WELCOME TO MY PRESENTATION

By Vaishali kesharwani

TELECOM CHURN CASE STUDY

PROBLEM STATEMENT

- 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 od a leading telecom firm build predictive models to identify customers at high risk of churn and identify the main indicator of churn.
- Retaining high profitable customers is the main business goal here .

STEPS WE CAN FOLLOW

- 1. Reading, understanding and visualizing the data.
- 2. preparing the data for modelling.
- 3. building the model
- 4.Evaluate the model

DATA PREPARATION

- **Reading and understanding the data.**
- **Handling missing values.**
- 1. Handling missing values in columns
- 2. Deleting the data column as the date column are not required in our analysis.
- 3. Dropping circle_id column as this column has only one unique value. Hence there till be no impact of this on the data analysis.
- **Filter high value customers.**
- 1. Creating columns avg_rech_amt_6_7 by summing up total recharge amount of month 6 and 7. then taking the avg of the sum.
- 2. Finding the 70^{th} percentile of the avg_reach_amt_6_7
- 3.Filters the customers, who have recharge more than or equal to X.

- **Handling missing values in row.**
- 1. Looks like MOU for all the type of call for the month of September(9) have missing value to gether for any particular record.
- 2. Lets check the records for the MOU for sep(9), in which these column have missing value together.
- 3. Looks like MOU for all the type of call for the month of Aug(8),have missing value to gether for any particular record.
- 4. lets check the records for the MOU for Aug(8), in which these column have missing value together
- 5. Looks like MOU for all the type of call for the month of July(7),have missing value to gather for any particular record.
- 6. Lets check the records for the MOU for July(7),in which these column have missing value together
- 7. Looks like MOU for all the type of call for the month of Jun(6),have missing value to gather for any particular record.
- 8. Lets check the records for the MOU for Jun(6),in which these column have missing value together
- We can see that we have lost almost 7% records. But we have enough number of records to do our analysis.

TAG CHURNERS

- Delete all the attributes correspoing to the churn phase.
- Checking churn percentage ----- 3.39
- There is very little percentage of churn rate. We will take care of the class imbalance later.

OUTLIER TREATMENT

In the filters dataset except mobile_number and churn column all the column are numeric types. Hence , converting mobile_number and churn datatype to object.

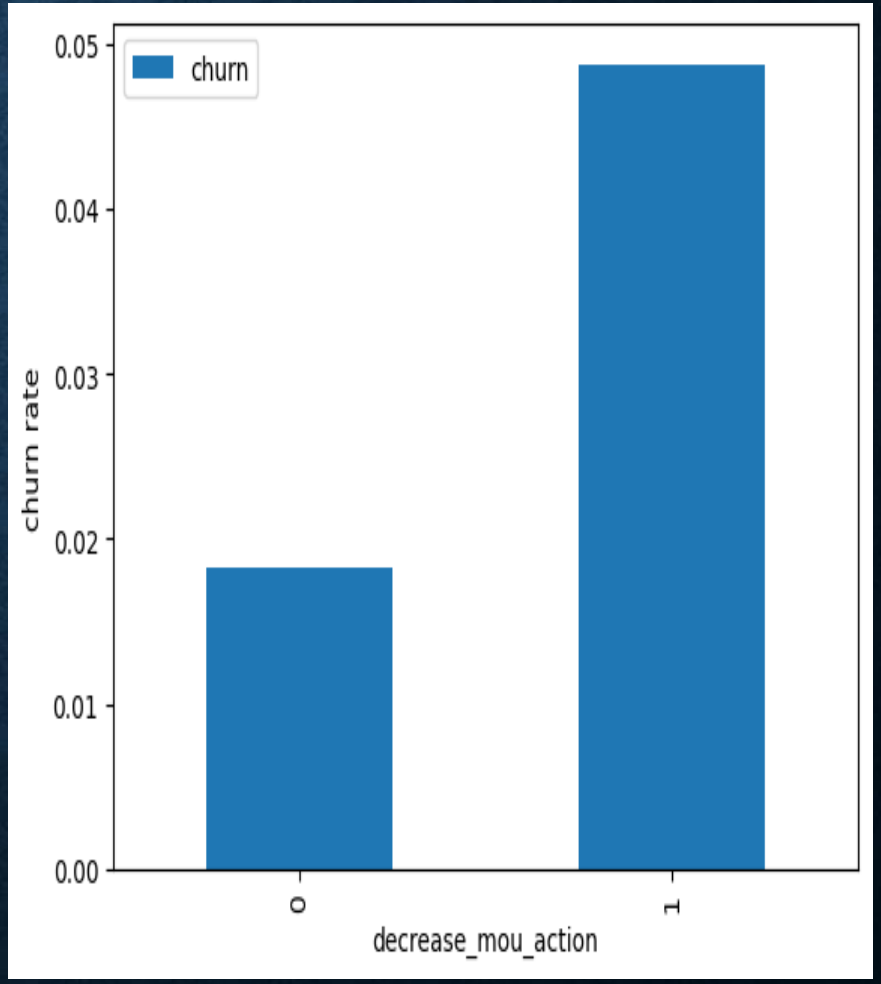
DERIVE NEW FEATURES

- Deriving new column decrease_mou_action
- Deriving new column decrease_rech_num_action
- Deriving new column decrease_rech_amt_action
- Deriving new column decrease_arpu_action
- Deriving new column decrease_vbc_action

EXPLORATORY DATA ANALYSIS [EDA]

UNIVARIATE ANALYSIS

- Churn rate on the basic whether the costumers decrease her/his MOU in action month.



- Analysis

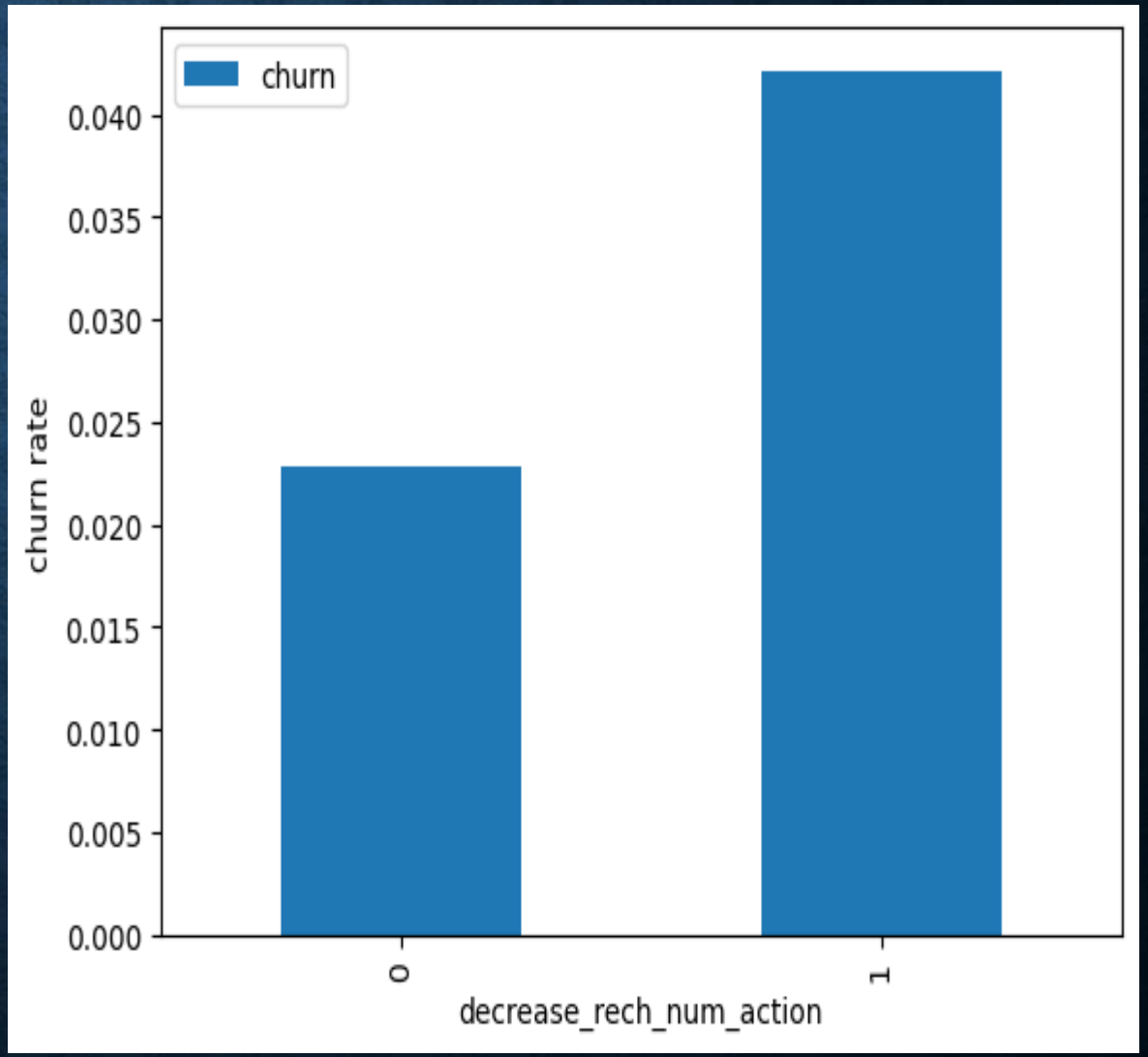
  we can see that the churn rate is more

  for the costumers, whose minute of usage

  (mou) decrease in the action phase than the good phase.

- Churn rate on the basis whether the costumers, decreased her/his number of recharge in action month.

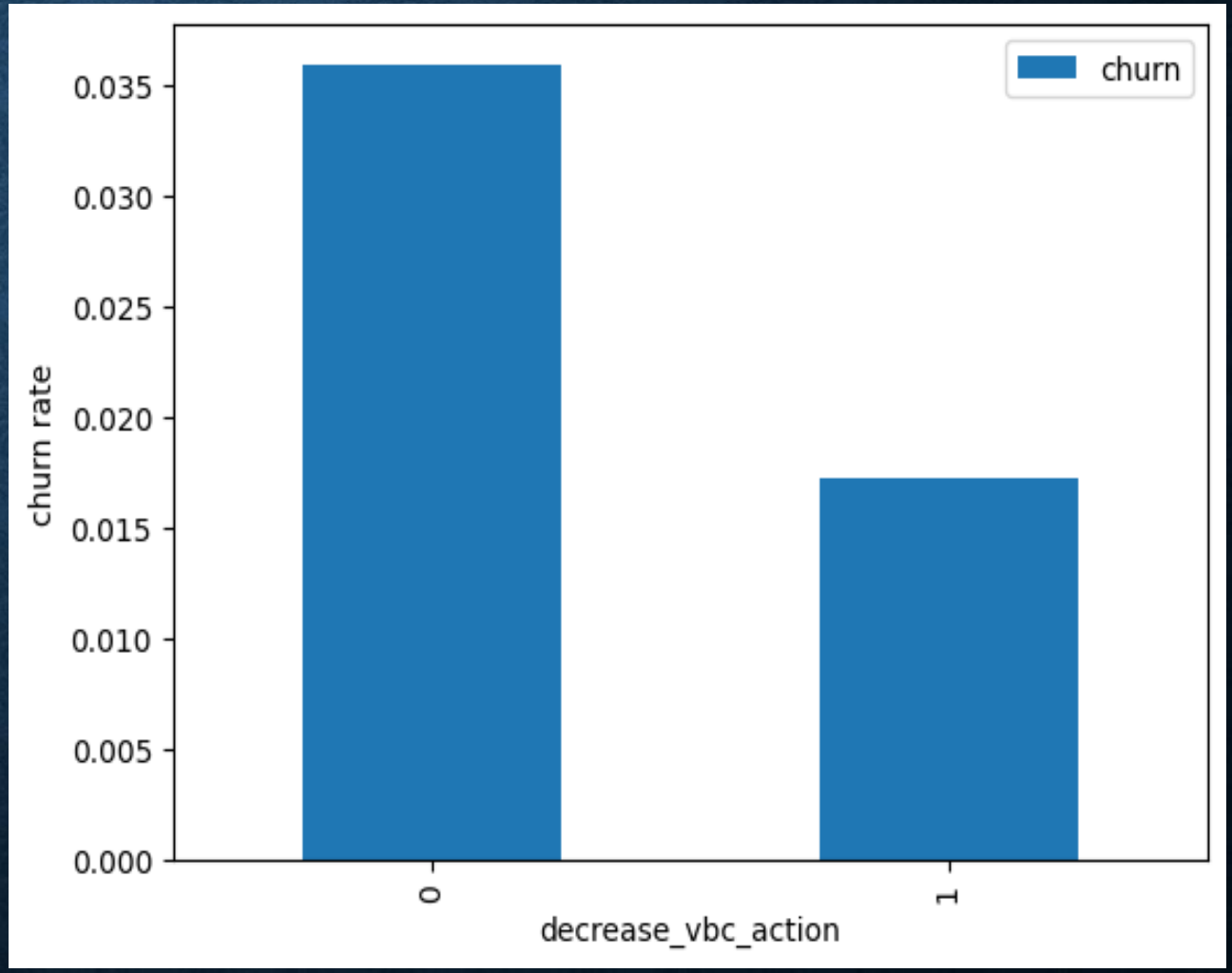
- Analysis
- As expected, the churn rate is more for the costumers whose number of recharge in the action phase is ;lesser than the number in good phase.



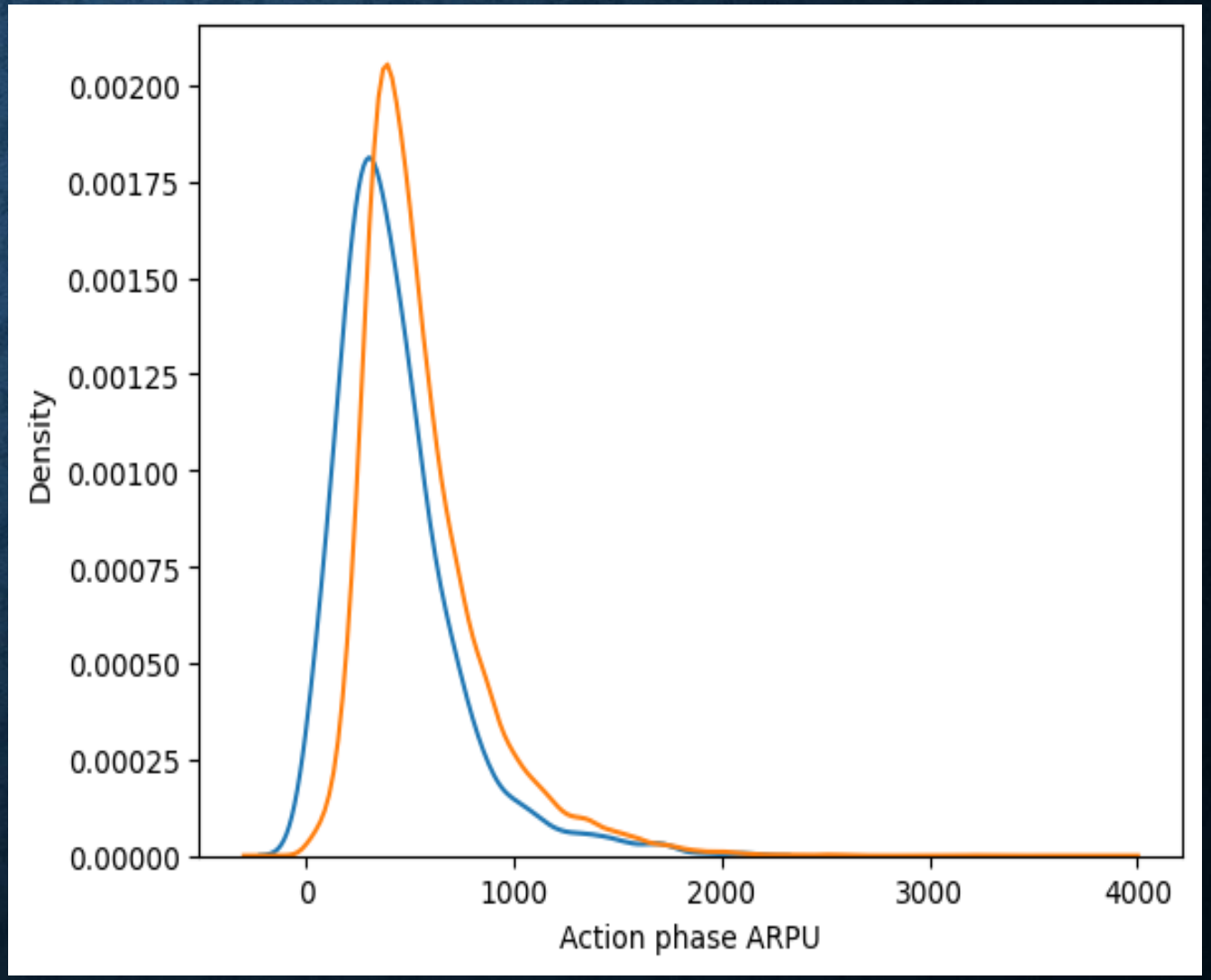
- Churn rate on the basis whether the customers decreased her/his volume based cost in action month.

- Analysis

Here we see the expected result. The column rate is more for the customers, whose volume based cost in action month in increased. That means the customers do not do the monthly recharge more when they are in the action phase.



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

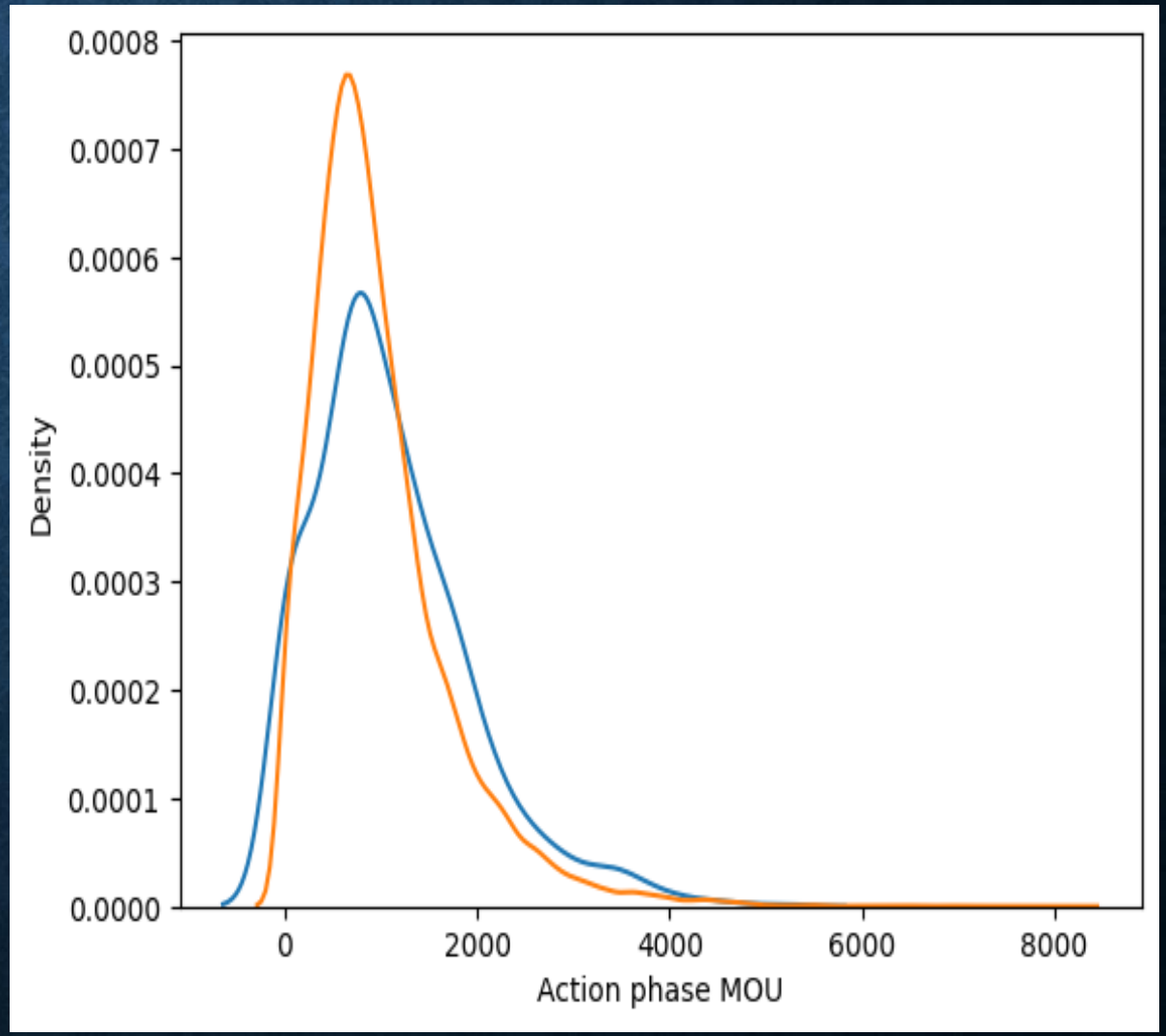


- Analysis
- Average revenue per user(ARPU) for the churn customers is mostly dense on the 0 to 900. the higher ARPU customer are less likely to be churned.

- ARPU for the not churned customers is
- mostly densed on the 0 to 1000.

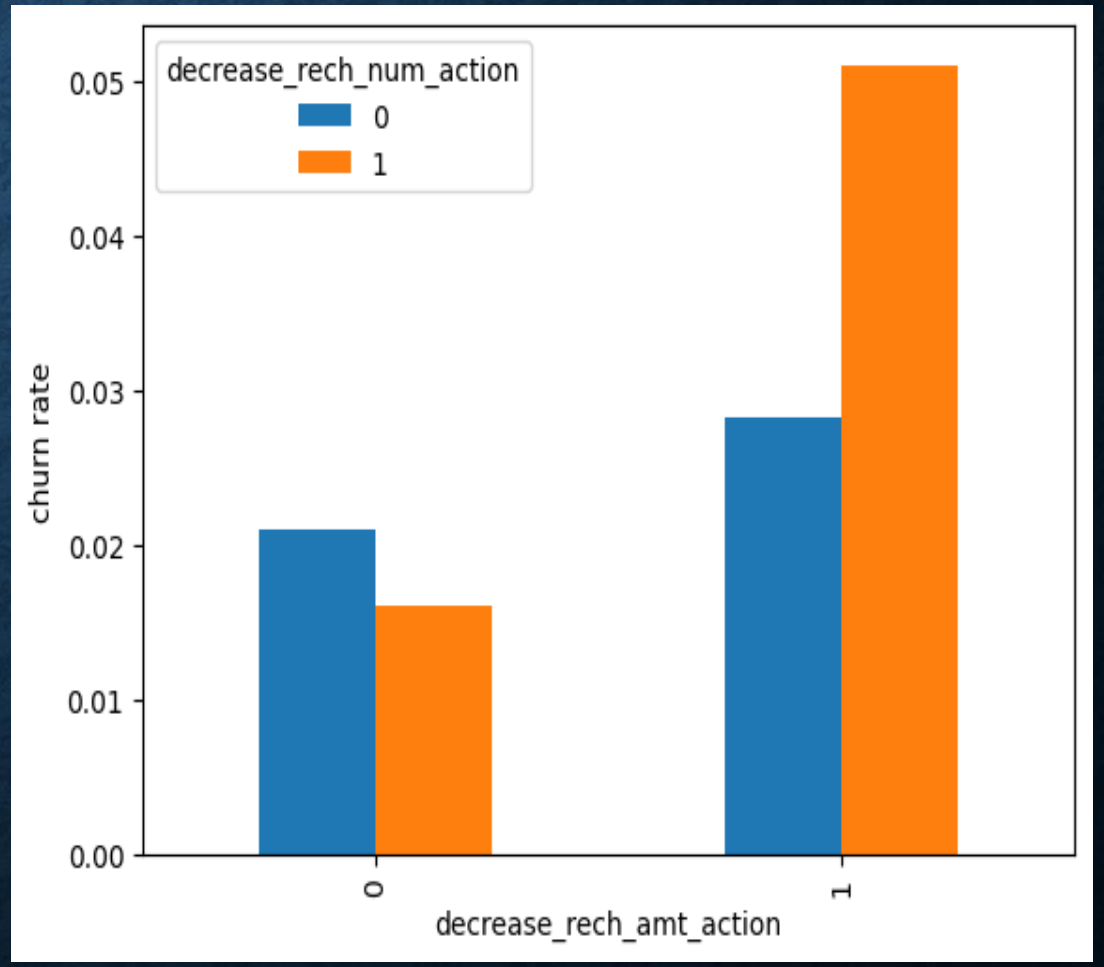
- Anakysis of the minute od usage MOU (churn and not chum) in the action phase.

- Analyse
- Minute of usage(MOU) of the churn customers is mostly populated on the 0 to 25000 rabnge. Higher the MOU, lesser the churn probability.



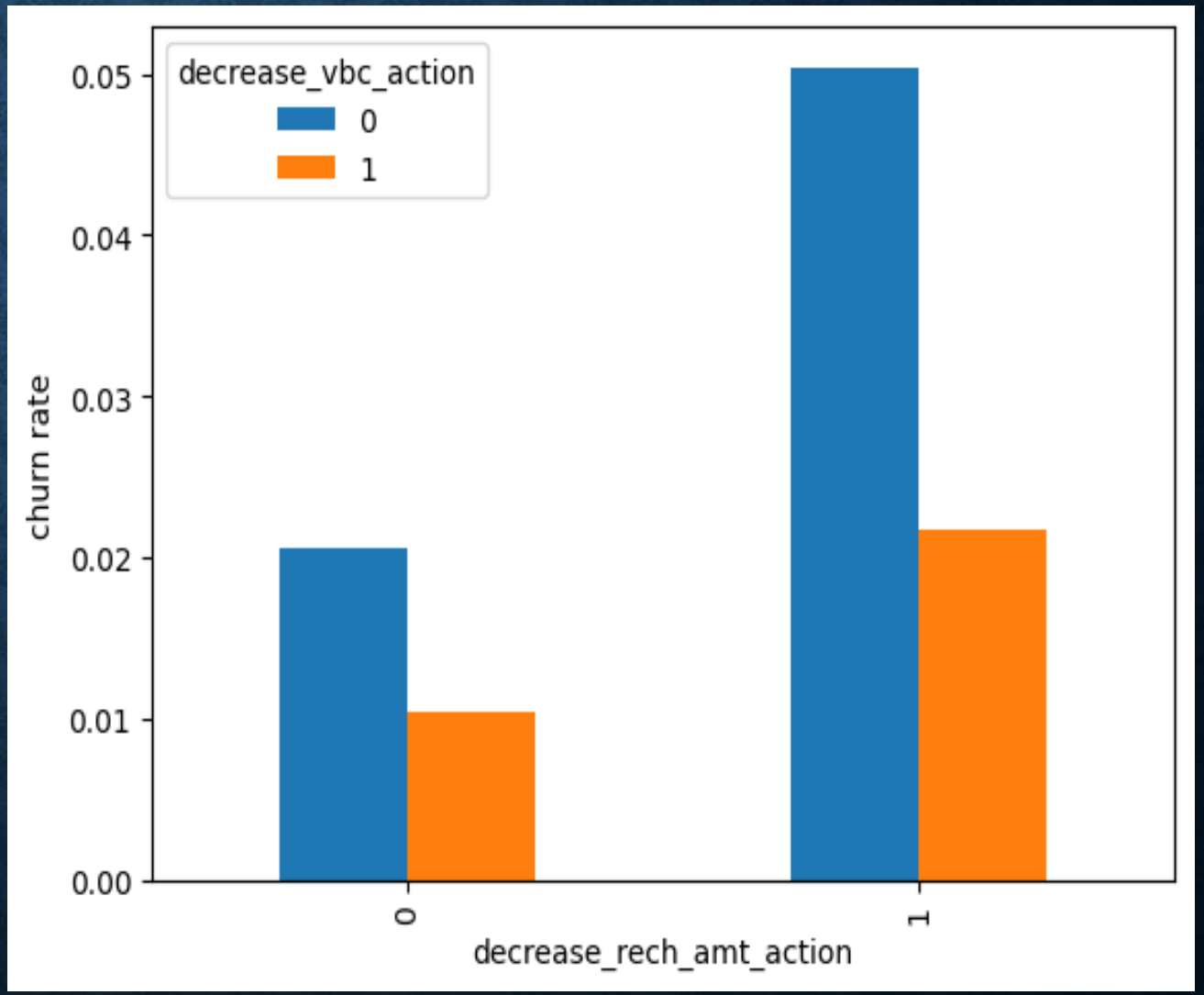
BIVARIATE ANALYSIS

- Analysis of churn rate by the decreasing recharge amount and number of recharge in the action phase.



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

- Analysis of churn rate by the decreasing recharge amount and volume based cost in the action phase.



- Analysis

we can see that the churn rate is more for the customers, whose recharge amount is decreased along with the volume based cost is increased in the action month.

TRAIN – TEST SPLIT

- Splitting data into tarin and test set 80:20.

  Dealing with data imbalance

  we are creating synthetic sample by doing up sampling using

  smote(synthetic minority oversampling technique).

  Feature scaling.

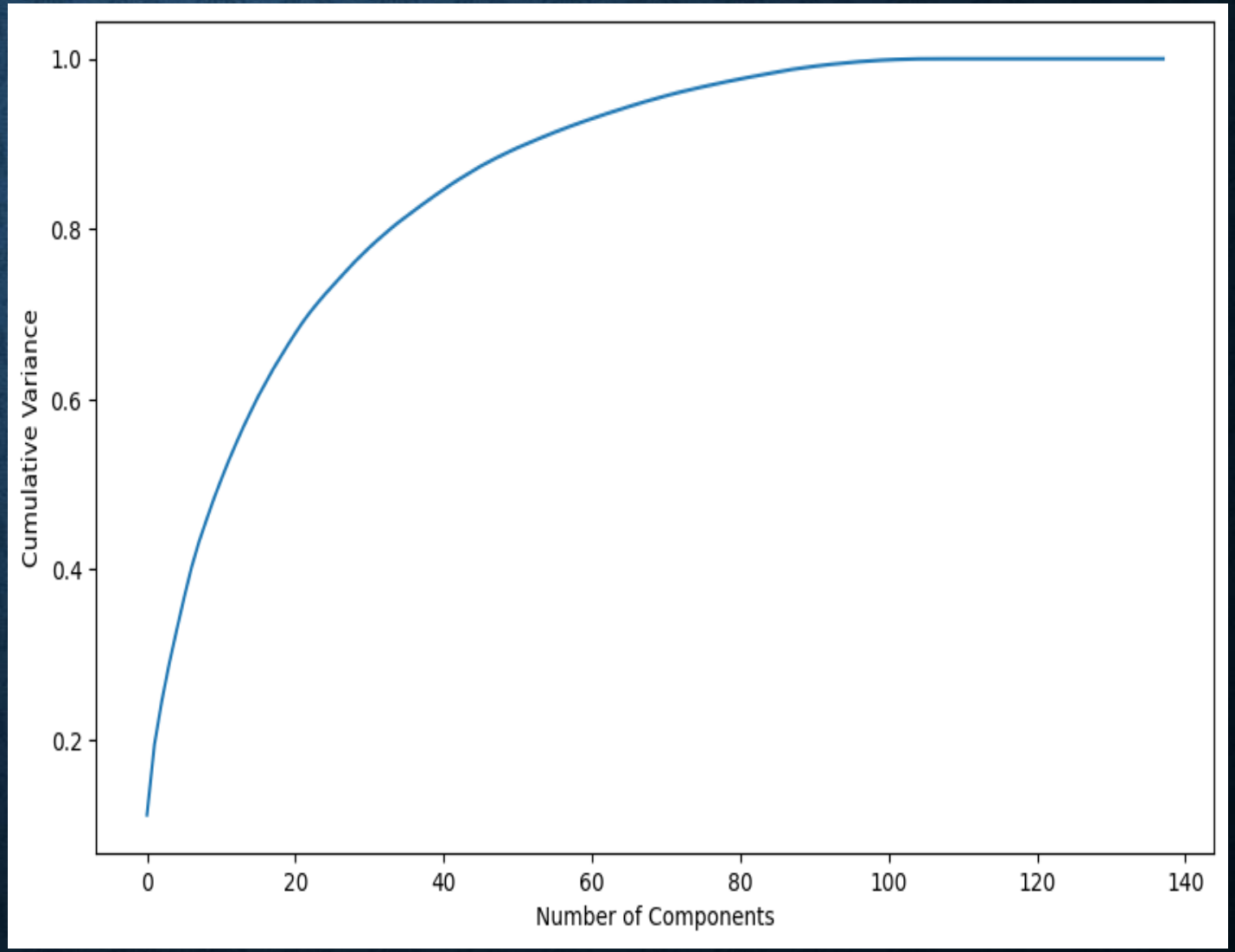
  Scaling the test set

we don’t fir scaler on the test set. We only transform the test set.

PCA(PRINCIPLE COMPONENT ANALYSIS)

We can see that 60 components explain almost more than 90% variance of the data.

So, we will perform PCA with 60 components.



- Applying transformation on the test set.

We are only doing transform in the test set not the4 fit transform. Because the fliting is already done on the train set.

So we just have to do the transformation with the already fitted data on the train set.

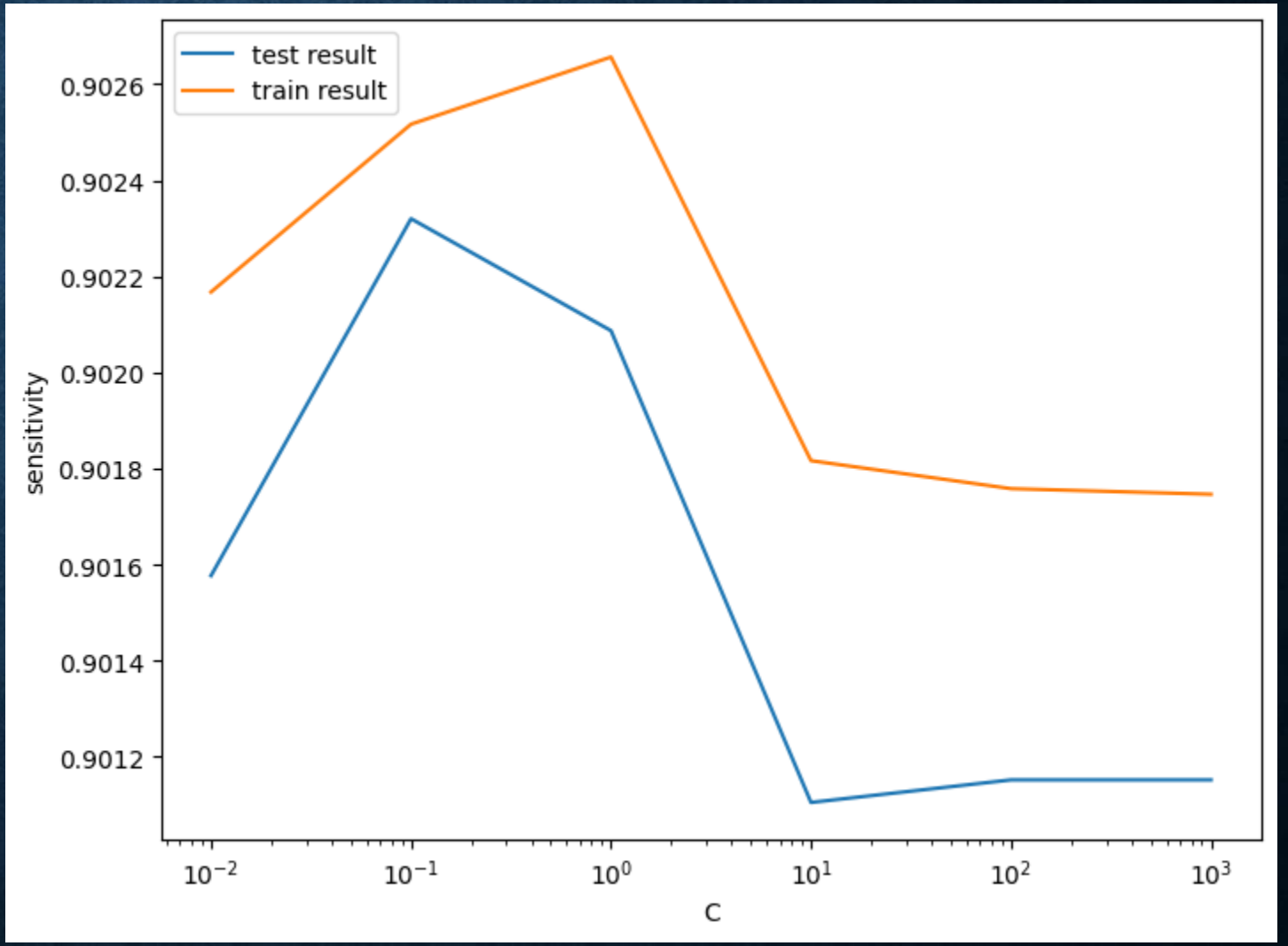
- Emphasize sensitivity/recall than accuracy

We are more focused on higher sensitivity/recall score than the accuracy.

MODEL HYPERPARAMETER TUNNING

- C is the inverse of regularization strength in logistic regression. Higher value of c correspond to less regularization.

- Logistic regression with optimal c.



PREDICTION ON THE TRAIN SET

- Confusion matrix

```
[[17886  3539
 [ 2078 19347]]
Accuracy:- 0.8689148191365228
Sensitivity:- 0.9030105017502917
Specificity:- 0.8348191365227537
```

PREDICTION ON THE TEST SET

- Confusion matrix.
- [[4443 905]
  [34 159]]

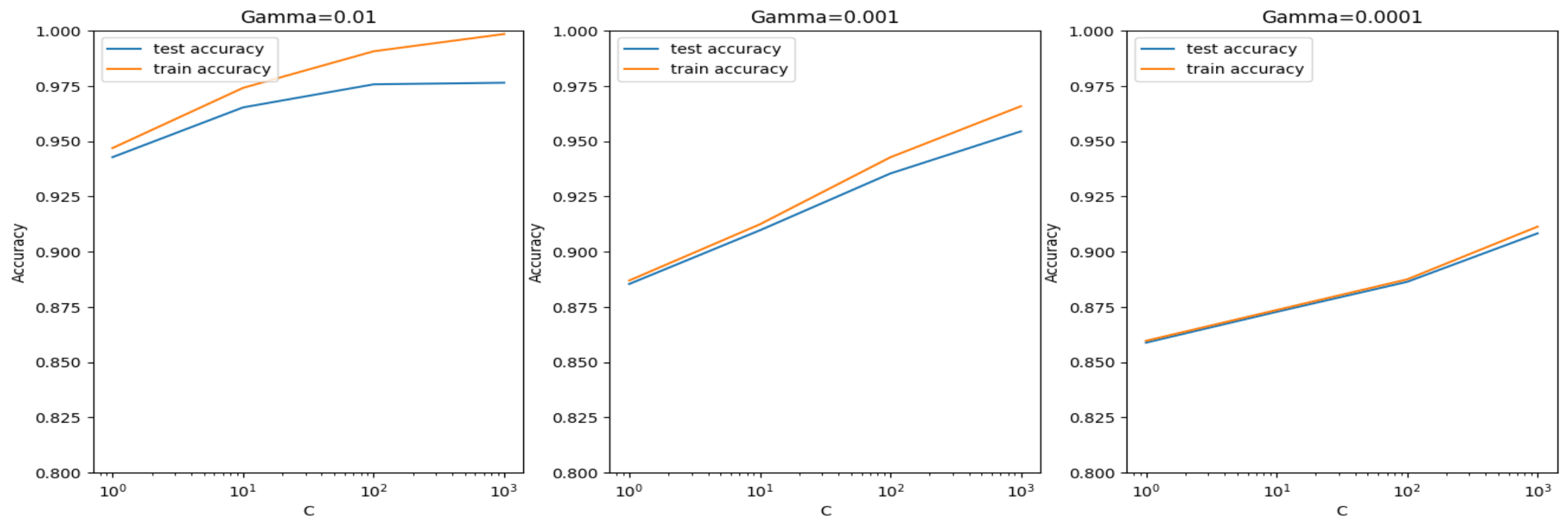
Accuracy:- 0.8305360043313481

Sensitivity:- 0.8238341968911918

Specificity:- 0.8307778608825729

SUPPORT VECTOR MACHINE(SVM) WITH PCA

- **Hyperparameter tuning**
- C:- Regularization parameter.
- gamma:- Handles non linear classifications.
- **Plotting the accuracy with various C and gamma values**



DECISION TREE WITH PCA

- Model with optimal hyperparameters
- Prediction on the train set
- Confusion matrix

[[18877 2548]

[1600 19825]]

Accuracy:- 0.9031971995332555

Sensitivity:- 0.9253208868144691

Specificity:- 0.881073512252042

Prediction on the test set

Confusion matrix

[[4606 742]

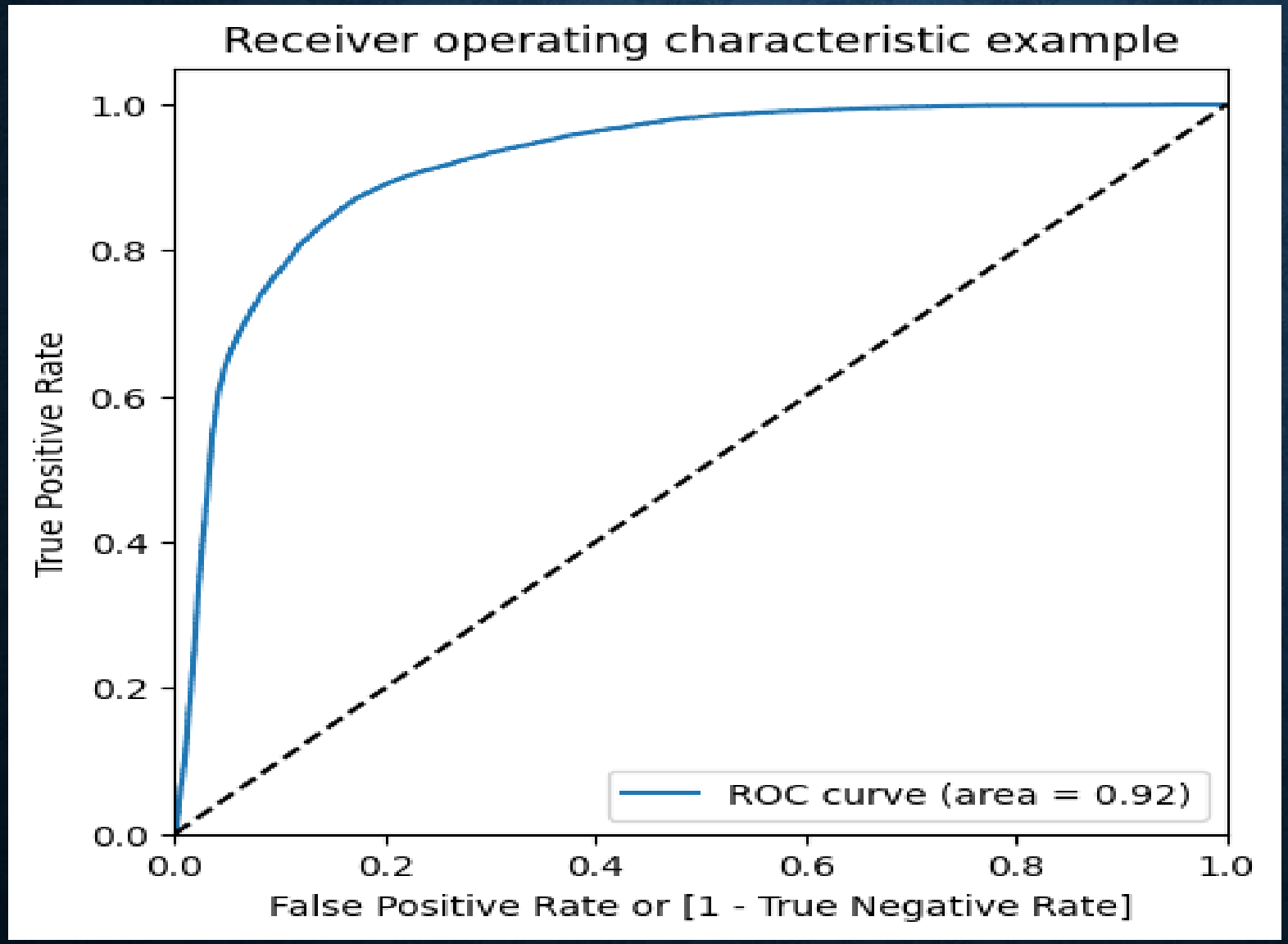
[53 140]]

Accuracy:- 0.8565240931239848

Sensitivity:- 0.7253886010362695

Specificity:- 0.8612565445026178

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



- **Final conclusion with no PCA**

- 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. Hence, the model is more relevant in terms of explaining to the business.

BUSINESS RECOMMENDATION

- **Top predictors**
- Below are few top variables selected in the logistic regression model.
- We can see most of the top variables have negative coefficients. That means, the variables are inversely correlated with the churn probability.

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
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

- If the local incoming minutes of usage (loc_ic_mou_8) is lesser in the month of August than any other month, then there is a higher chance that the customer is likely to churn.

RECOMMENDATION

- 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.
- Customers, whose monthly 3G recharge in August is more, are likely to be churned.
- Customers having decreasing 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 decreasing incoming minutes of usage for operators T to fixed lines of T for August are more likely to churn.
- roam_og_mou_8 variables have positive coefficients (0.7135). That means for the customers, whose roaming outgoing minutes of usage is increasing are more likely to churn.

THANK YOU !