Telecom Churn Case Study

BY

PRIYA KM

Business Objective

Introduction and Problem Statement:

Business Problem Overview
In the telecom industry, customer churn is a critical issue due to high competition and significant acquisition costs. Retaining high-profit customers has become a top priority for telecom operators.

Objective
The objective of this project is to analyze customer-level data from a leading telecom firm, build predictive models to identify customers at high risk of churn, and determine the key indicators of churn.

Understanding and Defining Churn
Churn in the telecom industry can be defined differently based on payment models (postpaid vs. prepaid). In this project, churn will be defined using the usage-based method for prepaid customers, focusing on those who have not used any service (calls, internet, etc.) for a specific period.

High-Value Churn

Approximately 80% of revenue in the Indian and Southeast Asian markets comes from the top 20% of customers (high-value customers). Identifying and reducing churn among high-value customers can significantly impact revenue retention.

Market Focus
This project is based on the Indian and Southeast Asian telecom markets, where the prepaid model is predominant.

Data Understanding

Business Objective and Dataset:
The dataset includes customer-level information over four consecutive months: June (6), July (7), August (8), and September (9).

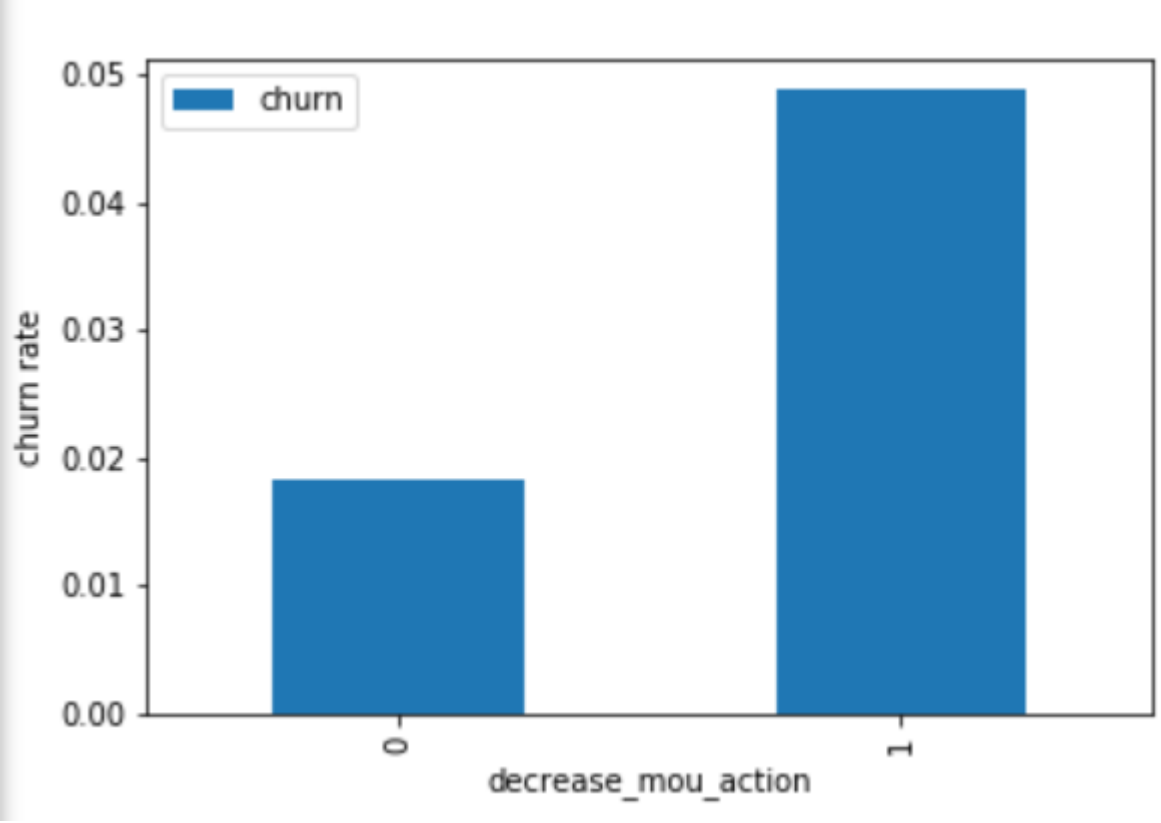
Business Objective:
Predict churn in the ninth month using data from the first three months. Understanding typical customer behavior during churn is crucial for this prediction.

Understanding Customer Behavior during Churn:
- **Good Phase**: Customer is satisfied and behaves normally.
- **Action Phase**: Customer experience deteriorates due to competitor offers, service issues, etc.
- **Churn Phase**: Customer decides to churn. This phase defines the churn outcome.

Data Dictionary:
The data dictionary provides meanings for abbreviations used in the dataset, such as loc (local), IC (incoming), OG (outgoing), T2T (telecom operator to telecom operator), T2O (telecom operator to another operator), RECH (recharge), etc.

Data Understanding

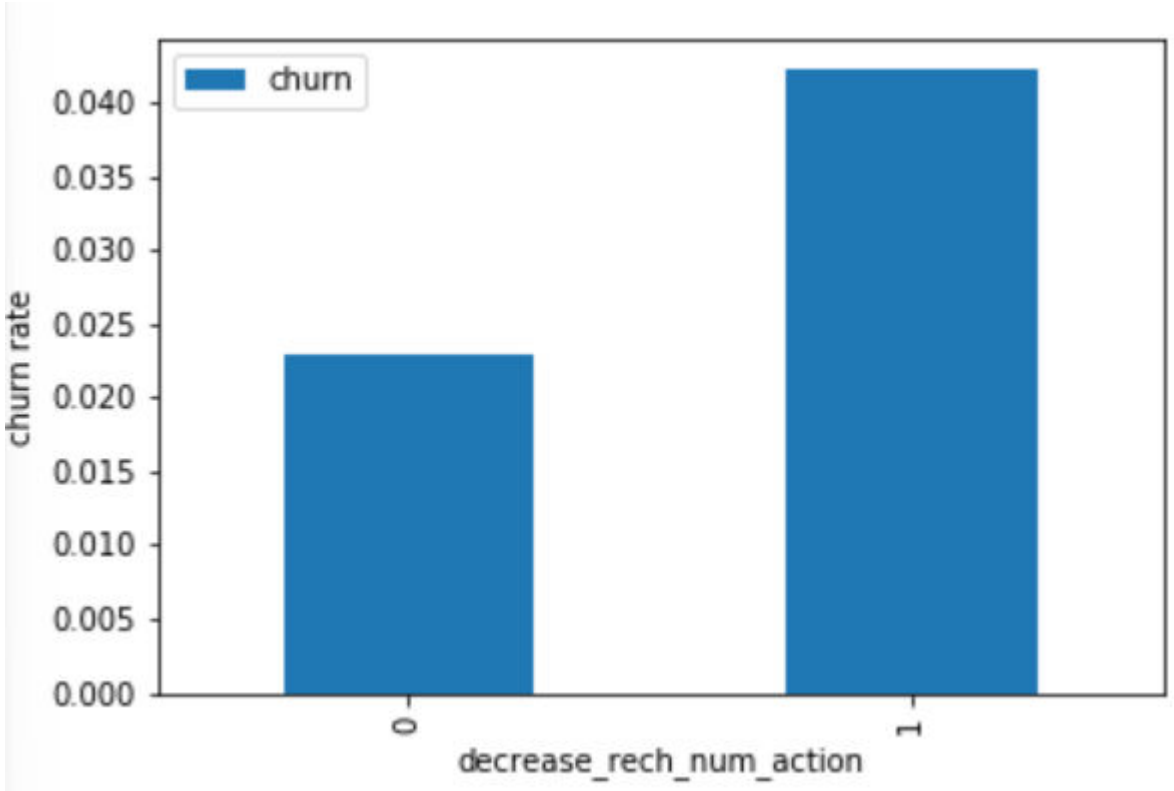
Churn rate based on whether the customer's MOU decreased in the action month:



We observe a higher churn rate among customers whose Minutes of Usage (MOU) decreased in the action phase compared to the good phase.

Data Understanding

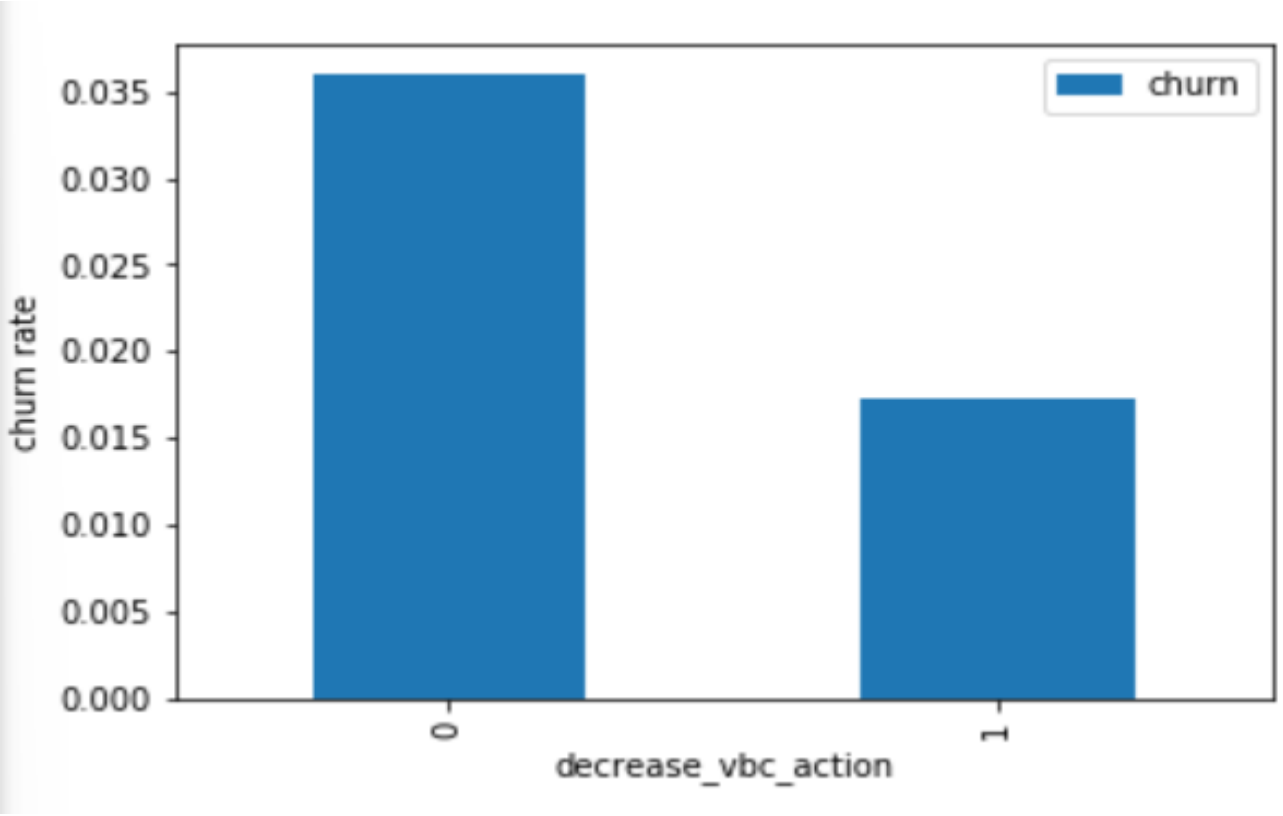
Churn rate on the basis whether the customer decreased her/his number of recharge in action month



The churn rate is higher among customers whose number of recharges decreased in the action phase compared to the good phase.

Data Understanding

Churn rate on the basis of whether the customer decreased her/his volume based cost in action month:



The churn rate is higher for customers whose volume-based cost in the action month has increased. This suggests that customers do not recharge more monthly when they are in the action phase.

Problem solving steps

- Perform data cleaning on the datasets to handle missing values, outliers, and other inconsistencies.
- Perform Exploratory Data Analysis on all the features of the datasets to identify the important predictor variables.
- Perform modeling with Principal component analysis (PCA).
- Perform Logistic Regression model with PCA and No PCA, optimal C.
- Tuning hyperparameter C
- Performing Support Vector Machine(SVM) with PCA.
- Perform some complex models like Decision tree and Random forest to see if the predictive power of the model is better than that of the logistic regression model.
- Access the financial benefits of the project by checking the underlying matrices that get optimized.
- Present all the results obtained in all the above steps to the management

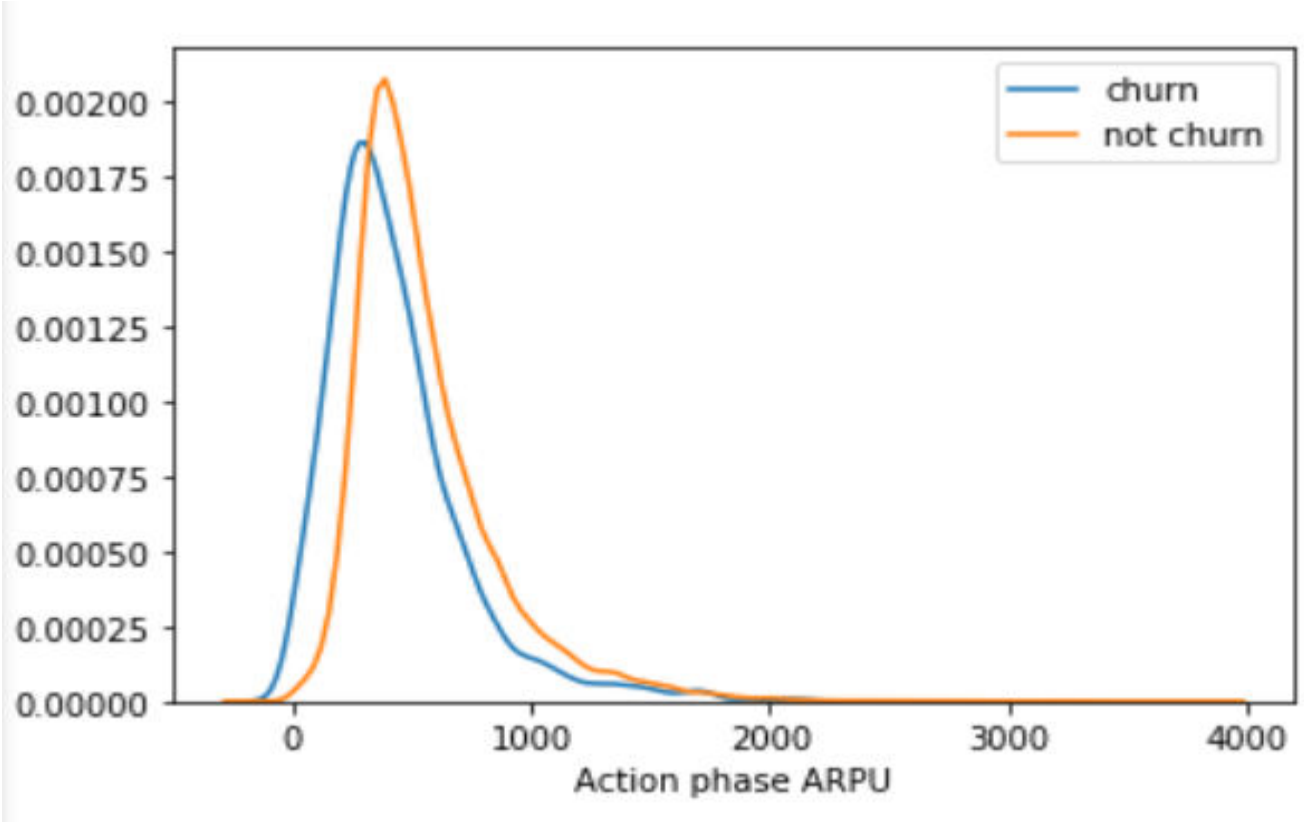
Criteria for model evaluation

All the models will be evaluated on the following parameters :

- Confusion matrix for each model.
- Sensitivity, specificity, accuracy curve for each model with different cut-offs.
- AUC-ROC curve for the model using cut-off values for each model.
- Precision and Recall curve for cut-off should be generated.
- Gini-Index needs to be evaluated for Tree based models like decision tree and random forest.
- Within each model type evaluation using Grid Search based on recall values should be done to get models with optimized hyper parameters.

Average Revenue Analysis

Action Phase Comparison Between Churn and Non-Churn Customers:



Average Revenue Per User (ARPU) Analysis

Churned Customers:

- ARPU is predominantly in the range of 0 to 900.
- Higher ARPU customers are less likely to churn.

Non-Churned Customers:

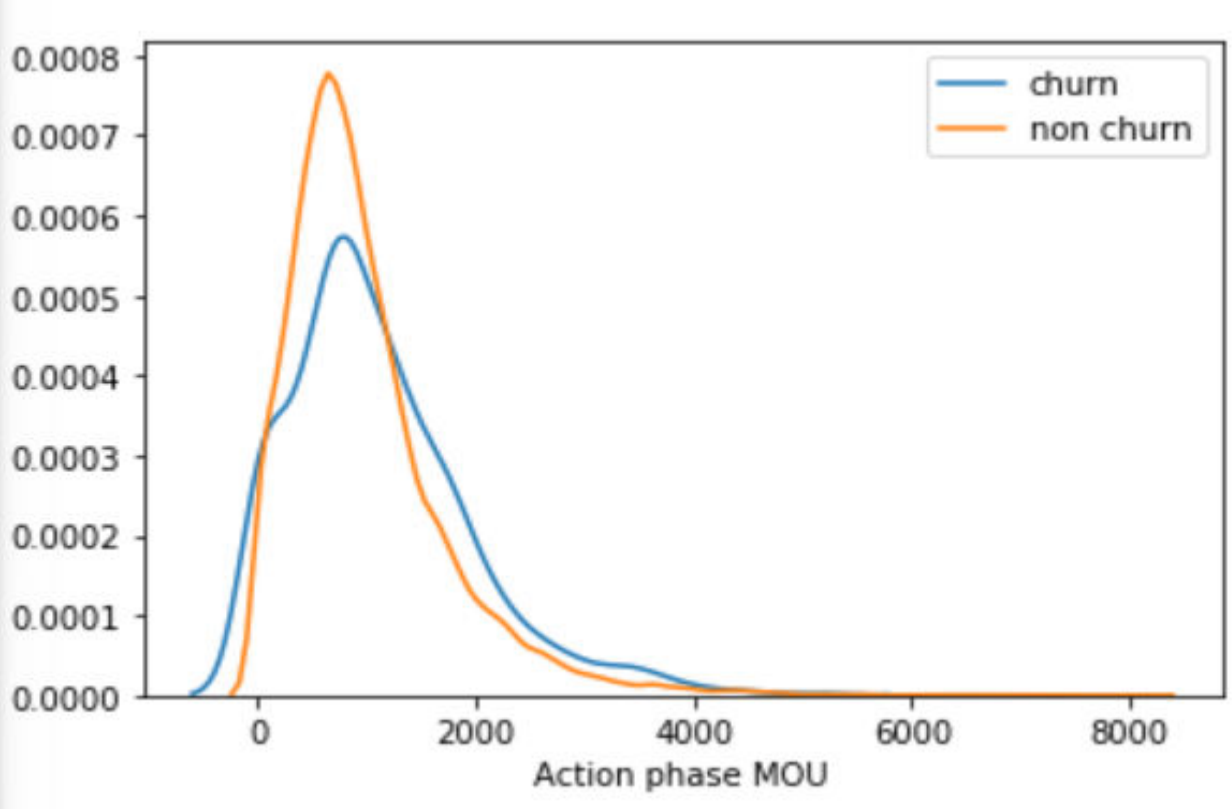
- ARPU is mostly concentrated in the range of 0 to 1000.

Insights:

- Churned customers tend to have lower ARPU compared to non-churned customers.
- Higher ARPU correlates with a lower likelihood of churn.

Minutes of Usage (MOU) Analysis

Analysis of the minutes of usage MOU (churn and not churn) in the action phase:



Minutes of Usage (MOU) Analysis:

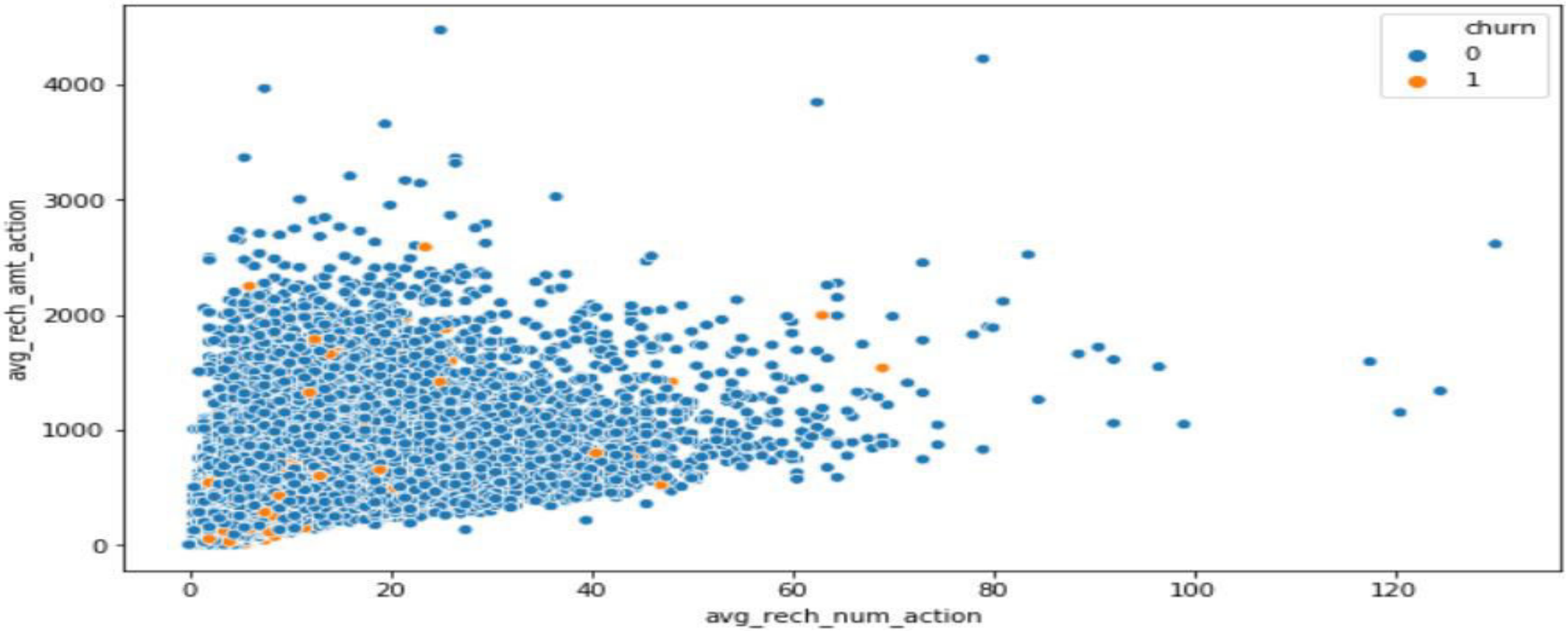
Churned customers:

- MOU is predominantly concentrated in the range of 0 to 2500.
- Higher MOU generally indicates a lower churn probability.

Insights:

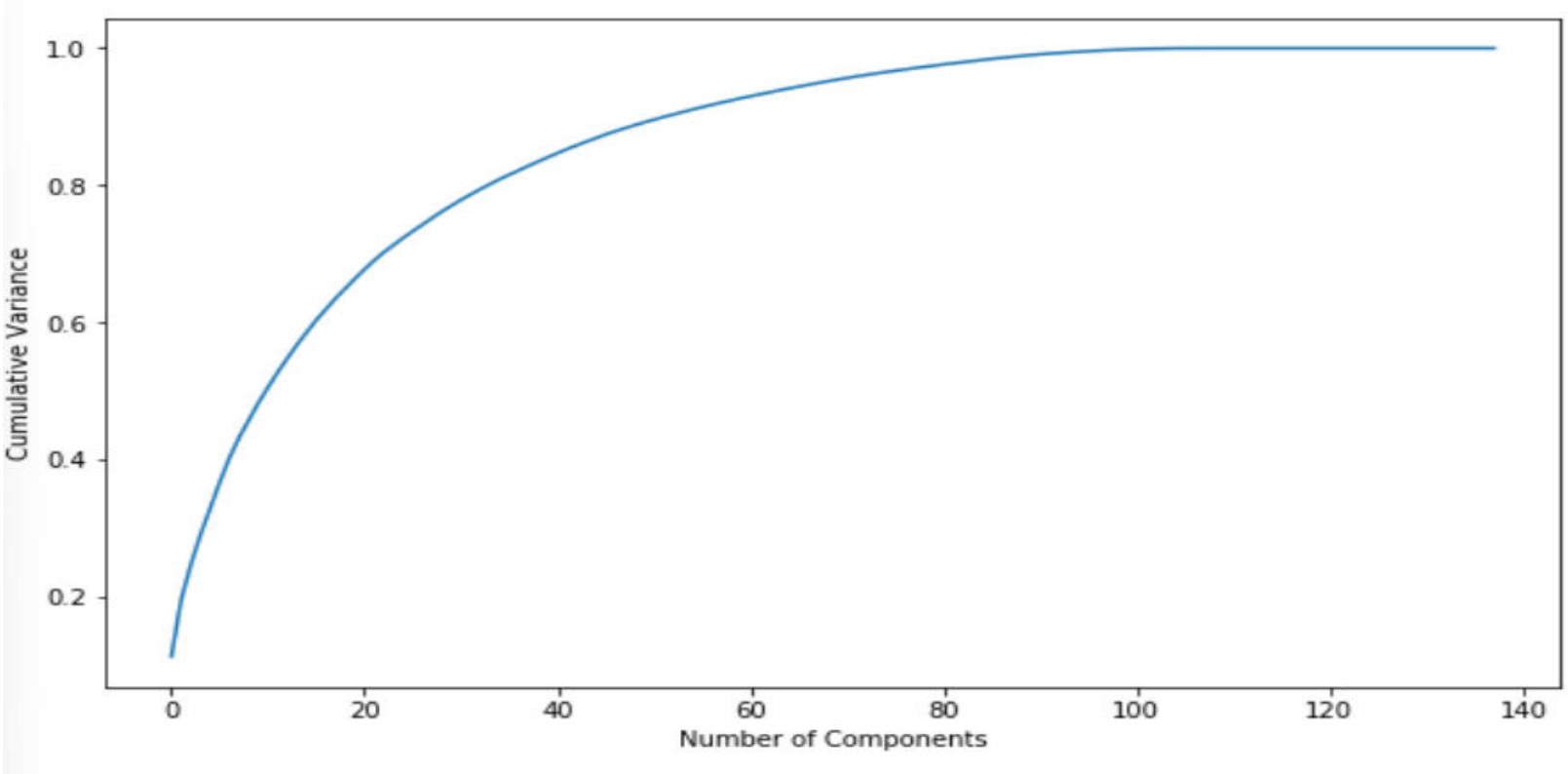
- Churned customers exhibit lower MOU compared to non-churned customers in the action phase.
- Non-churned customers have a more varied distribution of MOU, suggesting higher engagement and satisfaction.

Analysis of Recharge Amount and Frequency in Action Month



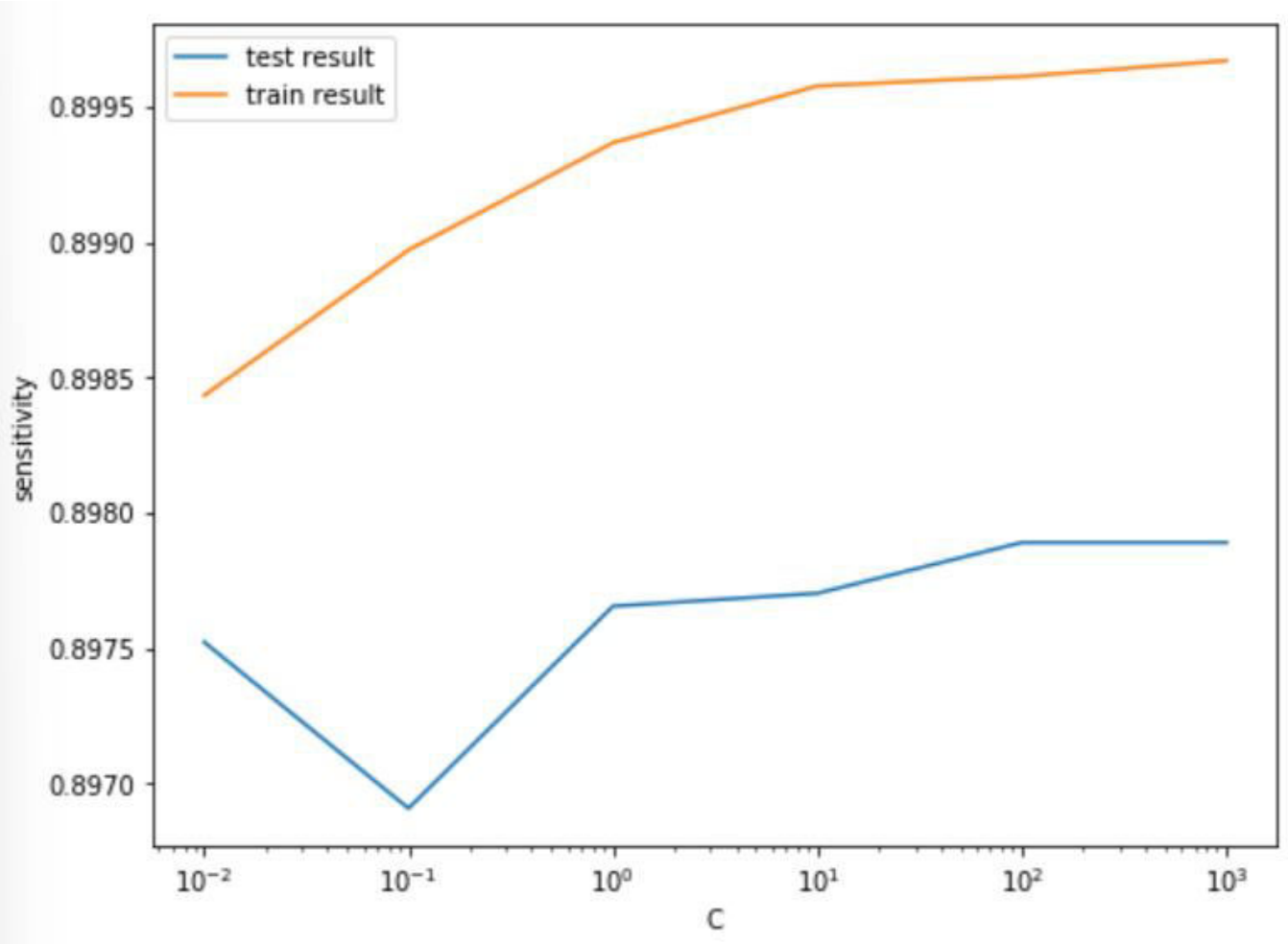
From the above pattern, we can observe that the number of recharges and the recharge amount are mostly proportional. The higher the number of recharges, the greater the recharge amount.

Model with PCA



Sixty components explain over 90% of the data's variance.

Logistic regression with PCA



Logistic regression with optimal C

Train set

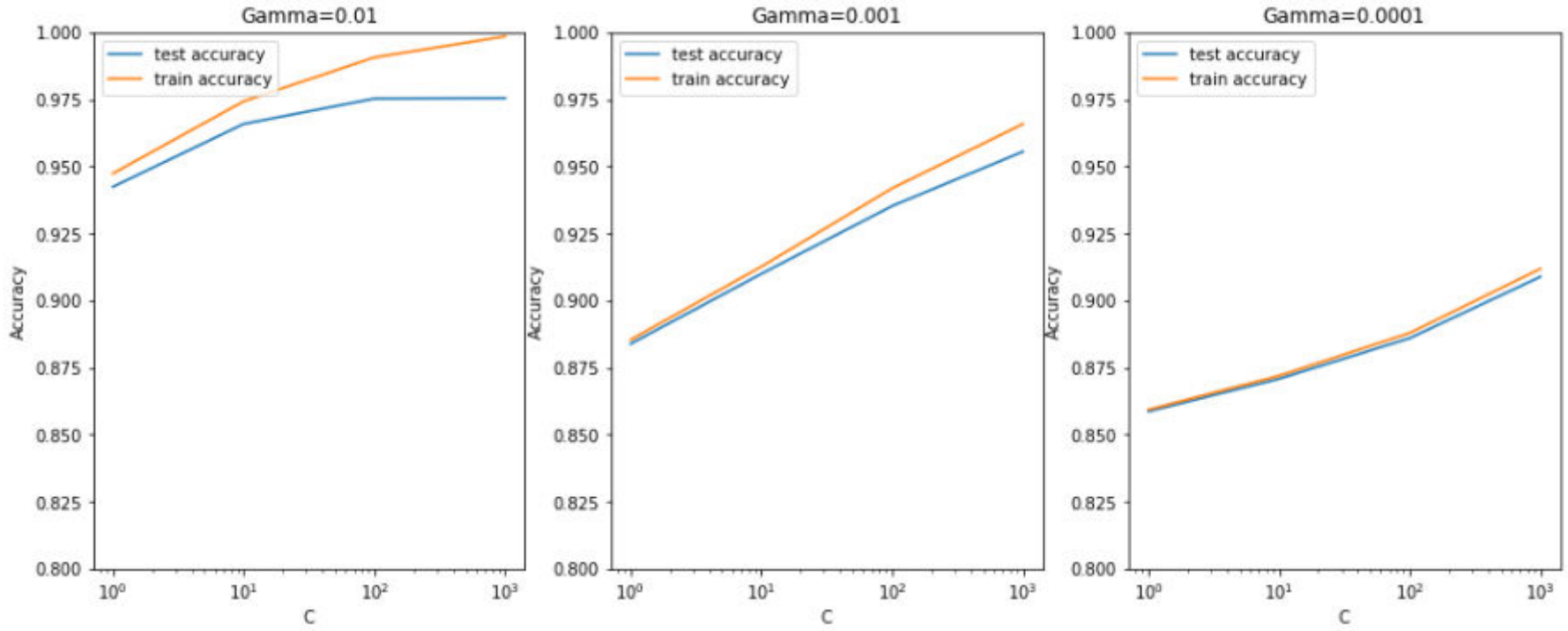
- Accuracy = 0.86
- Sensitivity = 0.89
- Specificity = 0.83

Test set

- Accuracy = 0.83
- Sensitivity = 0.81
- Specificity = 0.83

The model is performing well in the test set, consistent with what it learned from the train set.

Support Vector Machine(SVM) with PCA



The best test score is 0.9754959911159373 corresponding to hyperparameters C: 1000, gamma: 0.01

Support Vector Machine(SVM) with PCA

- Higher gamma values lead to overfitting; lower gamma (0.0001) results in similar train and test accuracy.
- At C = 100, we achieve good accuracy with comparable train and test scores.
- sklearn suggests optimal scores at gamma = 0.01 and C = 1000, but simpler models with gamma = 0.0001 and higher C can be better.
- Comparable average test accuracy (~90%) can be achieved with gamma = 0.0001 by increasing C.
- Tradeoff between high gamma (high non-linearity) and average C vs. low gamma (low non-linearity) and high C.
- A simpler model with gamma = 0.0001 and C = 100 is preferred for less non-linearity.

Decision tree with PCA

Train set:

- Accuracy = 0.90
- Sensitivity = 0.91
- Specificity = 0.88

Test set:

- Accuracy = 0.86
- Sensitivity = 0.70
- Specificity = 0.87

We can see from the model performance that the sensitivity has decreased while evaluating the model on the test set. However, the accuracy and specificity are quite good in the test set.

Random forest with PCA

Train set:

- Accuracy = 0.84
- Sensitivity = 0.88
- Specificity = 0.80

Test set:

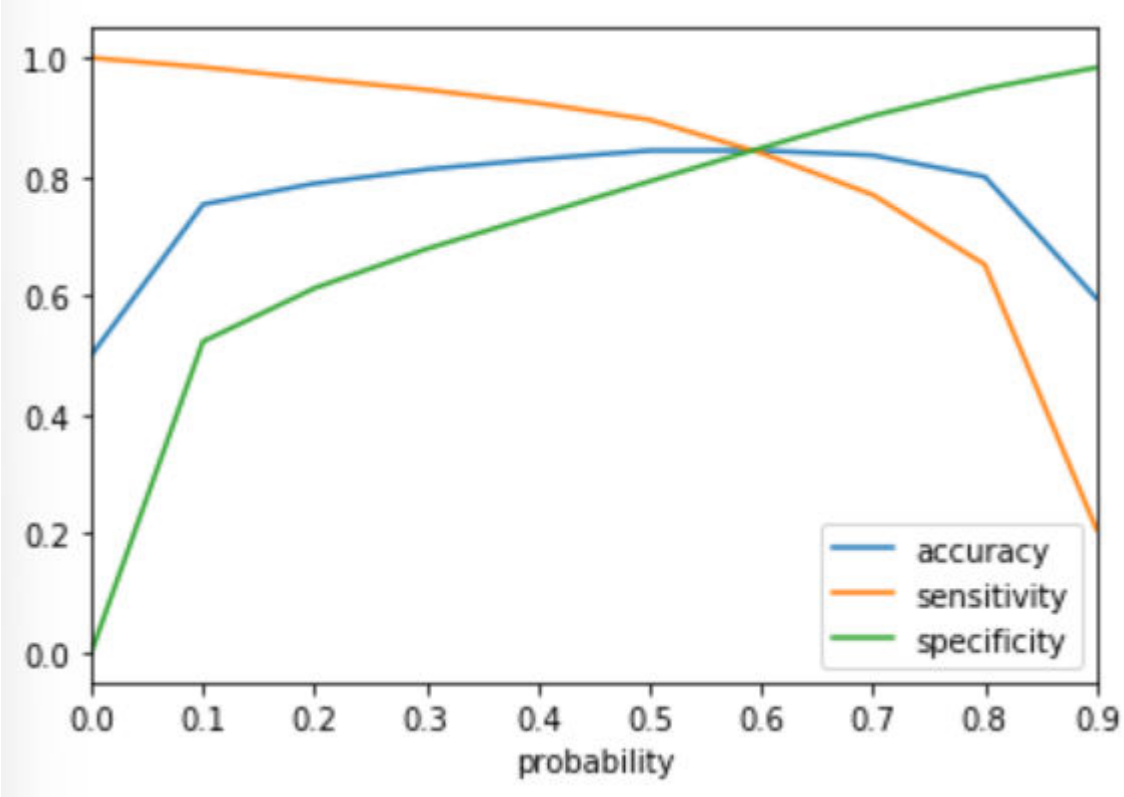
- Accuracy = 0.80
- Sensitivity = 0.75
- Specificity = 0.80

We can see from the model performance that the sensitivity has decreased while evaluating the model on the test set. However, the accuracy and specificity are quite good in the test set.

Final Conclusion with PCA

After testing several models, it is evident that for achieving the best sensitivity, which was our primary goal, the classic Logistic Regression and SVM models perform well. Both models achieved a sensitivity of approximately 81% and an accuracy of around 85%.

Logistic regression with No PCA



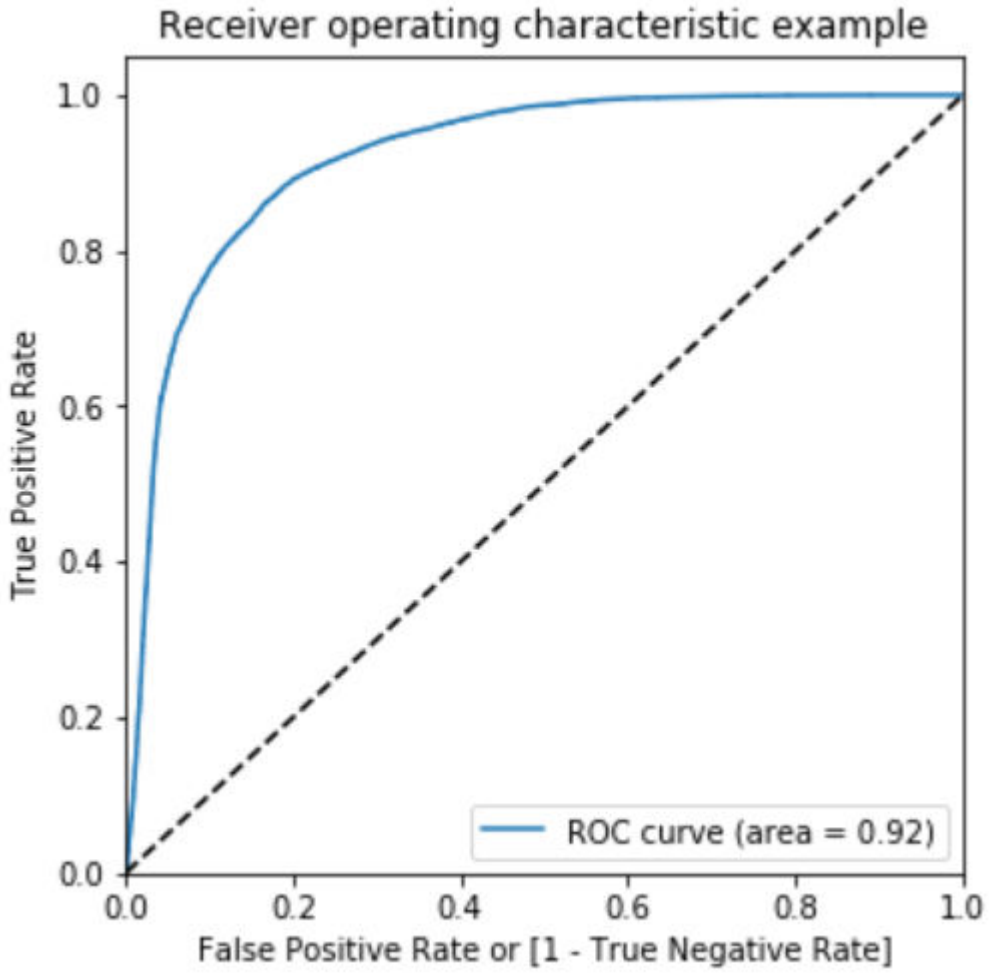
Analysis of the Curve:

- **Accuracy**: Stabilizes around 0.6.
- **Sensitivity**: Decreases with the increased probability cutoff.
- **Specificity**: Increases with the increasing probability cutoff.

At the 0.6 mark, where the three parameters intersect, there is a balance between sensitivity and specificity with good accuracy.

However, since our primary goal is to achieve better sensitivity, we opt for a probability cutoff of 0.5, despite the curve suggesting 0.6 as optimal. This choice allows us to prioritize higher sensitivity over accuracy and specificity.

Plotting the ROC Curve (Trade off between sensitivity & specificity)



We can see the area of the ROC curve is closer to 1, which is the Gini of the model.

Plotting the ROC Curve (Trade off between sensitivity & specificity)

Model Summary

Train set:

- Accuracy = 0.84
- Sensitivity = 0.81
- Specificity = 0.83

Test set:

- Accuracy = 0.78
- Sensitivity = 0.82
- Specificity = 0.78

Overall, the model is performing well in the test set, consistent with what it learned from the train set.

Final Conclusion without PCA:

The logistic regression model without PCA exhibits good sensitivity and accuracy, comparable to models with PCA. Thus, opting for the simpler logistic regression model without PCA is advisable as it effectively explains important predictor variables and their significance. Additionally, this model aids in identifying actionable variables for decision-making regarding potentially churned customers, making it more relevant for business interpretation.

Business recommendation

Top Predictors

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

Variables	Coefficients
loc_ic_mou_8	-3.3287
og_others_7	-2.4711
ic_others_8	-1.5131
isd_og_mou_8	-1.3811
decrease_vbc_action	-1.3293

Variables	Coefficients
monthly_3g_8	-1.0943
std_ic_t2f_mou_8	-0.9503
monthly_2g_8	-0.9279
loc_ic_t2f_mou_8	-0.7102
roam_og_mou_8	0.7135

We can see that most of the top variables have negative coefficients, indicating an inverse correlation with the churn probability.

Business recommendation

Example Interpretations:

- If the local incoming minutes of usage (loc_ic_mou_8) are lower in the month of August compared to previous months, there is a higher likelihood that the customer will churn.
- Customers with higher outgoing charges to other operators in July and lower incoming charges from other operators in August are more likely to churn.
- Customers with increased value-based costs in the action phase (August) are more likely to churn.
- Customers with higher monthly 3G recharges in August are likely to churn.
- Customers with decreasing STD incoming minutes of usage for operators T to fixed lines of T in August are more likely to churn.
- Customers with decreasing monthly 2G usage in August are more likely to churn.
- 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.

Business recommendation

Recommendations:

- Target customers whose minutes of usage for incoming local calls and outgoing ISD calls are less in the action phase (mostly in 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 likely to churn.
- Customers with decreasing STD incoming minutes of usage for operators T to fixed lines of T in August are more likely to churn.
- Customers with decreasing monthly 2G usage in August are likely to churn.
- 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

BY

PRIYA KM