Domain Oriented Case Study

VEDANT DUBEY

VIJENDRA KUMAR GUPTA

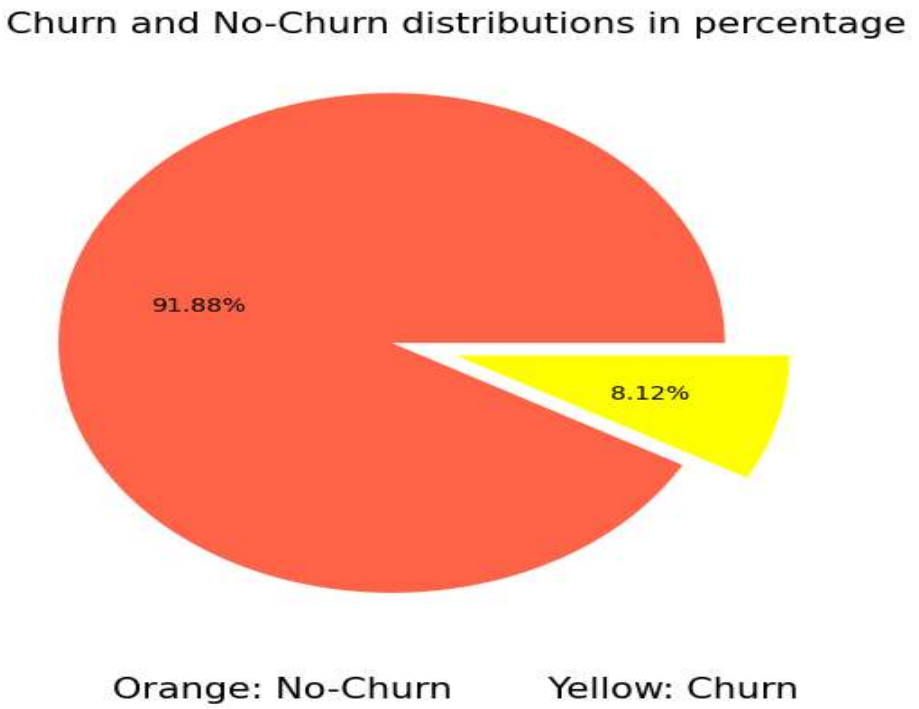
Business Problem Overview

- 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.
- 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.

Steps Involved

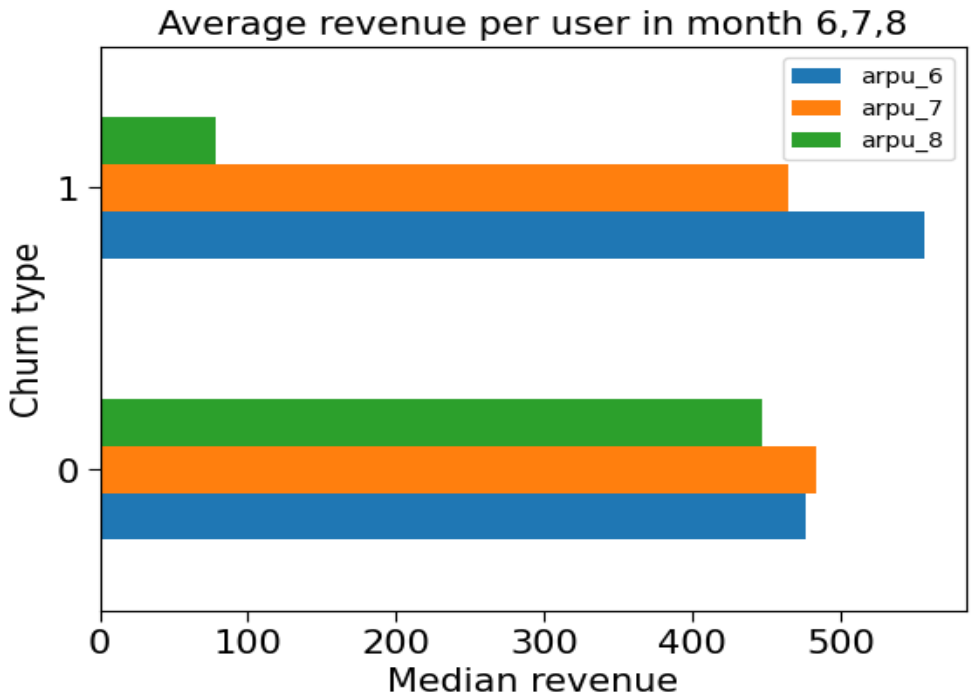
- Importing the Data
  - Importing the data from CSV file and also importing various libraries
- Data Cleaning
  - Dropping columns with more than 40% of missing values.
  - Fixing the missing rows or column by either imputing with 0 or NA.
- Exploratory Data Analysis
  - Used EDA to get the preliminary sense of the data.
- Model Building
  - Built and analyzed various models to find the best model for estimation.
- Conclusion/ Recommendation

Exploratory Data Analysis
Churn vs No Churn



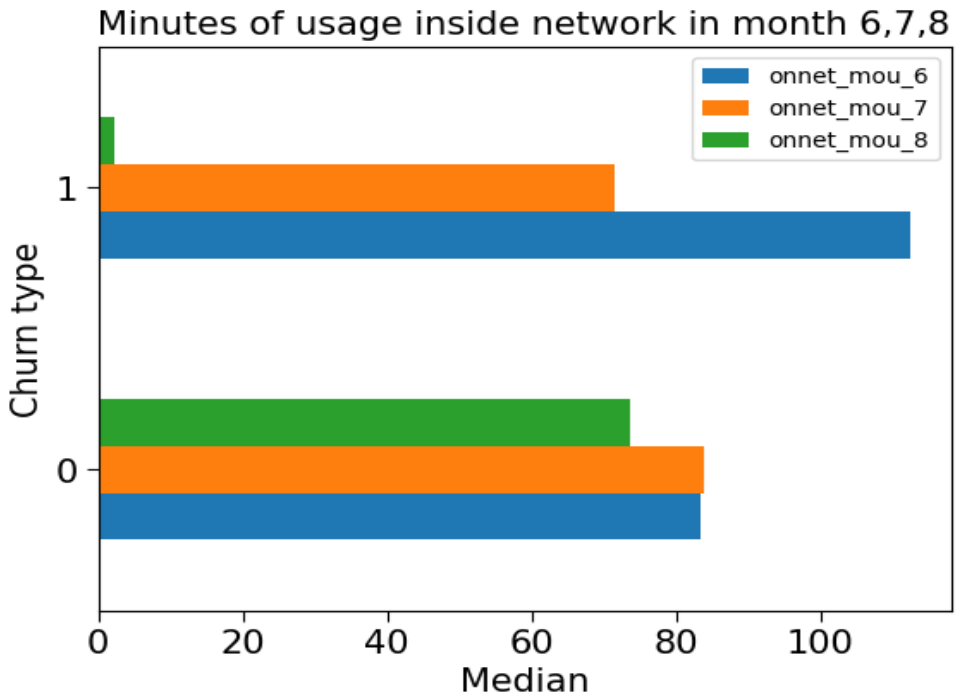
- Observation:
- ~92% of the customers belong to No-churn and ~8% belong to Churn groups respectively.

Average revenue per user in month 6,7,8



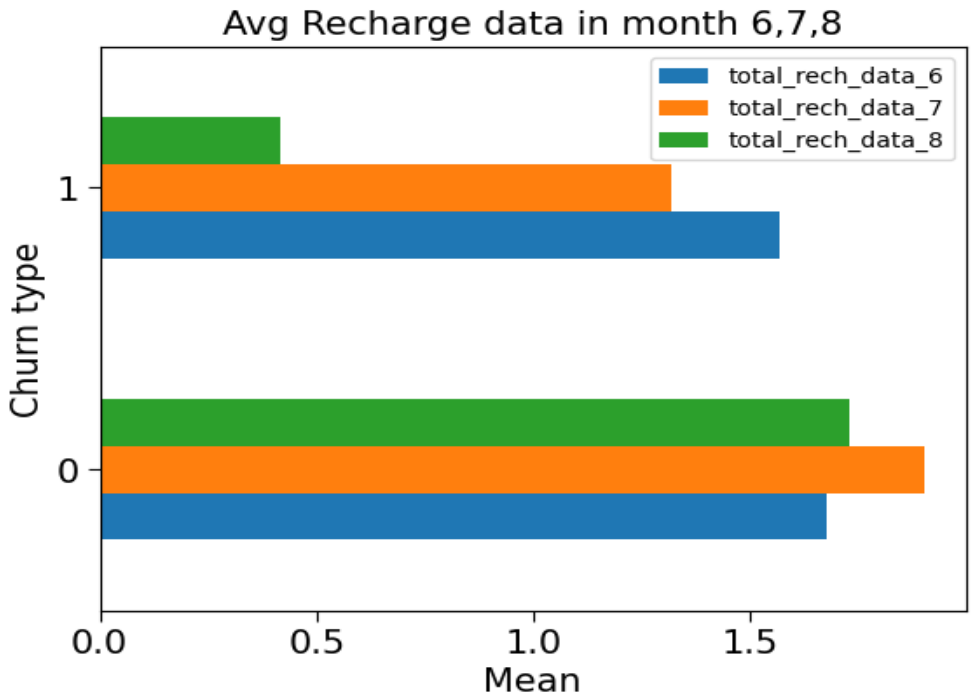
- Observation:
- Average revenue per user is more in month 6 for customer who will Churn and ARPU is more in month 7 for non Churn customers
- The variance in monthly ARPU is high in churn customers viz-a-viz non churn customers

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



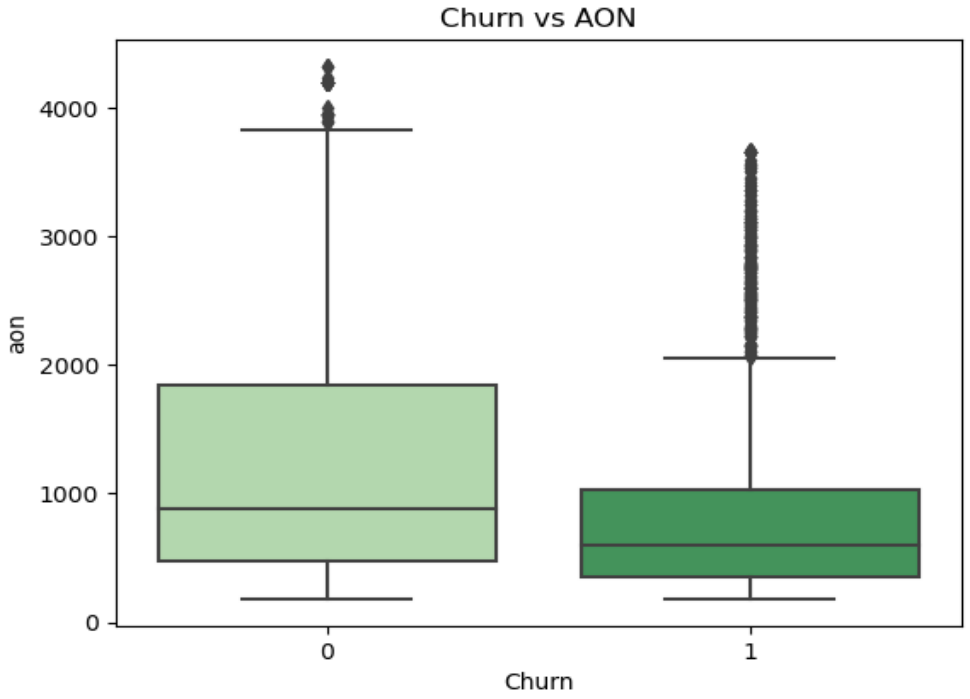
- Observation:
- MOU is more in month 6 for churn customers and MOU is more in month 7 for non churn customers
- MOU for non churn customers are lower than of churn customers
- The variance in monthly MOU is high in churn customers viz-a-viz non churn customers

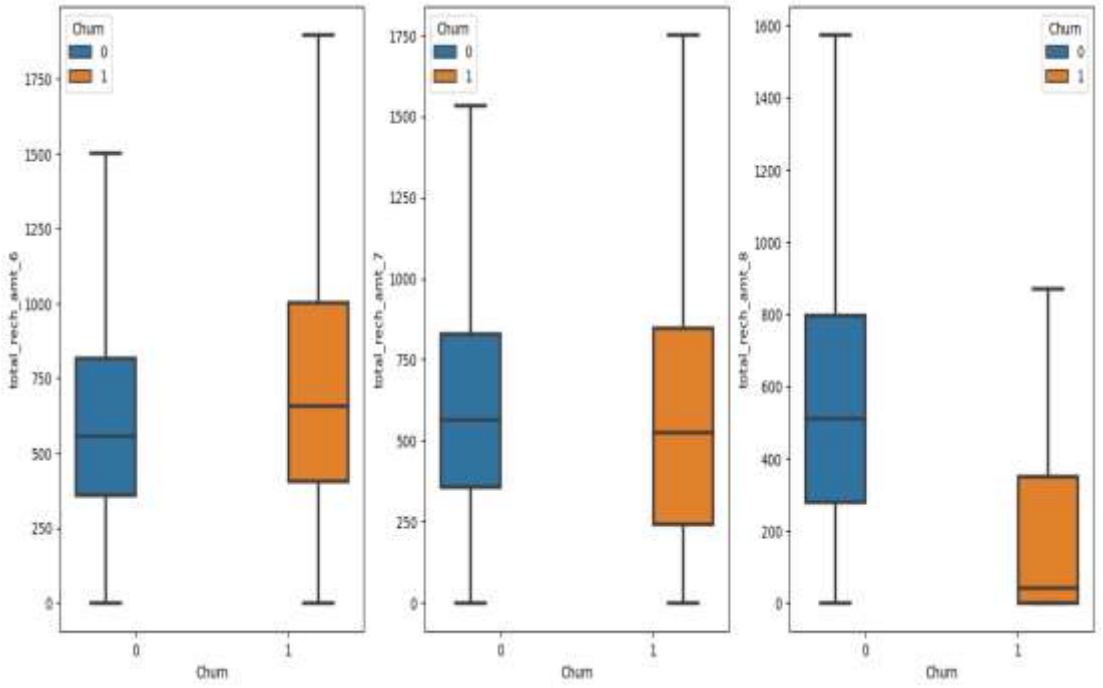
Avg Recharge data in month 6,7,8



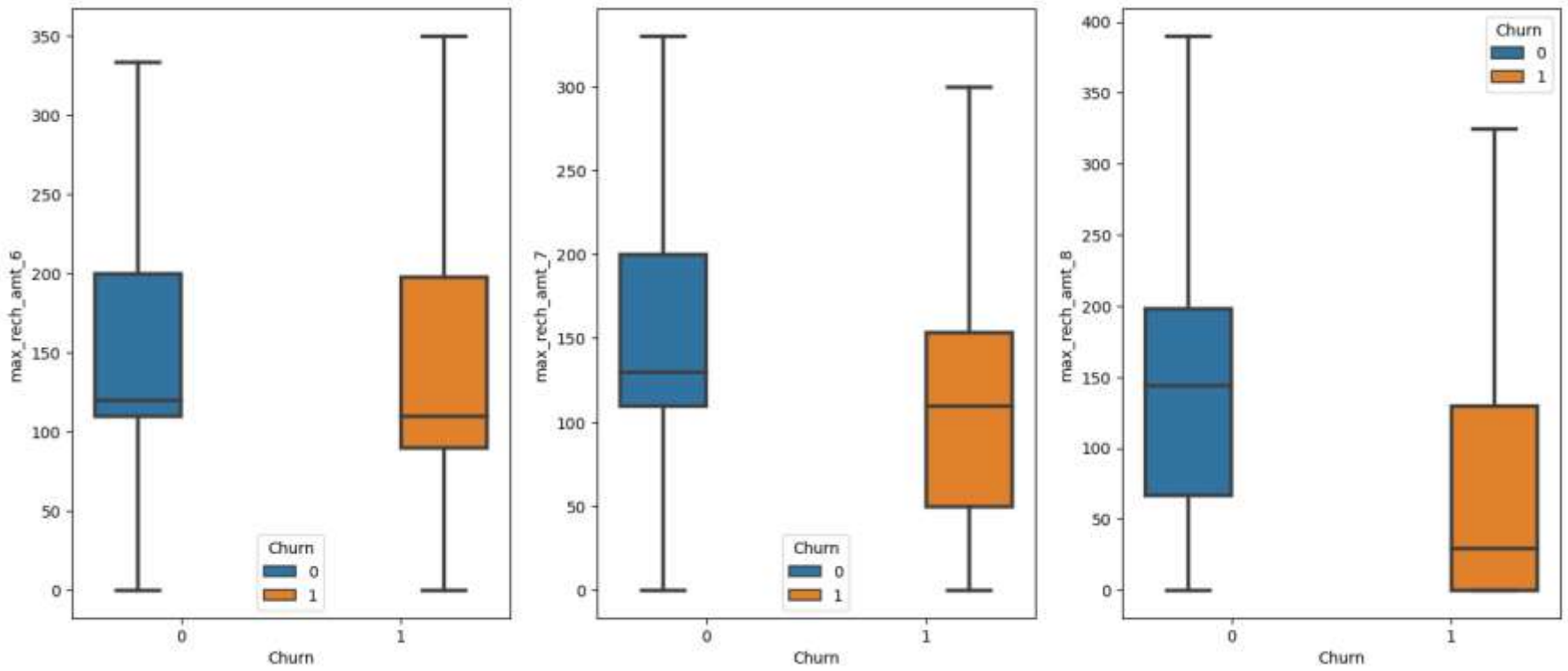
- Observation:
- Average recharge amount by non churn customers is higher than churn customers
- The variance in the average monthly recharge amount is higher in churn customers than non churn customers

Churn vs AON and Churn vs total_rech_amt





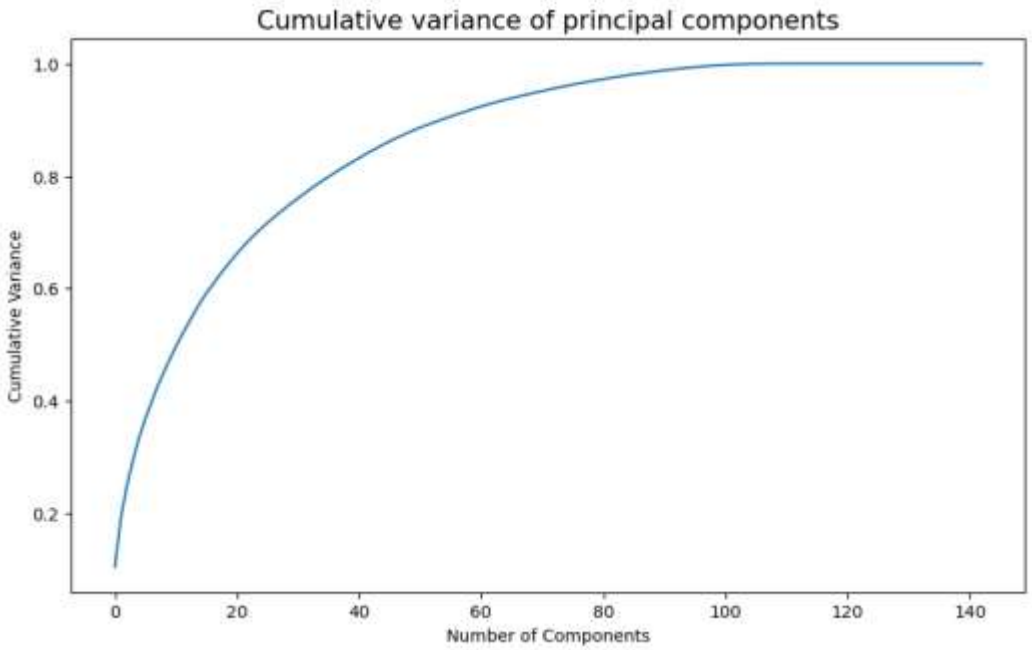
Churn vs max_rech_amt



Steps for Model fitting

- Splitting the data into Train and Test.
  - Train data is 70% while 30% is the test dataset.
- Dealing with class imbalance.
  - Dealing with the class imbalance using SMOTE (Synthetic Minority Oversampling Technique).
- Feature Scaling.
  - Scaling the feature using Standardization method.
- PCA for feature selection.

Results of PCA



- Observation:
- Here nearly 70 features explain more than 90% of the variance.
- So will perform PCA using 70 features.

Model 1: Logistic Regression with PCA

- Metrics using this model
- For Train Dataset
  - Accuracy = 86.77%
  - Sensitivity = 89.28%
  - Specificity = 84.26%
- For Test Dataset
  - Accuracy = 83.49%
  - Sensitivity = 79.15%
  - Specificity = 83.85%
- Conclusion: The model has performed fairly well on the test dataset.

Model 2: Decision Tree with PCA

- Metrics using this model
- For Train Dataset
  - Accuracy = 86.45%
  - Sensitivity = 87.54%
  - Specificity = 85.36%
- For Test Dataset
  - Accuracy = 80.80%
  - Sensitivity = 68.57%
  - Specificity = 81.81%
- Conclusion: Here the Accuracy and the specificity is good as compared to train data but the Sensitivity is reduced.

Model 3: Random Forest using PCA

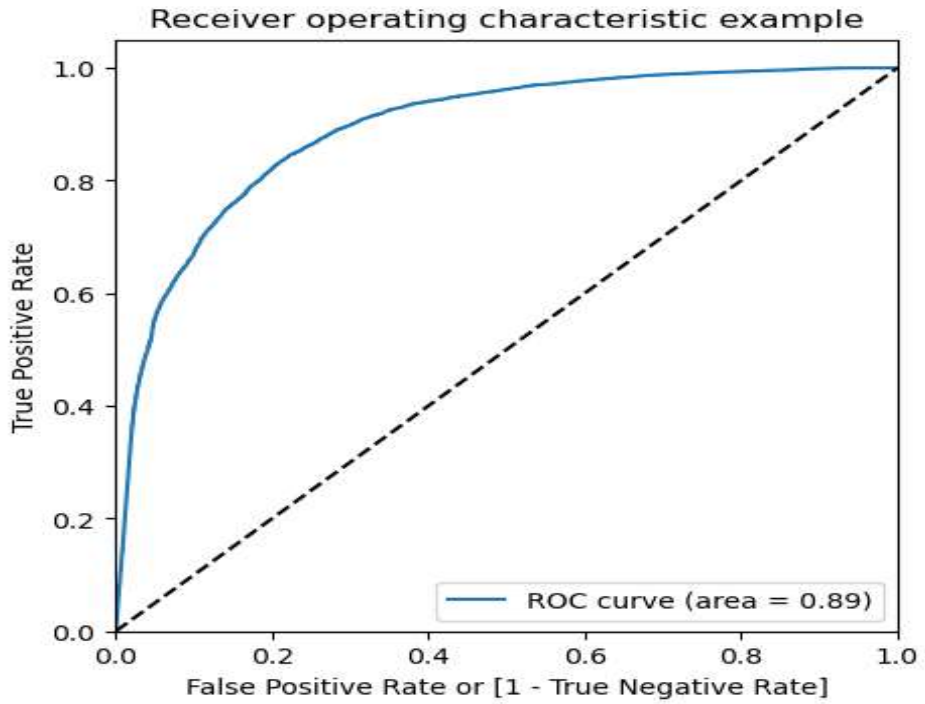
- Metrics using this model
- Best parameters:
  - 'max_depth': 15, 'max_features': 8, 'min_samples_leaf': 50, 'min_samples_split': 50, 'n_estimators': 200
- For Test Dataset
  - Accuracy = 80.80%
  - Sensitivity = 68.57%
  - Specificity = 81.81%

Model 4: Logistic Regression without PCA

- Best Features
- loc_ic_mou_8, monthly_2g_7, sachet_2g_6, sachet_2g_8, monthly_3g_6, monthly_3g_7, monthly_3g_8, sachet_3g_6, sachet_3g_8
- For Train Dataset
  - Accuracy = 80.53%
  - Sensitivity = 87.60%
  - Specificity = 73.45%
- For Test Dataset
  - Accuracy = 73.44%
  - Sensitivity = 78.41%
  - Specificity = 73.03%

Model 4: Logistic Regression without PCA

- ROC Curve



- Conclusion:
- Overall, the model is performing well in the test set, what it had learnt from the train set.

Recommendations

- Target the customers, whose minutes of usage of the incoming local calls and outgoing ISD calls are less in the action phase (mostly in the month of August).
- Target the customers, whose outgoing others charge in July and incoming others on August are less.
- Also, the customers having value based cost in the action phase increased are more likely to churn than the other customers. Hence, these customers may be a good target to provide offer.
- Max Recharge Amount is a strong feature to predict churn.
- Customers with tenure less than 4 yr are more likely to churn.