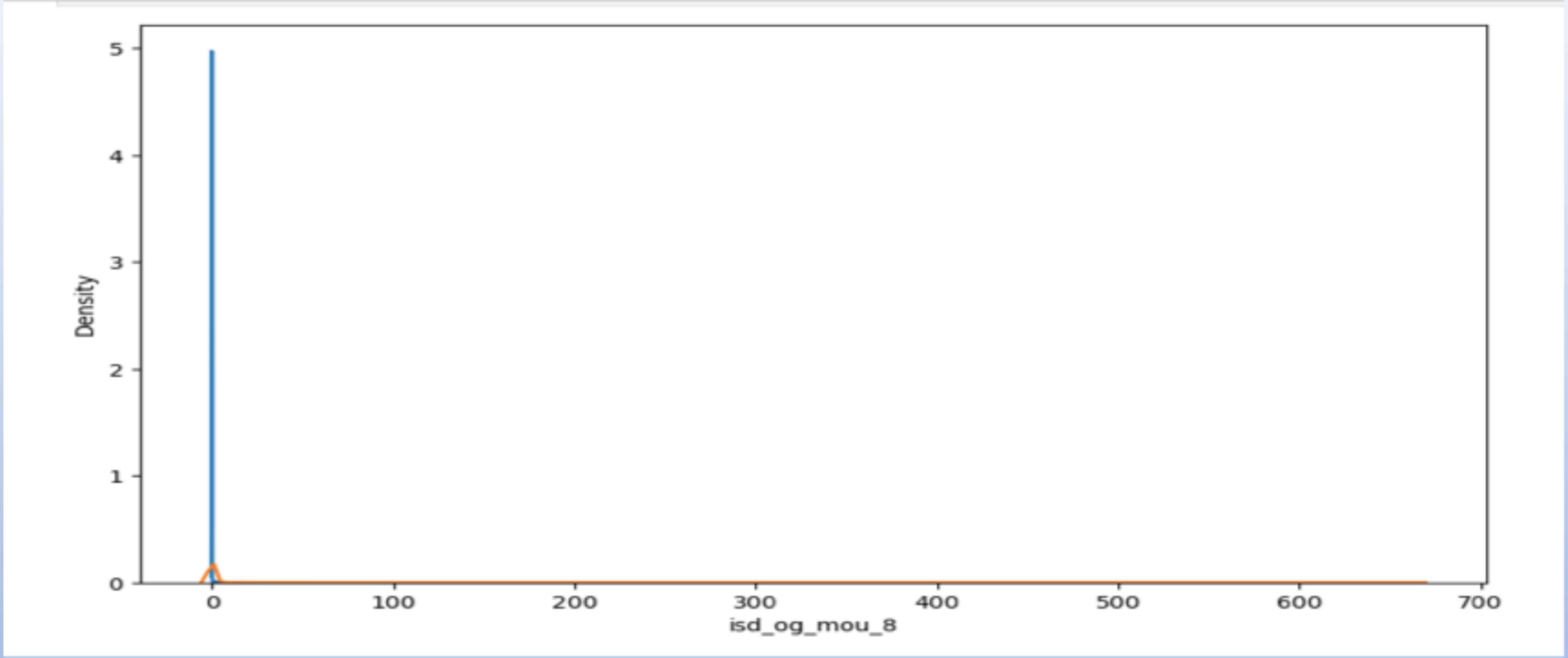
	Accuracy	Sensitivity	Specificity
Logistic Regression with PCA	0.85	0.90	0.81
Decision Tree	0.93	0.11	0.96
Random Forest	0.96	0.80	1.0
Logistic regression without PCA	0.89	0.91	0.86

We can see that the logistic model with no PCA has good sensitivity and accuracy, which are comparable to the models with PCA. So, we can go for the more simplistic model such as logistic regression with PCA as it explains the important predictor variables as well as the significance of each variable. The model also helps us to identify the variables which should be act upon for making the decision of the to be churned customers

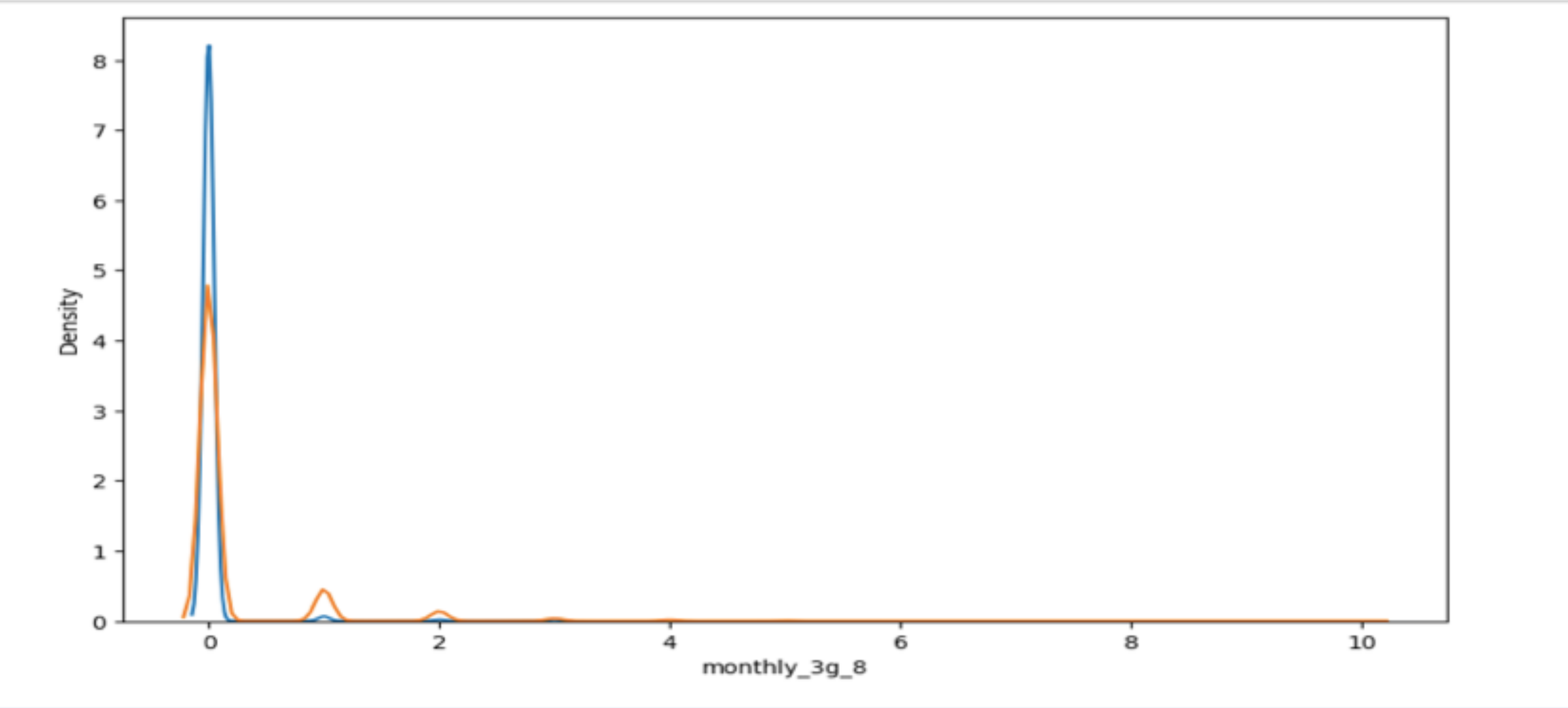
Plots of important predictors for churn and non churn customers



We can see that for the churn customers the minutes of usage for the month of August is mostly populated on the lower side than the non churn customers.



We can see that the ISD outgoing minutes of usage for the month of August for churn customers is densed approximately to zero. On the onther hand for the non churn customers it is little more than the churn customers.



The number of mothly 3g data for August for the churn customers are very much populated aroud 1, whereas of non churn customers it spreaded accross various numbers.

Business Recommendations

Top predictors

Below are few top variables selected in the logistic regression model.

Variables.	Coefficients
loc_ic_mou_8.	-1.9744
ic_others_8	-1.4913
decrease_vbc_action	-1.3078
og_others_7	-1.1915
isd_og_mou_8	-1.0212
monthly_3g_8	- 0.9871
monthly_2g_8	-0.9031
std_ic_t2f_mou_8	-0.7922
loc_ic_t2f_mou_8.	-0.7547
roam_og_mou_8	1.2482

Business Recommendations

Interpretations of The Customers Who are more likely to churn are :

- If the local incoming minutes of usage (loc_ic_mou_8) are lower in the month of August compared to previous months.
- Customers with higher outgoing charges to other operators in July and lower incoming charges from the operators in August.
- Customers with higher monthly 3G recharges in August.
- Customers with decreasing STD incoming minutes of usage for operators X to fixed lines of X in August.
- Customers with decreasing monthly 2G usage in August.
- Customers with decreasing incoming minutes of usage for operators T to fixed lines of T in August.
- Customers with increasing roaming outgoing minutes of usage in August.

Business Recommendations

Recommendations:

- Target Customers whose minutes of usage for incoming local calls and outgoing ISD calls are less in the action phase mostly in the month of August.
- Target Customers whose outgoing others charge in July and incoming others charge in August are less .
- Customers with increased value-based cost in August are more likely to churn, consider providing offers to retain them.
- Customers with higher monthly 3G recharge in August are more likely to churn.
- Customers with decreasing STD incoming minutes of usage for operators X to fixed lines of X in August.
- Customers with decreasing monthly 2G usage in August
- Customers with decreasing incoming minutes of usage for operators T to fixed lines of T in August are more likely to churn.
- Customers with increasing roaming outgoing minutes of usage in August are more likely to churn.

These Insights can help in developing targeted retention strategies to reduce customer churn.

THANK YOU