Domain Oriented:
Telecom Churn Case Study

Presented by:

- Saleha Razvi
- Debabrata Panda
- Ramisetty Maneesha

Problem Statement

- 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.

Business Objective

- 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.
- define high-value customers based on a certain metric and predict churn only on high-value customers.

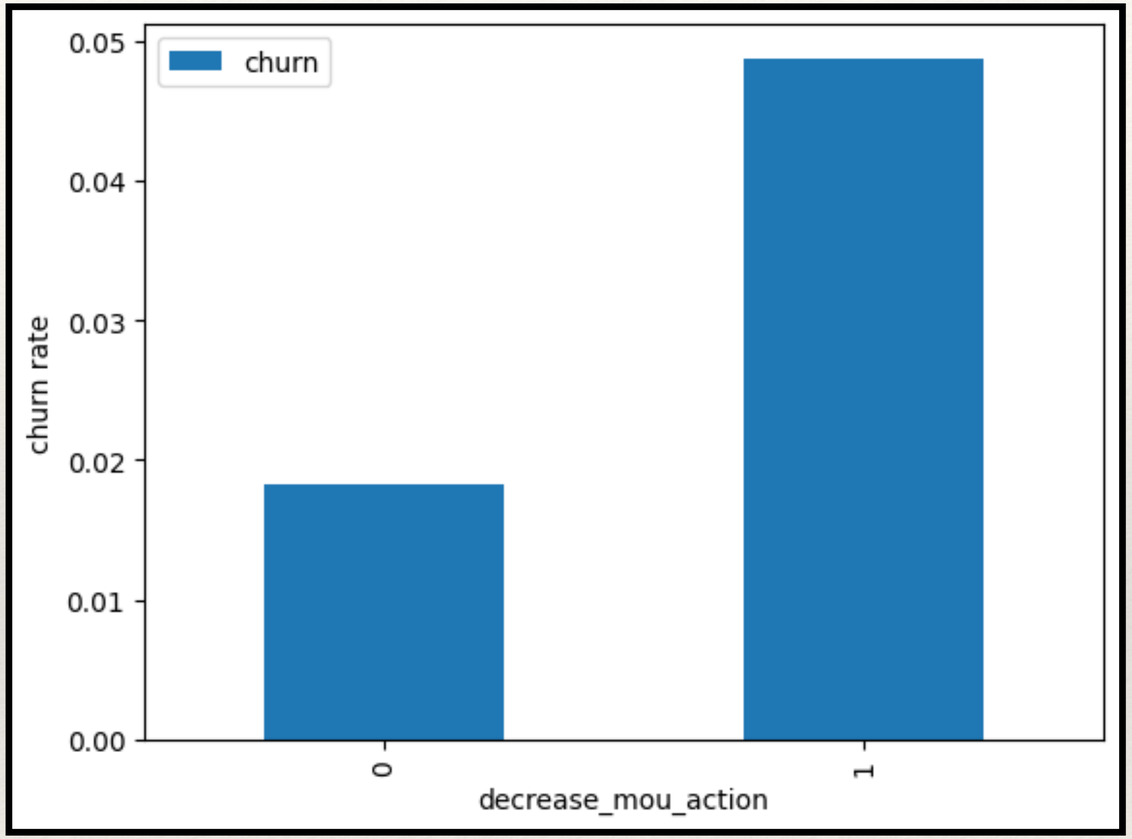
Data Analysis

- Total number of rows and columns are 99999, 226.
- We can see there are a lot of column which have high number of missing values which are greater than 70% so these columns are of no use so we can drop them
- We can further drop a few more columns such as date,circle_id, date of last recharge etc.
- After this we can filter out the high value customers by By summing up total recharge amount of months June and July

EDA

Churn rate on the basis whether the customer decreased her/his MOU in action month

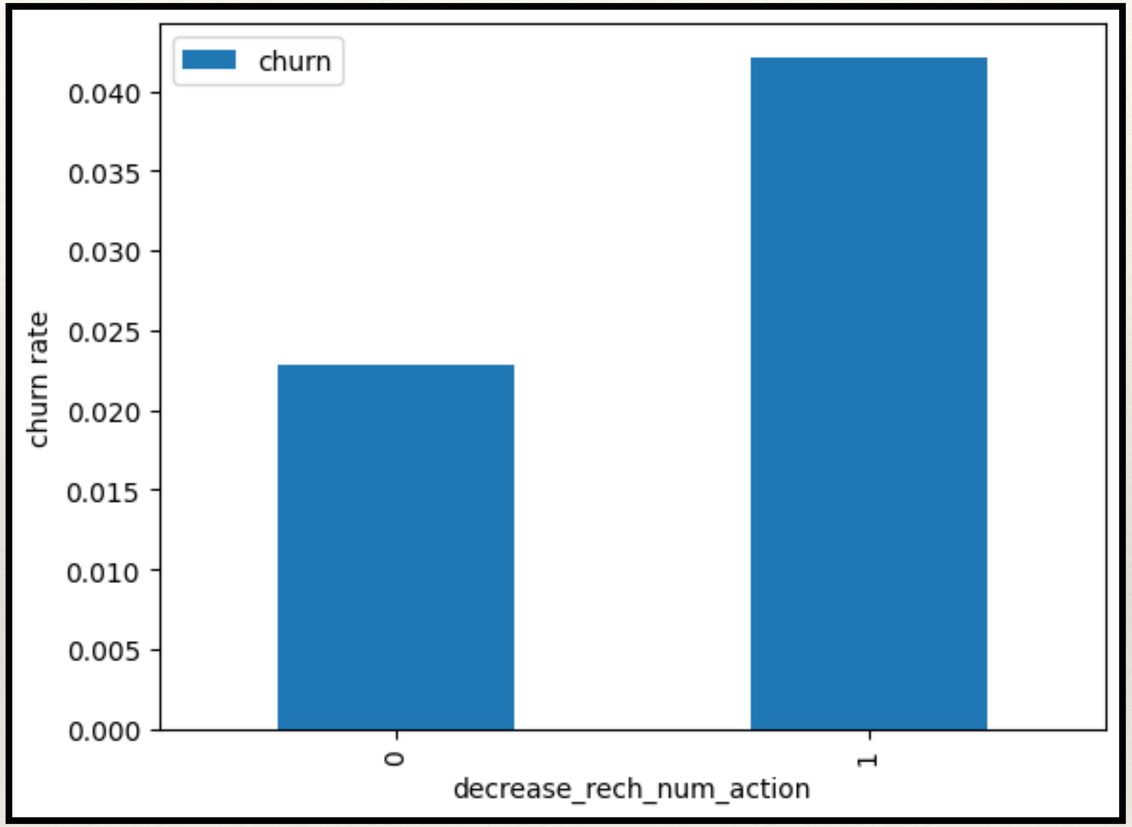
- From the above graph, we can see that the churn rate is more for the customers, whose minutes of usage(MOU) decreased in the action phase than in the good phase.



EDA

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

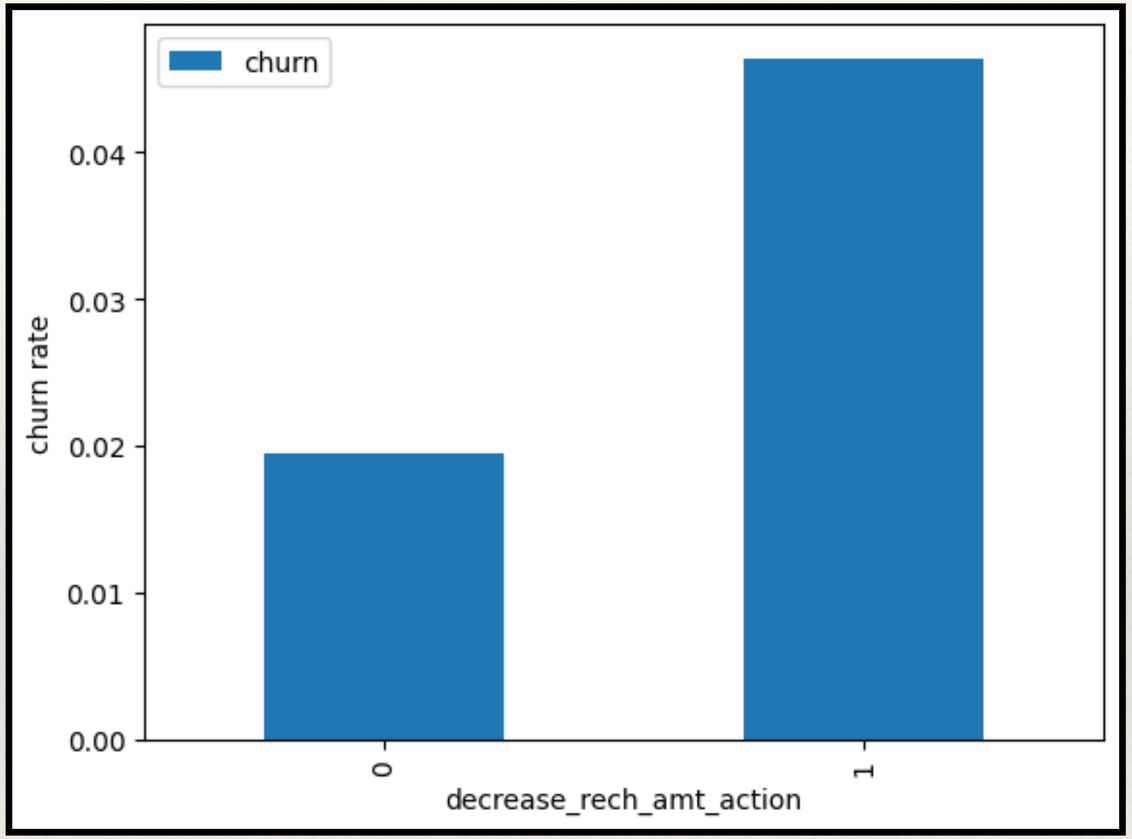
- From the graph, it can be concluded that the churn rate is more for the customers whose number of recharges in the action phase is lesser than the number in the good phase.



EDA

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

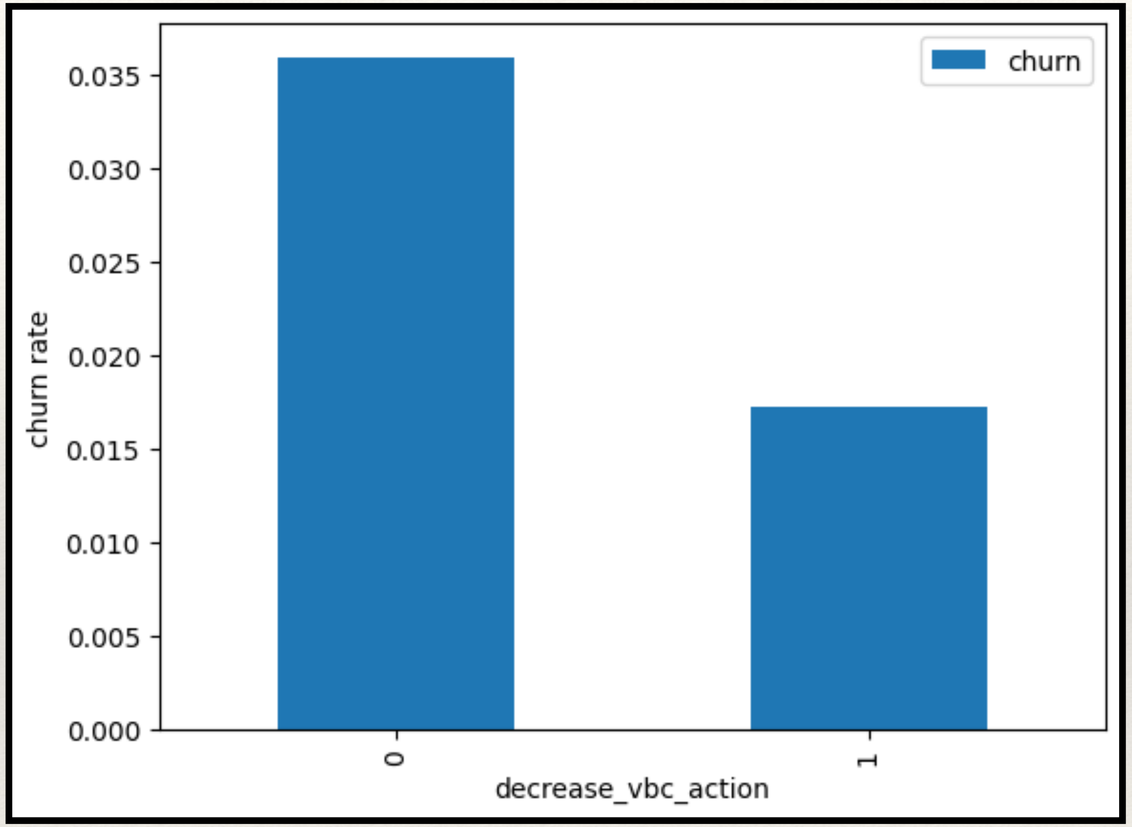
- From the above graph it can be concluded that the same behaviour is followed i.e. the churn rate is more for the customers whose amount of recharge in the action phase is lesser than the amount in the good phase.



EDA

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

- From the graph, it can be concluded the expected result. The churn rate is more for the customers, whose volume-based cost in action month is increased. That means the customers do not do the monthly recharge more when they are in the action phase.

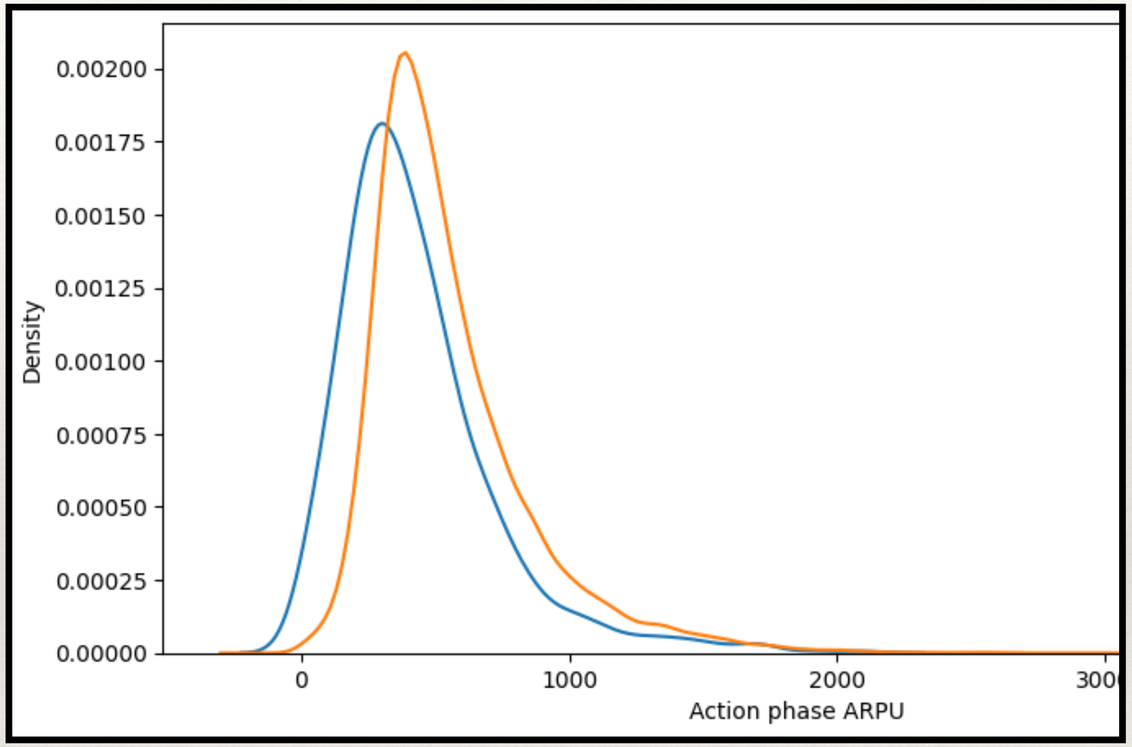


EDA

Analysis of the average revenue per customer (churn and not churn) in the action phase

- From the above graph, it can be concluded that the Average revenue per user (ARPU) for the churned customers is most dense in the 0 to 900. The higher ARPU customers are less likely to be churned.

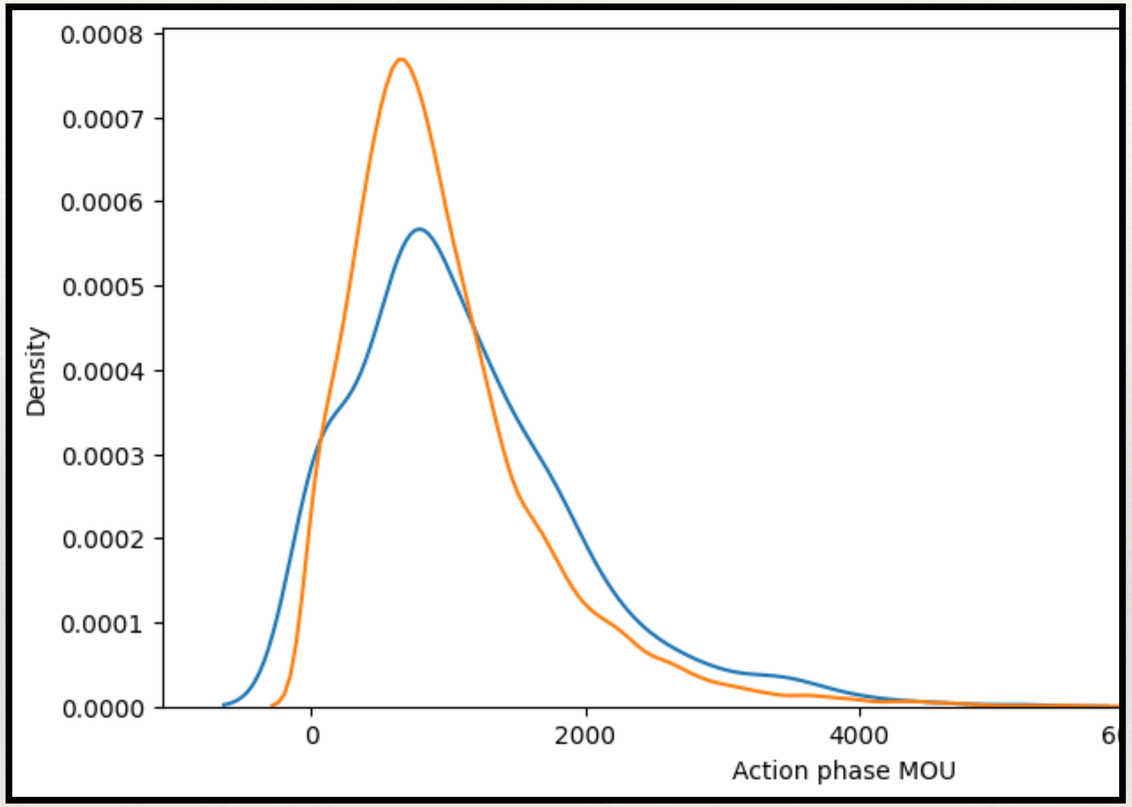
- ARPU for the not churned customers is most dense on the 0 to 1000.



EDA

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

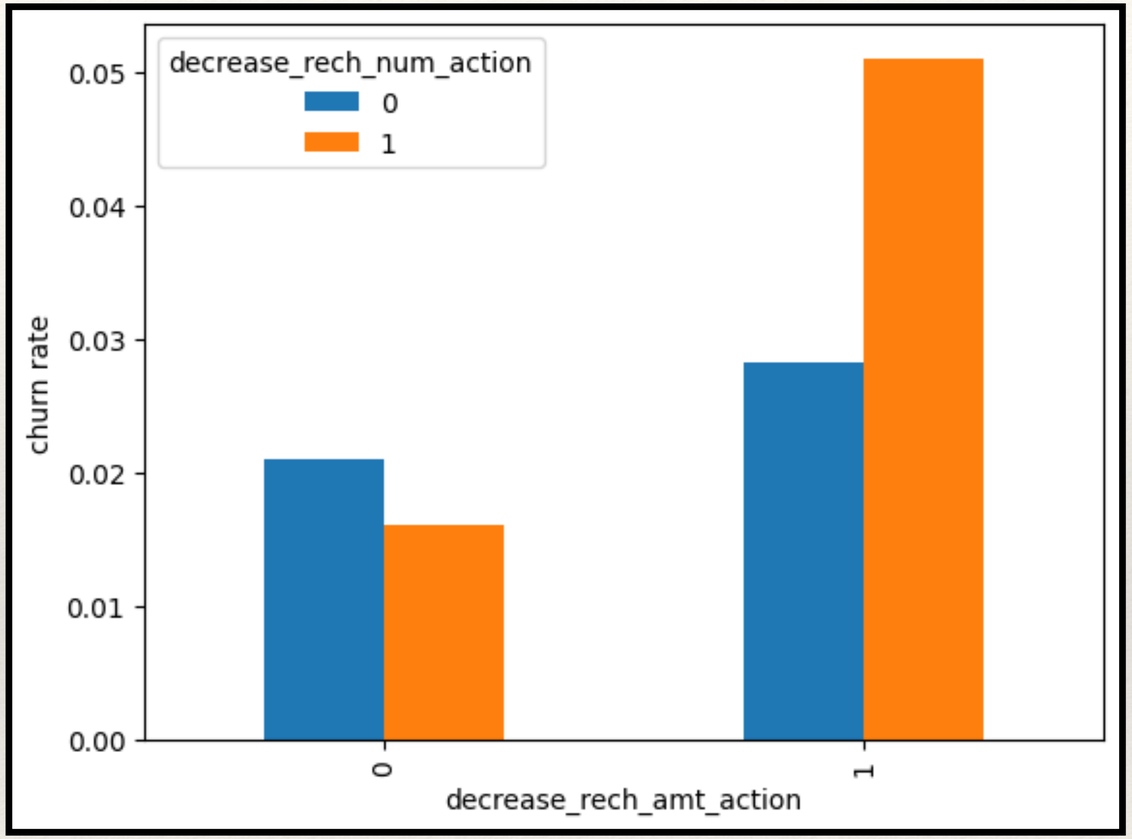
- From the above graph, it can be concluded that the minutes of usage of the churn customers are mostly populated in the 0 to 2500 range. Higher the MOU, the lesser the churn probability.



Bivariate Analysis

Decreasing recharge amount and number of recharge in action phase

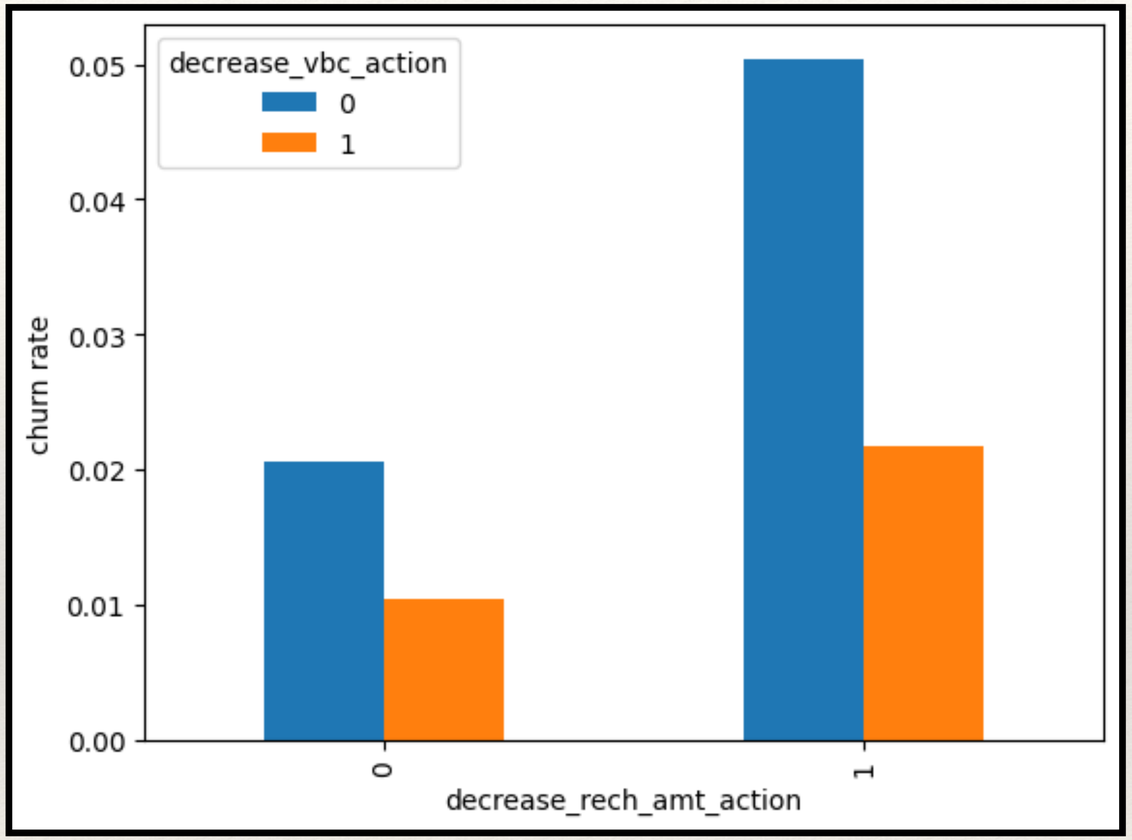
- From the above graph, it can be concluded that the churn rate is more for the customers, whose recharge amount as well as the number of recharges have decreased in the action phase as compared to the good phase.



Bivariate Analysis

Decreased in recharge amount and volumne based cost in action phase

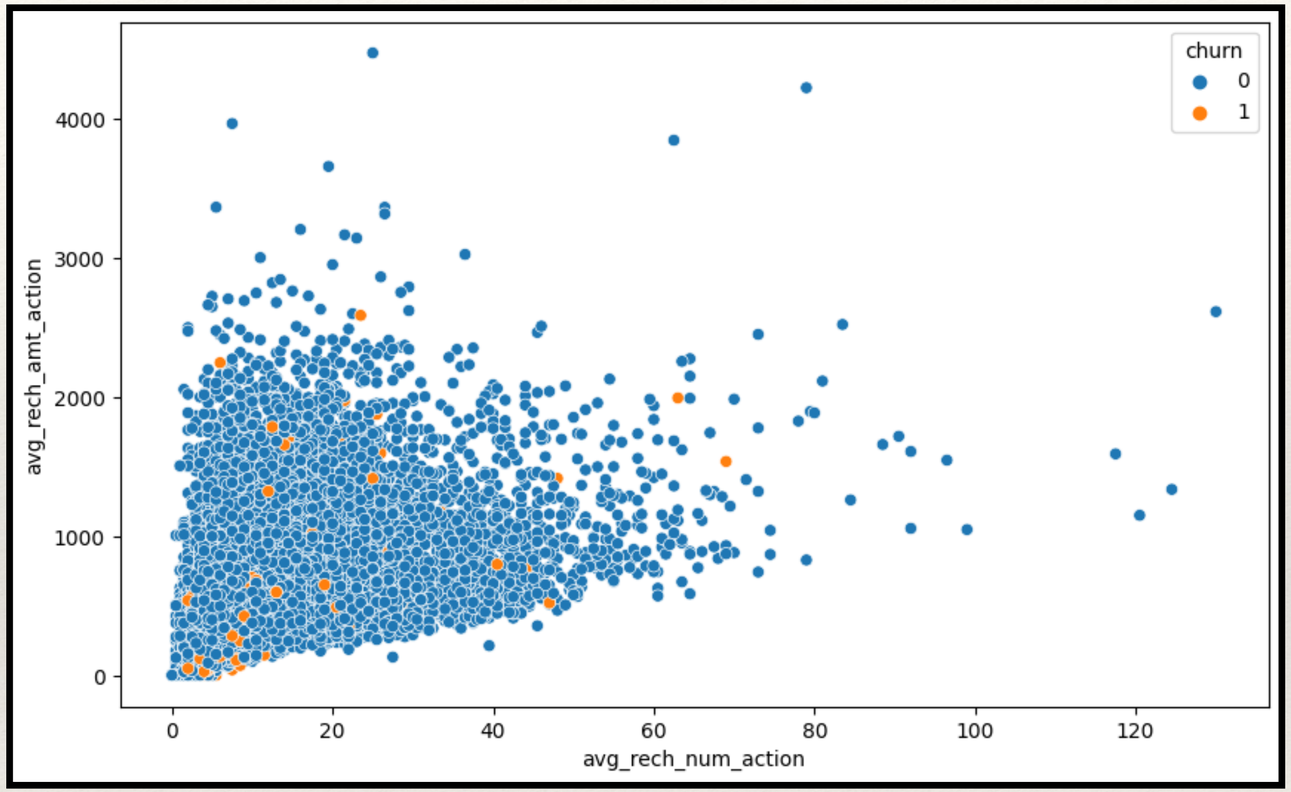
- From the above graph, it can be concluded that the churn rate is more for the customers whose recharge amount is decreased along with the volume-based cost increase in the action phase.



Bivariate Analysis

Recharge amount and number of recharge in action month

- From the above graph, it can be concluded the recharge number and the recharge amount are mostly proportional. More the number of recharges, the more the amount of the recharge.

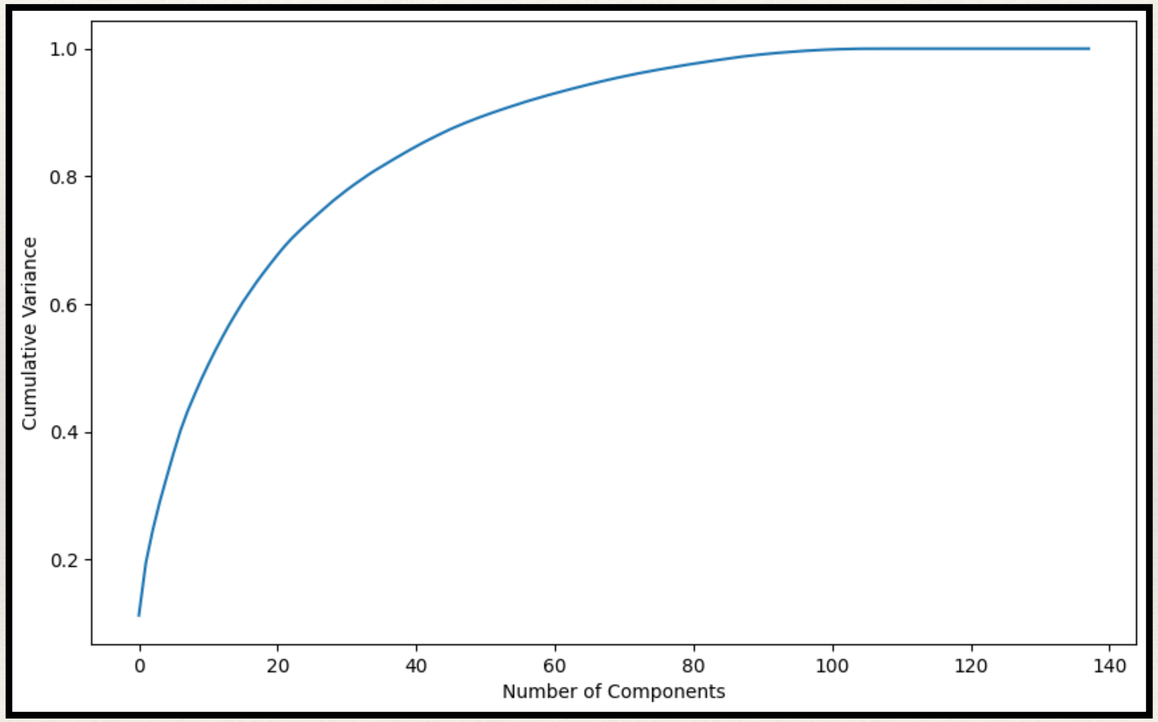


Post data cleanup

- From our above analysis of the graph and data we can drop the derived columns as they are not needed for analysis which include total_mou_good', 'avg_mou_action', 'diff_mou', 'avg_rech_num_action' etc.
- So now after dropping all the non necessary information we are finally left with 27705 rows and 140 columns
- We can see that we are left with only 27.7% of the initial value but these are the high value customers which bring majority of the revenue
- Now we can proceed by doing a train test split of 20-80 to build the model

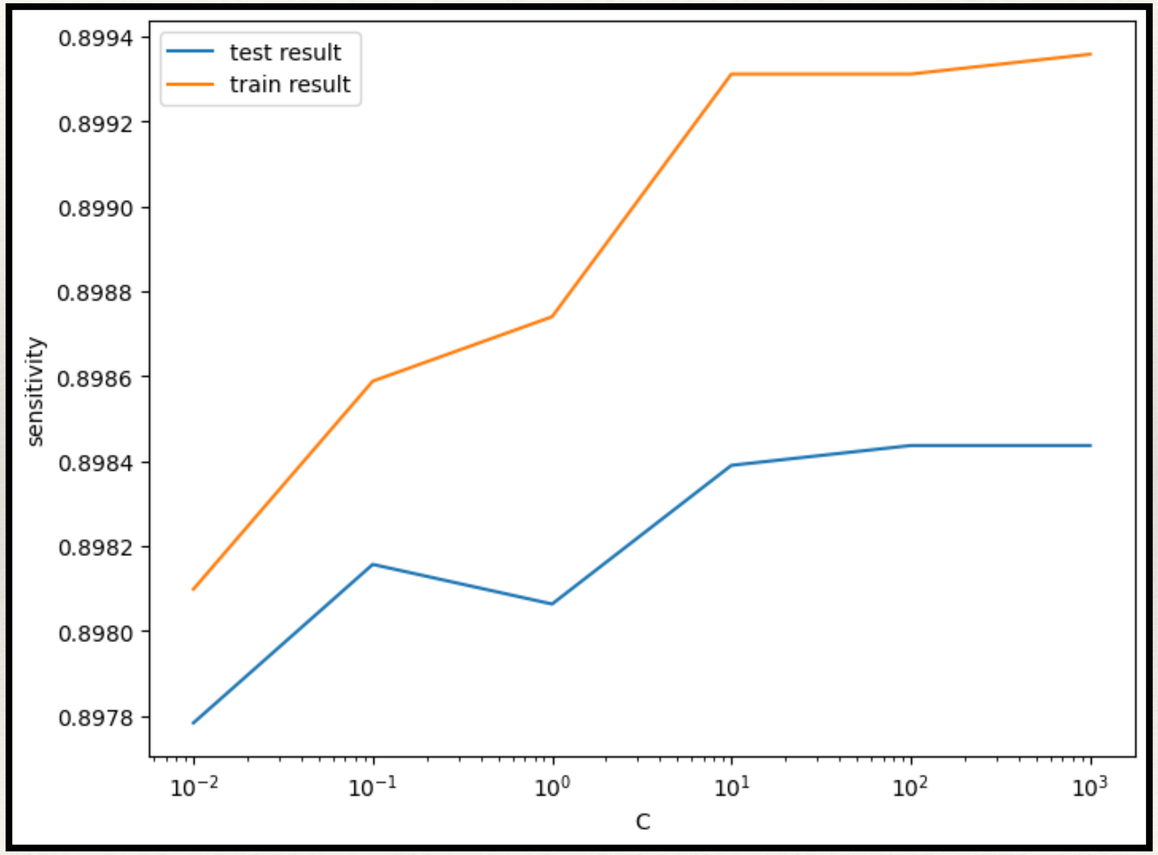
Model Building - 1 (Model with PCA)

- First we will build our model with the PCA module and will chose our random state to be 42
- By plotting the scree plot we can see that 60 components explain almost more than 90% variance of the data. So, the PCA can be performed on 60 components.
- Hence we need to perform PCA with 60 components
- Also our focus is more on higher Sensitivity/Recall score than the Accuracy. This is because we care more about the churn cases than the not churn cases. The main goal is to retain the customers, who have possibility to churn. There should not be a problem, if we consider few not churn customers as churn customers and provide them some incentives for retaining them. Hence, the sensitivity score is more important here.



Logistic Regression with PCA

- From our graph we can observe that the highest test sensitivity is 0.898 at C = 100
- After building our model using this c value we can see that our metrics for the model are
- Train Set:
  - Accuracy = 0.86
  - Sensitivity = 0.89
  - Specificity = 0.83
- Test Set:
  - Accuracy = 0.83
  - Sensitivity = 0.81
  - Specificiy = 0.83



Decision Tree with PCA

- By performing hyper parameter tuning we can infer that
- The best sensitivity is present at :- 0.9007234539089849
- And the best DecisionTreeClassifier for our data would have the following features
  - max_depth=10,
  - min_samples_leaf=50
  - min_samples_split=100

Decision Tree with PCA

- After finishing the model we can get the confusion matrix and do predictions on the data which provides us with a model summary of
- Train Set:
  - Accuracy = 0.90
  - Sensitivity = 0.91
  - Specificity = 0.88
- Test Set:
  - Accuracy = 0.86
  - Sensitivity = 0.70
  - Specificity = 0.87

Random Forest

- By performing hyper parameter tuning we can infer that
- The best AUC is 0.9990003390296671 and the Best hyperparameters are {'criterion': 'entropy', 'max_features': ‘auto'}
- Now doing predictions we find out that the values are
  - Sensitivity: 0.58
  - Specificity: 0.97
  - AUC: 0.96

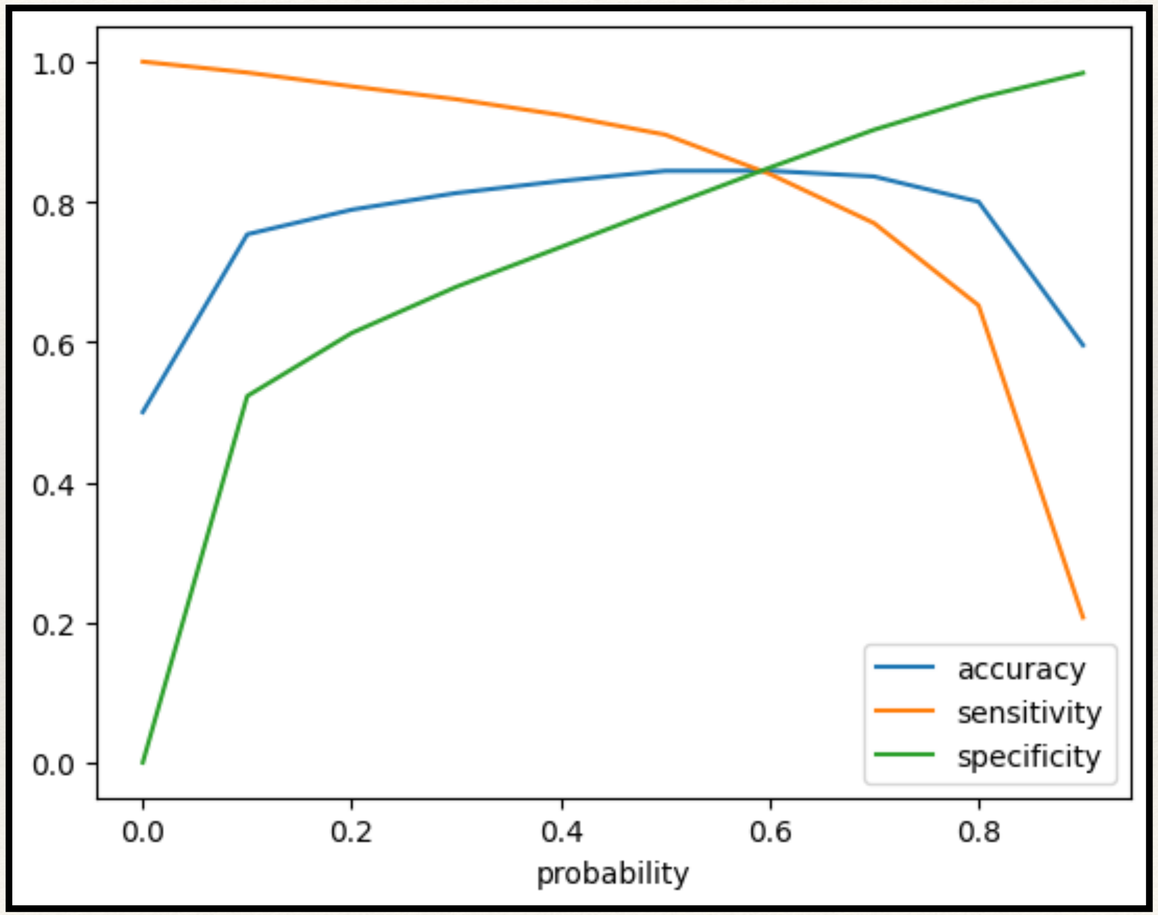
Model Building - II(Model with no PCA)

Logistic Regression with no PCA

- From the analysis, we can see that there are a few features that have positive coefficients and a few that have negative ones.
- Many features have higher p-values and hence became insignificant in the model.
- Now we will build a RFE with a step size of 15
- Now we will check and eliminate all columns with a p value of more than 0.05 and VIF greater than 5

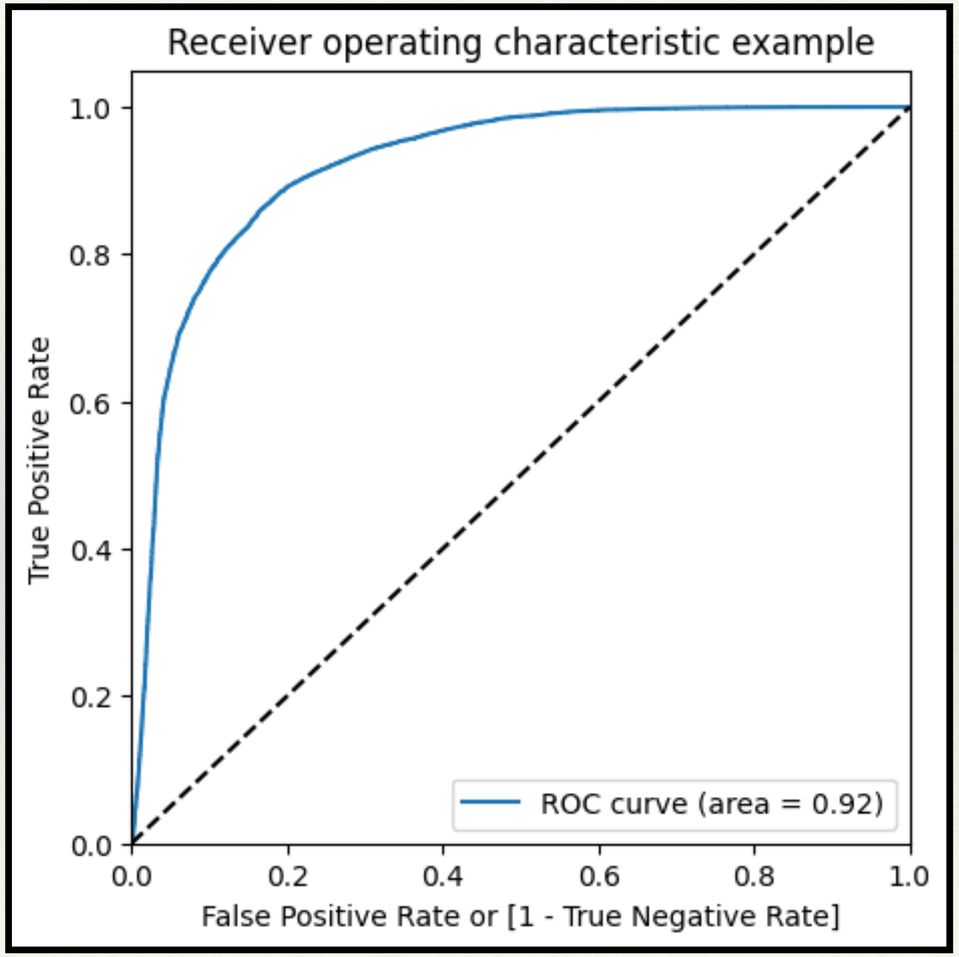
Model Performance on Train Set

- Now we will create a dataframe with the actual churn and the predicted probabilities
- By using multiple cutoff values we can calculate multiple values of accuracy sensitivity and specificity for our purpose
- From which we can conclude that
- Accuracy - is stable around 0.6
- Sensitivity - Decreases with the increased probability
- Specificity - Increases with the increasing probability.
- At point 0.6 where the three parameters cut each other, we can see a good balance, but since we want to achieve better sensitivity than accuracy and specificity, hence we are taking 0.5 to achieve higher sensitivity, which is our main goal.



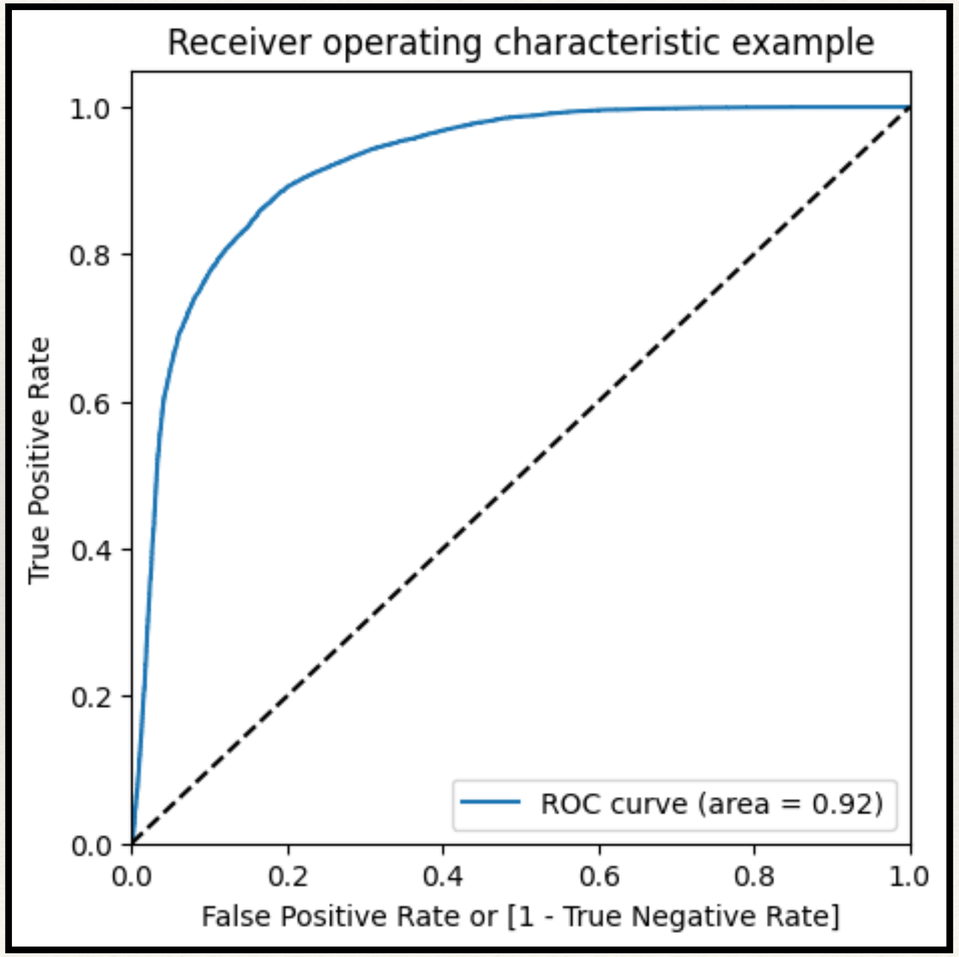
Model Performance on Train Set

- According to our chosen cutoff the values that we get are
- Accuracy:- 0.84
- Sensitivity:- 0.89
- Specificity:- 0.79
- We have got good accuracy, sensitivity and specificity on the train set prediction.
- Also plotting the ROC Curve which is the trade off between sensitivity & specificity we can see that We can see the area of the ROC curve is 0.92 which is very close to 1, which is the Gini of the model.



Model Performance on test set

- After preparing the confusion matrix on the test set we can observe that
- Accuracy = 0.78
- Sensitivity = 0.99
- Specificity = 0.12
- We have got a good model where our test results are very close to the train result



Business Recommendation

- The variables in the logistic regression model that are top predictors are as follows:
  - loc_ic_mou_8
  - og_others_7
  - ic_others_8
  - isd_og_mou_8
  - decrease_vbc_action
  - monthly_3g_8
  - std_ic_t2f_mou_8
  - monthly_2g_8
  - oc_ic_t2f_mou_8
  - roam_og_mou_8

Conclusion

- From the analysis, we found out the fact that
- Std Outgoing Calls and Revenue Per Customer are strong indicators of Churn,
- similarly Incoming and Outgoing Calls for the 8th month and the average revenue in the 8th month are the most important columns to predict churn.
- The customers with a tenure of less than 4 years are more likely to churn.
- Max Recharge Amount is a strong feature to predict churn.

Business Recommendations

- Target customers whose minutes of usage of incoming local calls and outgoing ISD calls are less in the action phase, mostly in August.
- Target customers whose outgoing other charges in July and incoming others in August are less.
- Customers having value-based costs in the action phase increased are more likely to churn. Hence, these customers may be a good target to provide offers.
- Customers whose monthly 3G recharge in August is more, are likely to be churned.
- Customers having to decrease STD incoming minutes of usage for operators T to fixed lines of T for the month of August are more likely to churn.
- Customers decreasing monthly 2G usage for August are most probable to churn.
- Customers having to decrease incoming minutes of usage for operators T to fixed lines of T for August are more likely to churn.
- Customers whose roaming outgoing minutes of usage are increasing are more likely to churn. The roam_og_mou_8 variables have positive coefficients (0.7135).
- These recommendations can help the business to identify customers who are at a higher risk of churn and take proactive measures to retain them by offering relevant deals and promotions.

Thank You