Telecom Customers Churn Assessment Project

Manish Dabhade
Samanth Reddy
Koushik

Business Problem Overview:

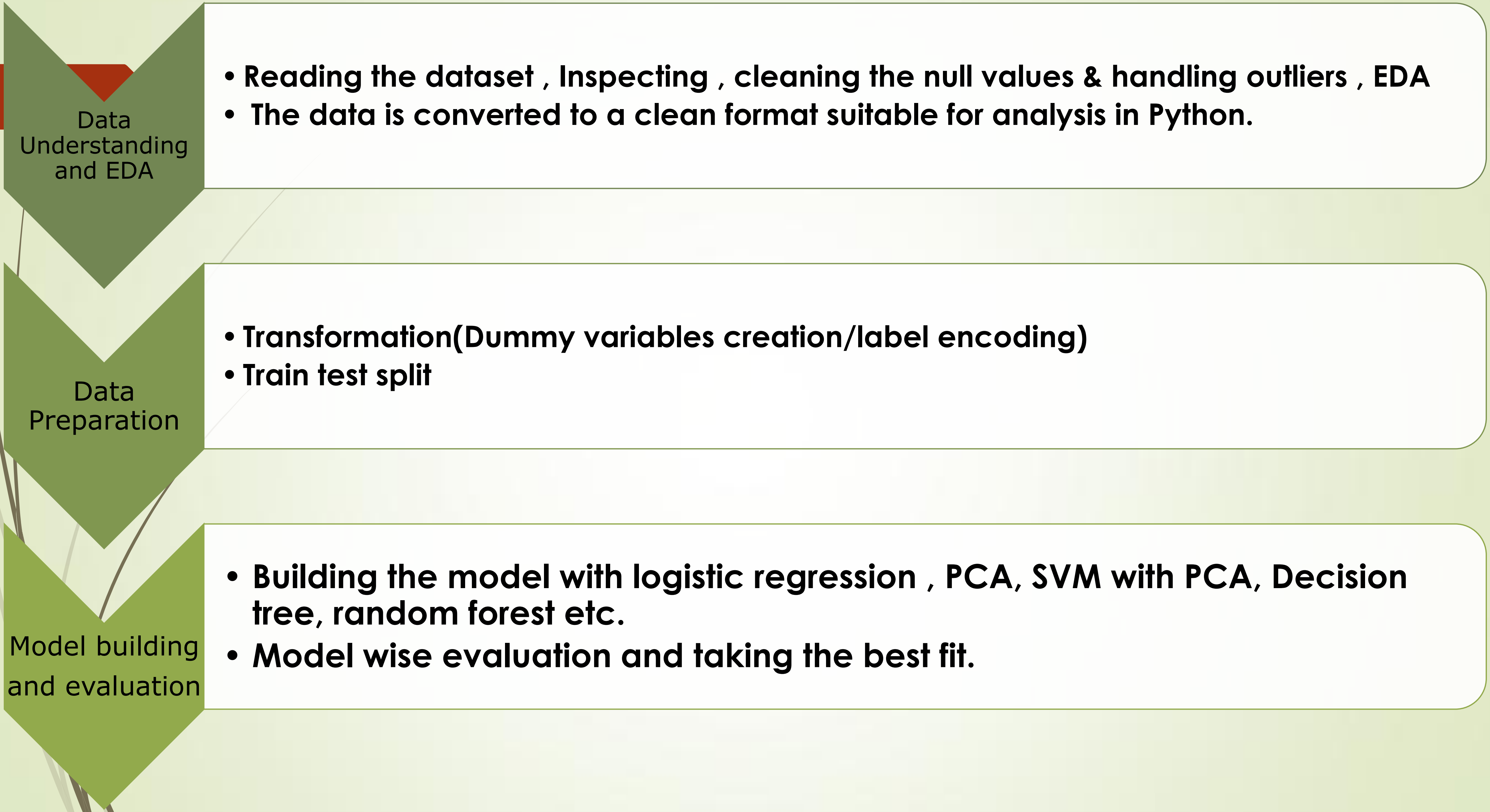
In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another. In this highly competitive market, the telecommunications industry experiences an average of 15-25% annual churn rate. Given the fact that it costs 5-10 times more to 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 analyse 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.

Understanding the business objective:

- *The business objective is to predict the churn in the last (i.e. the ninth) month using the data (features) from the previous three months(i.e 6,7,8).*

Steps Followed

- Reading, understanding, cleaning and visualizing the Data Set.
- Preparing the data for modelling.
- Building the Model.
- Evaluating the Model.



Data Understanding and EDA

- Reading the dataset , Inspecting , cleaning the null values & handling outliers , EDA
- The data is converted to a clean format suitable for analysis in Python.

Data Preparation

- Transformation(Dummy variables creation/label encoding)
- Train test split

Model building and evaluation

- Building the model with logistic regression , PCA, SVM with PCA, Decision tree, random forest etc.
- Model wise evaluation and taking the best fit.

Data Understanding and EDA

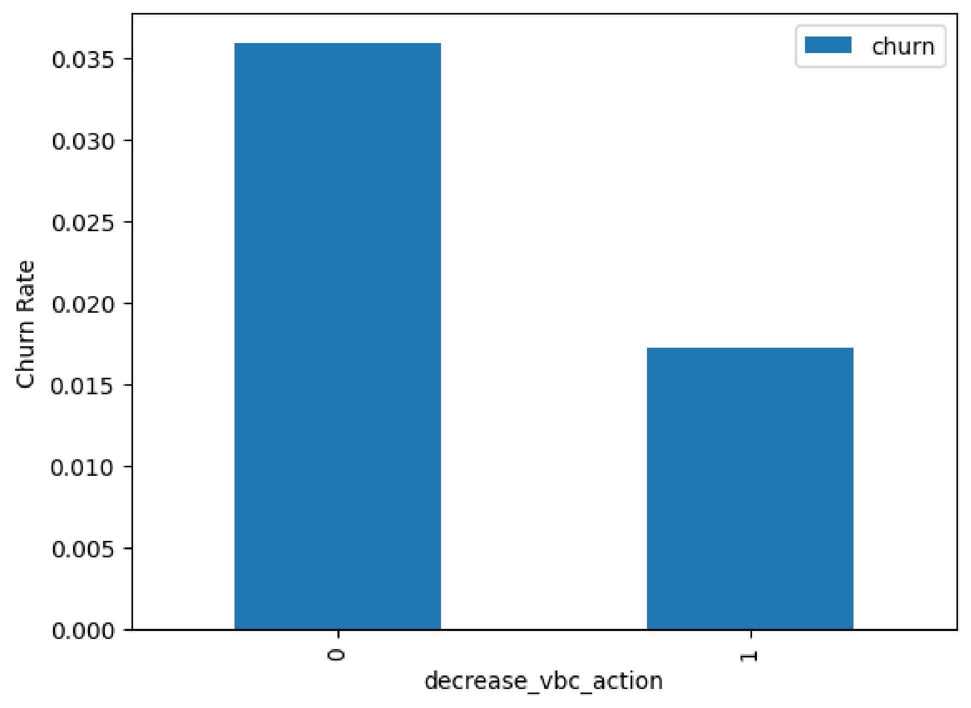
- Imported all the necessary libraries.
- The dataset consists of 99999 Rows and 226 columns.
- Dropped the columns having missing values of more than 30%. After which we are left with 186 columns.
- Dropped different types of date columns and we are left with 178 columns.
- Identified high valued customers by average recharge amount for 6th and 7th month combined and separated the 70th percentile which came out to be Rs.368.5
- Only taken the customers who are above recharging with a minimum of this Rs.368.5 and we are left with 30011 rows.
- Further deleted the rows and columns which have high missing values.
- Created the data frame consisting minutes of usage for the 9th month as null and dropped them. Dropped the columns with missing values in minutes of usage for 8th, 7th and 6th months also.
- Finally we are left with 27991 rows which is approximately 93% of the valued customers data.

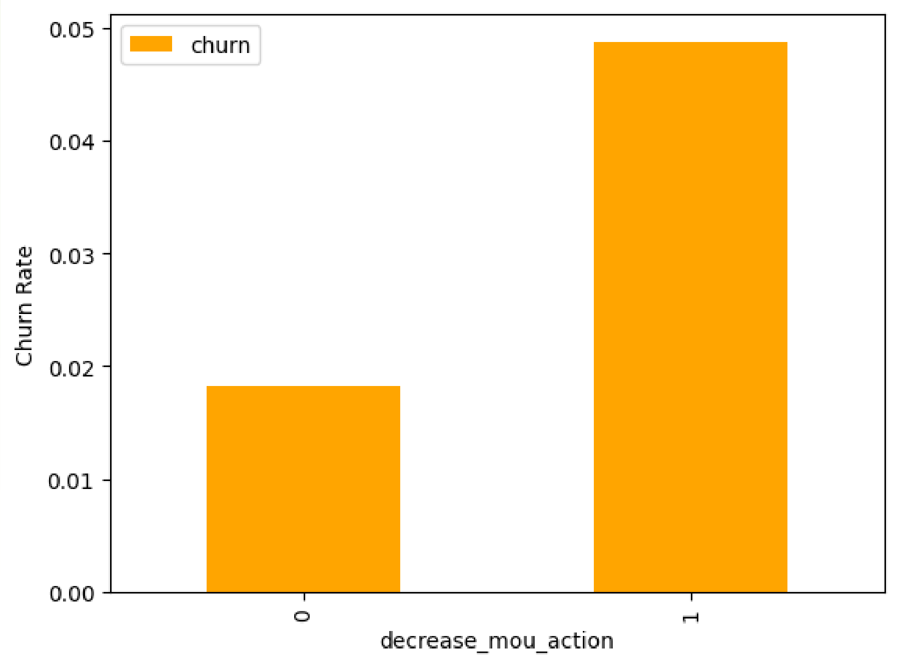
Churned Customers

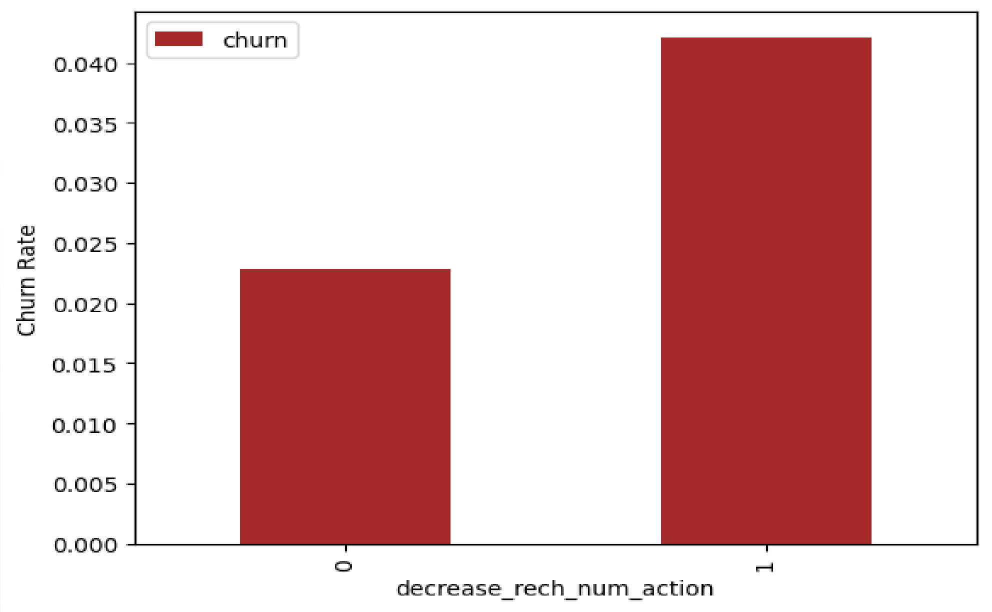
- Churned customers are picked from the 9th month data by the convention that churn = 0 is the ones who made call or used data, churn = 1 the ones who have not.
- Deleted all the attributes which are related to churn month that are ending with 9 and then drop those columns.
- Identified the numeric columns and removed the outliers below 10percentile and above 90percentile and left with 27705 rows and 136 columns.
- Created extra columns by identifying the decrease in different parameters like minutes of usage, data, arpu, data, number, amount, vbc with 1s(for decrease) 0s(no decrease).

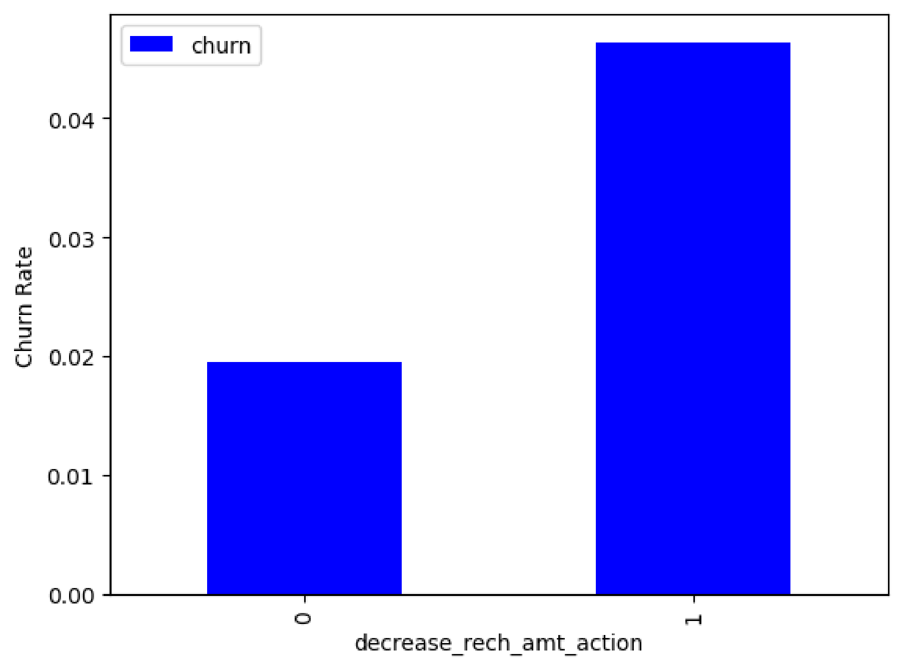
Exploratory Data Analysis (Univariate)

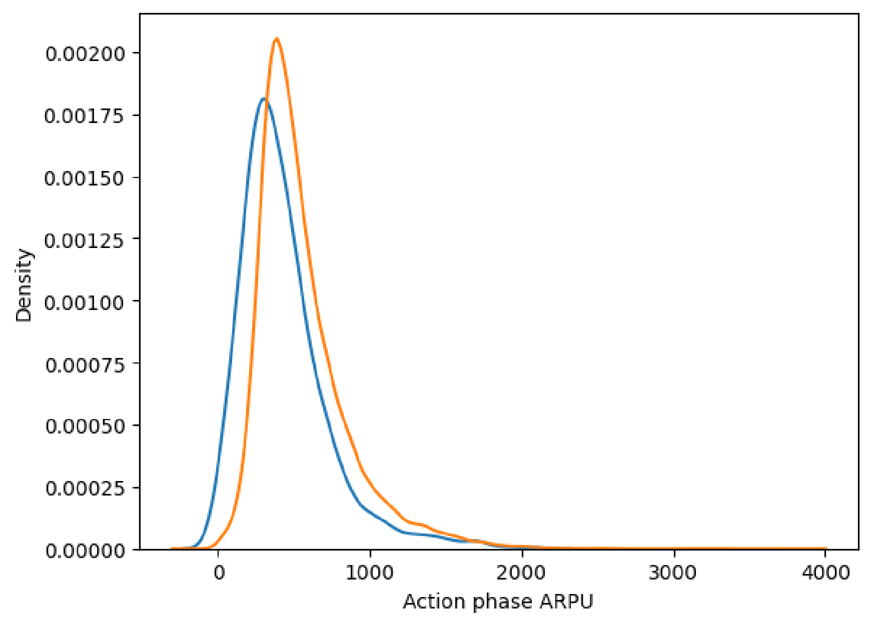
- Churn rate is more for the customers, whose minutes of usage(mou) decreased in the action phase than the good phase.
- Churn rate is more for the customers, whose number of recharge in the action phase is lesser than the number in good phase.
- Churn rate is more for the customers, whose amount of recharge in the action phase is lesser than the amount in good phase.
- 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.
- Average revenue per user (ARPU) for the churned customers is mostly dense on the 0 to 900. The higher ARPU customers are less likely to be churned. ARPU for the not churned customers is mostly dense on the 0 to 1000.
- Minutes of usage(MOU) of the churn customers is mostly populated on the 0 to 2500 range. Higher the MOU, lesser the churn probability.

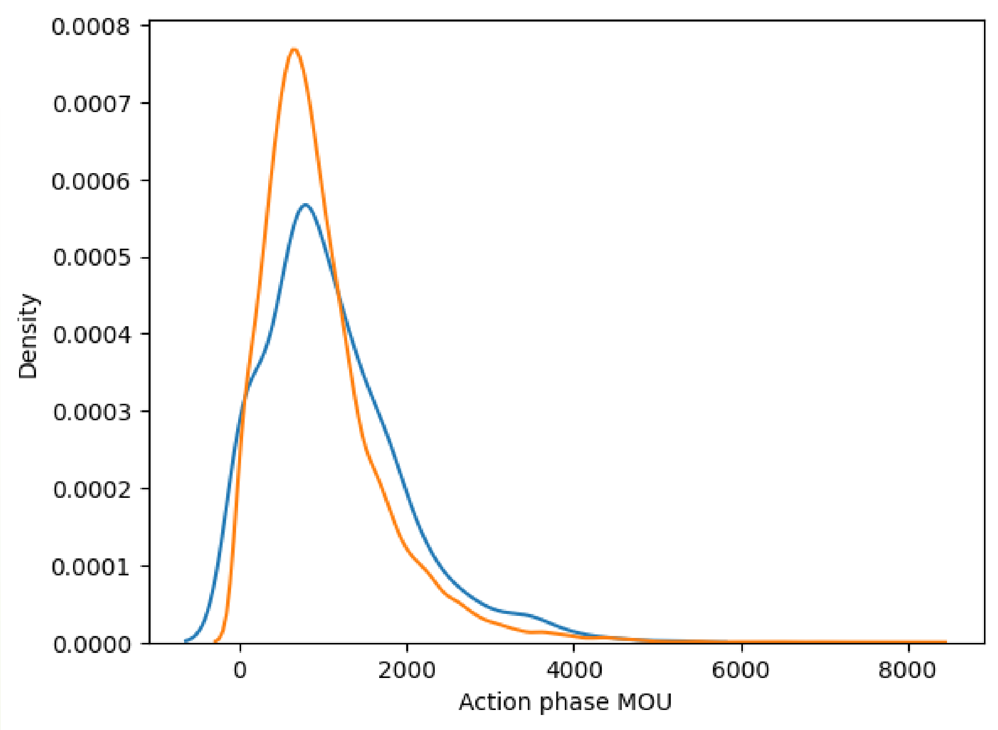

churn
Churn Rate
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
1
decrease_vbc_action


churn
Churn Rate
0.05
0.04
0.03
0.02
0.01
0.00
0
1
decrease_mou_action


churn
Churn Rate
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
1
decrease_rech_num_action

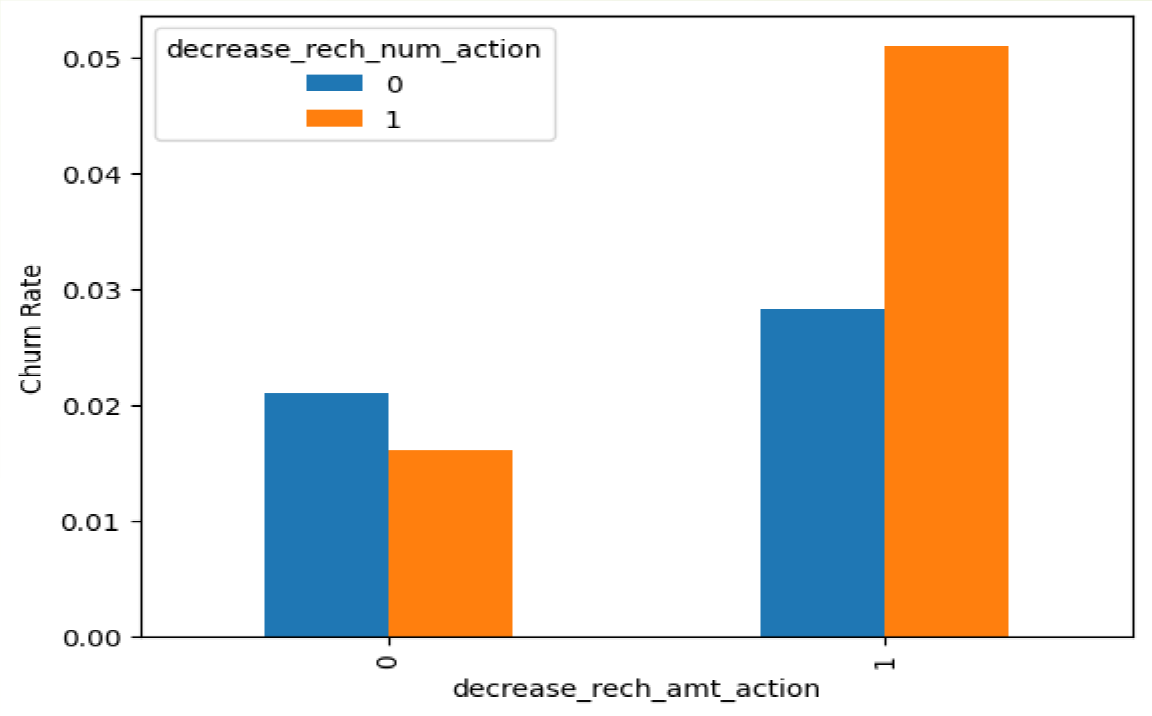

churn
Churn Rate
0.00
0.01
0.02
0.03
0.04
0
1
decrease_rech_amt_action

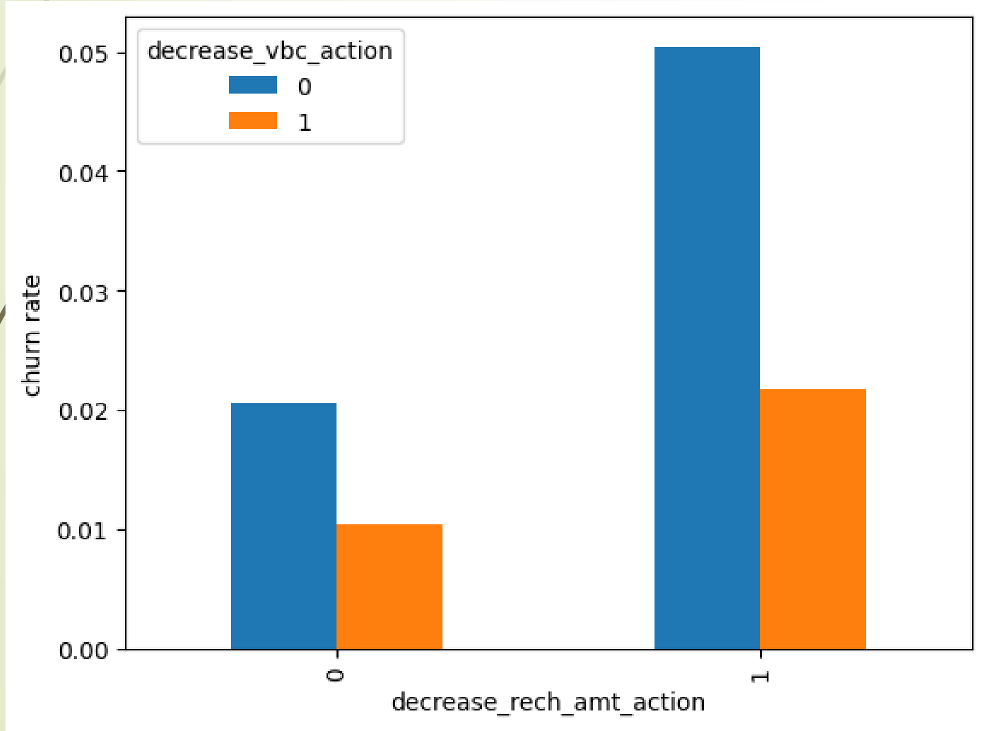

Density
0.00000
0.00025
0.00050
0.00075
0.00100
0.00125
0.00150
0.00175
0.00200
0
1000
2000
3000
4000
Action phase ARPU


Density
0.0000
0.0001
0.0002
0.0003
0.0004
0.0005
0.0006
0.0007
0.0008
0
2000
4000
6000
8000
Action phase MOU

Exploratory Data Analysis (Bivariate)

- 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.
- Churn rate is more for the customers, whose recharge amount is decreased along with the volume based cost is increased in the action month.
- From the scatter plot, the recharge number and the recharge amount are mostly proportional. More the number of recharge, more the amount of the recharge.


decrease_rech_num_action
🟦 0
🟧 1
Churn Rate
0.05
0.04
0.03
0.02
0.01
0.00
0
1
decrease_rech_amt_action


decrease_vbc_action
🟦 0
🟧 1
churn rate
0.05
0.04
0.03
0.02
0.01
0.00
0
1
decrease_rech_amt_action

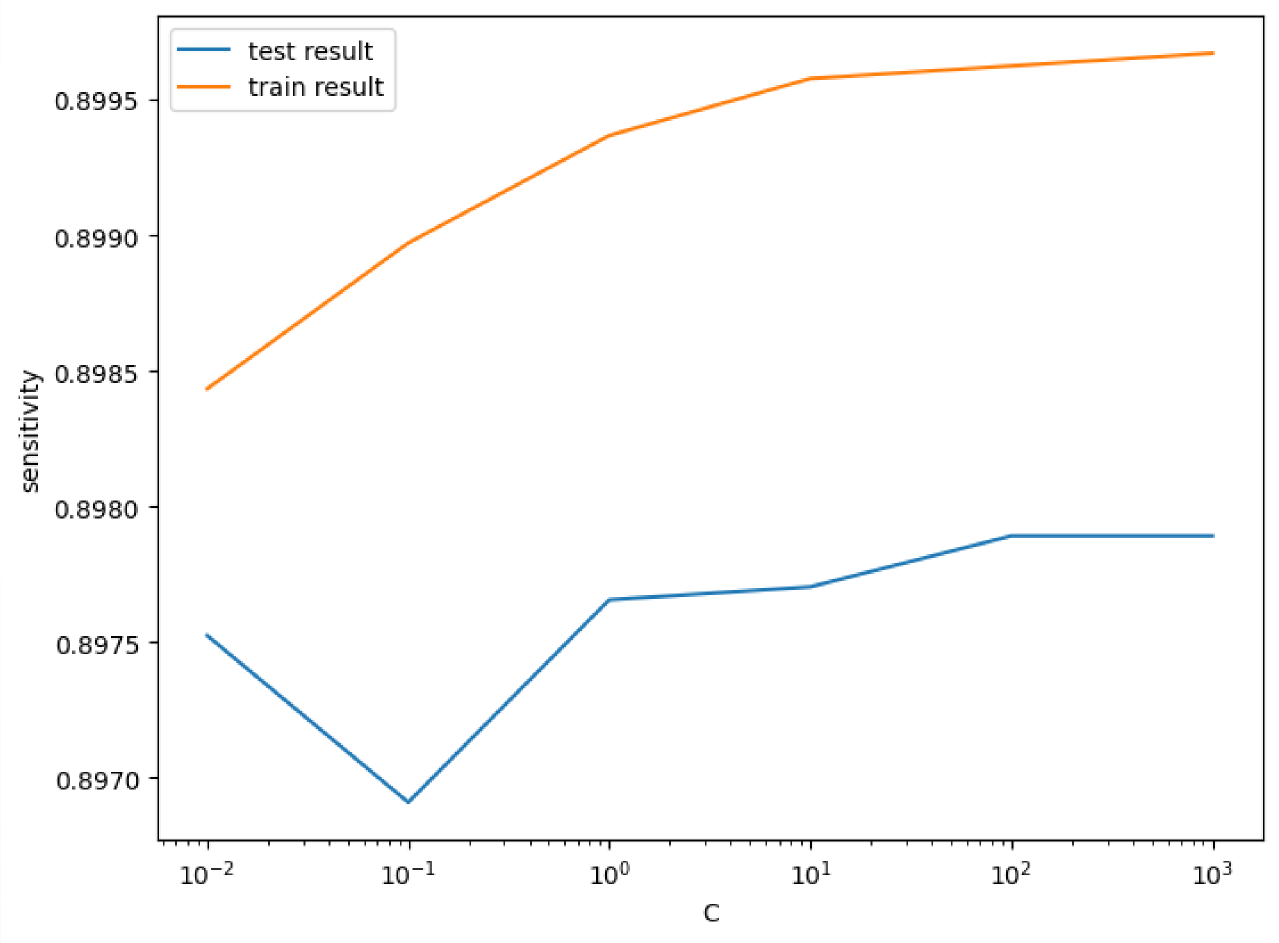
- The dataset is split into 80:20 for train and test.
- Necessary libraries imported for smote.
- To deal with data imbalance we created synthetic samples by doing up sampling using SMOTE(Synthetic Minority Oversampling Technique).

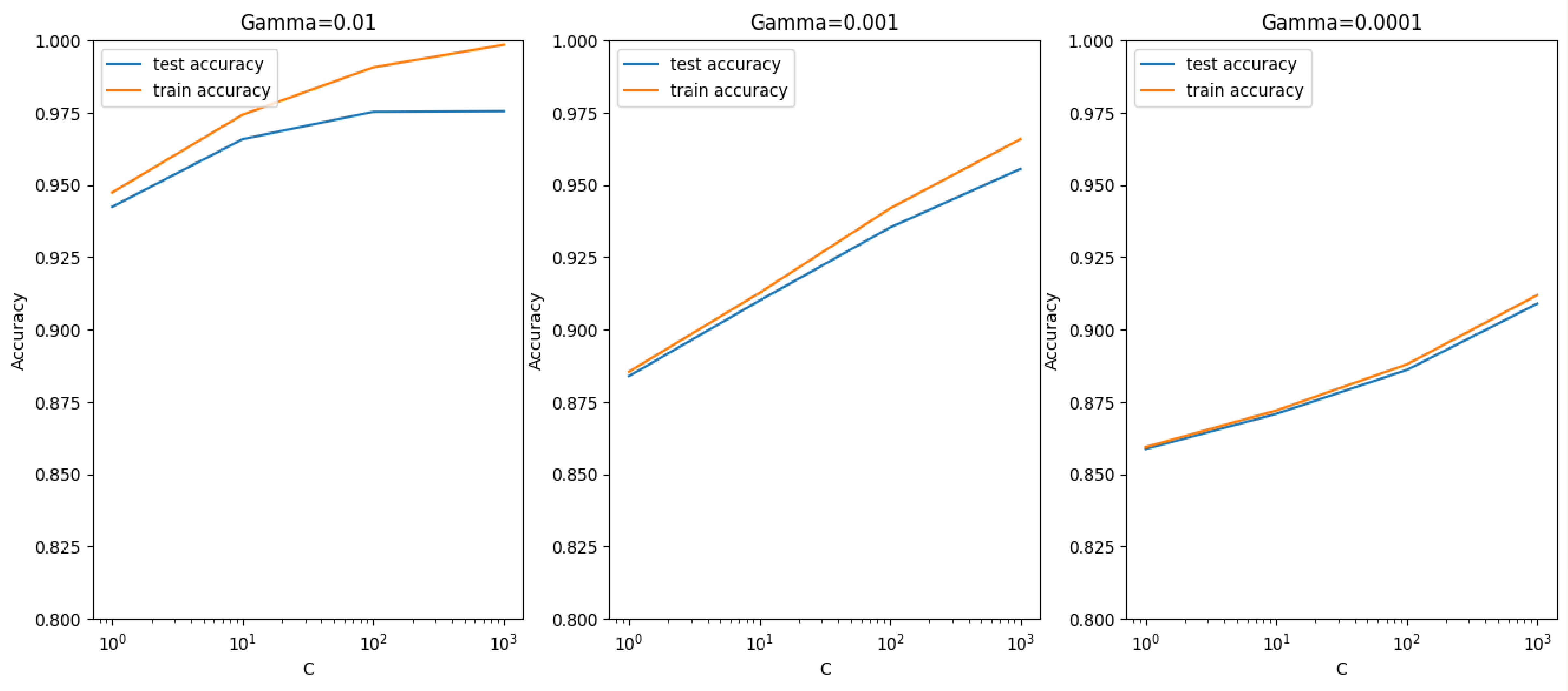
Modeling

After PCA, Logistic regression with PCA and PCA 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




Gamma=0.01
Gamma=0.001
Gamma=0.0001
test accuracy
train accuracy
Accuracy
C
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.800
10^0
10^1
10^2
10^3

Support Vector Machine(SVM) with PCA

C -- Regularization parameter.
gamma -- Handles non linear classifications.

- The best test score is 0.9754959911159373 corresponding to hyper parameters {'C': 1000, 'gamma': 0.01}

High gamma (i.e. high non-linearity) and average value of C
Low gamma (i.e. less non-linearity) and high value of C

Model will be simpler if it has as less non-linearity as possible, so we choose gamma=0.0001 and a high C=100.

Model building with Optimal parameters.

Train set

- Accuracy = 0.89
- Sensitivity = 0.92
- Specificity = 0.85

Test set

- Accuracy = 0.85
- Sensitivity = 0.81
- Specificity = 0.85

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

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

Final conclusion with PCA

From the said models we can see that for achieving the best sensitivity, which was our objective,

- the classic Logistic regression or the SVM models performs well. For both the models the sensitivity was approximately 81%.

- We have good accuracy of approximately 85%.

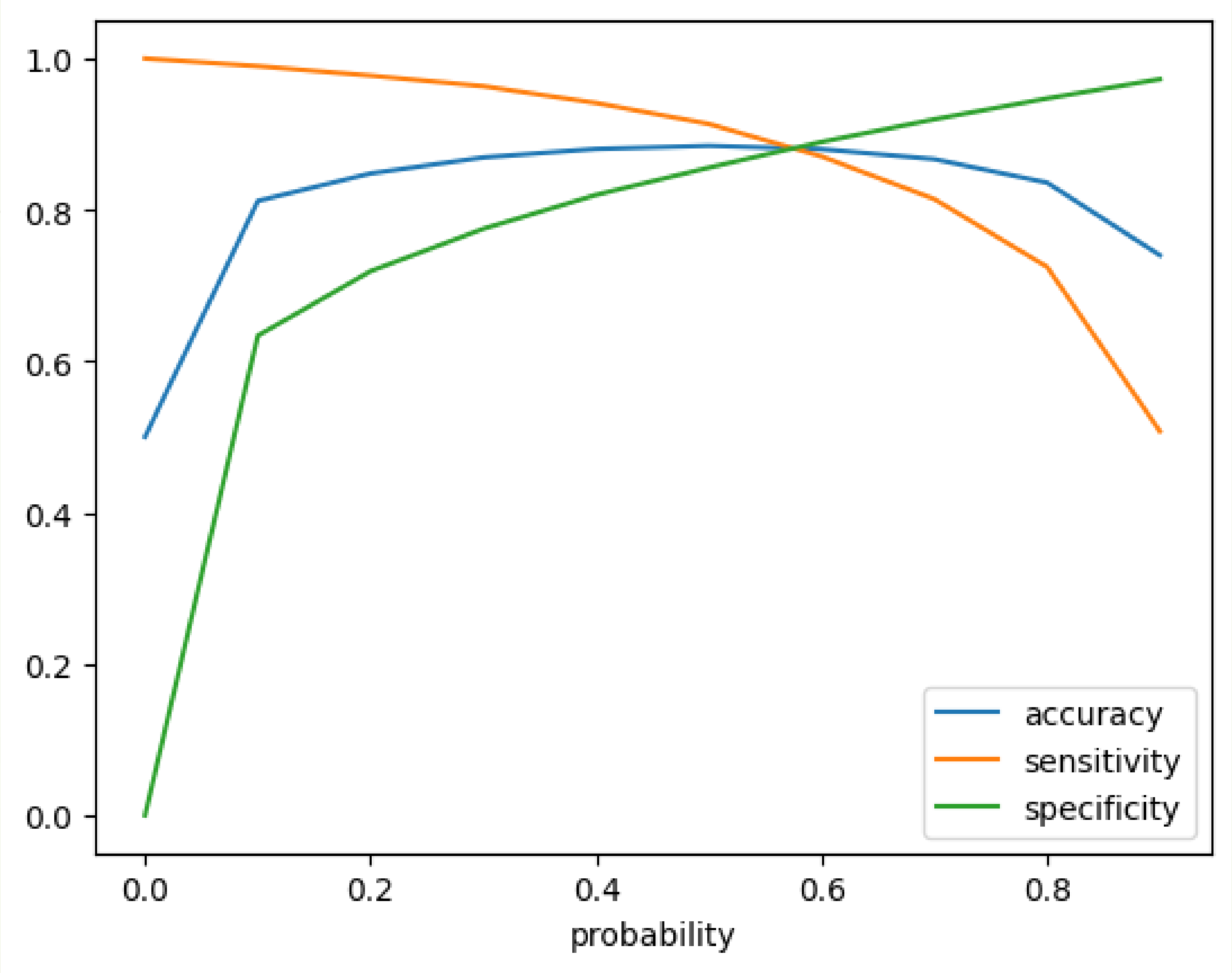
Logistic Regression without PCA

Using RFE

- **With Model 1** : We can remove column **og_others_8**, which is insignificant as it has the highest **p-value 0.99**

- **With Model 2**: The variable **offnet_mou_8** column has the highest **VIF 7.45**. Hence, deleting offnet_mou_8 column.

- **With Model 3**: VIF list we can observe that all the variables are significant and there is no multi collinearity among the variables. This model is finalised.


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
probability
accuracy
sensitivity
specificity

Evaluation

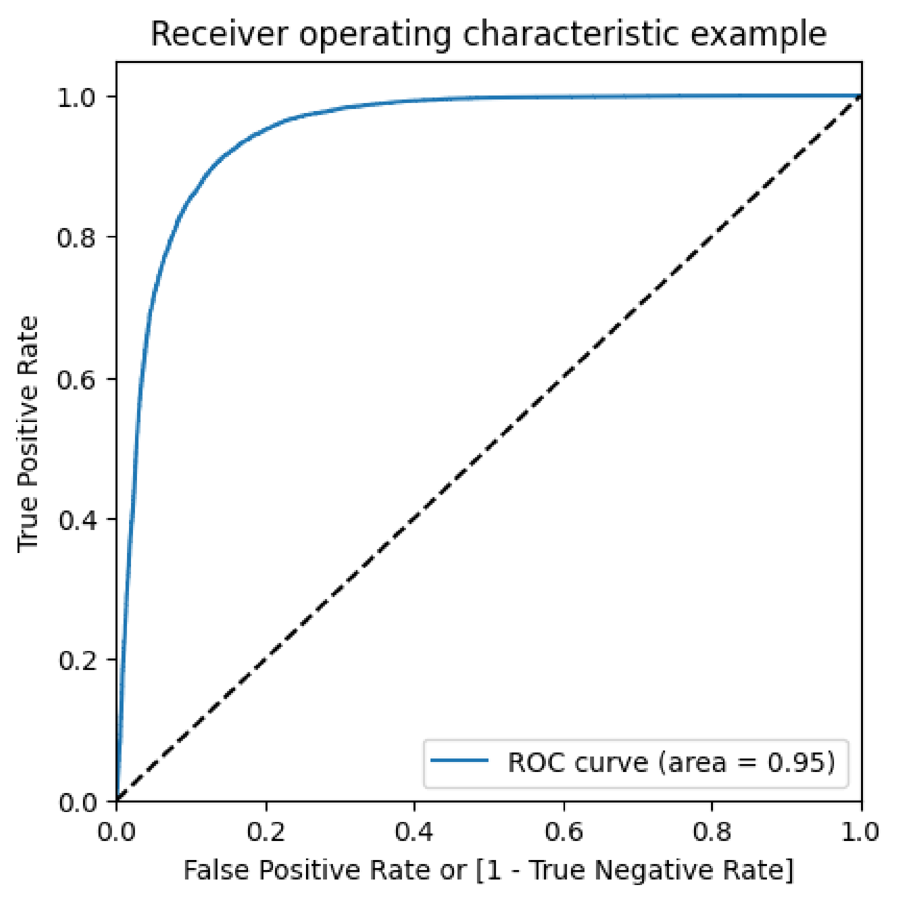
- Accuracy - Becomes stable around 0.6
- Sensitivity - Decreases with the increased probability.
- Specificity - Increases with the increasing probability.
- At probability 0.6, the three parameters(accuracy, sensitivity, specificity) cut each other. There is a good balance.
- We are intended to achieve better sensitivity than accuracy and specificity. We should actually take 0.6 as the optimum probability cut-off, but we took 0.5 for achieving higher sensitivity.

Metrics on train data

From confusion metrics

- Accuracy:- 0.884480746791 1319
- Sensitivity:- 0.913325554259 0432
- Specificity:- 0.855635939323 2206

Plotting ROC – area under the curve is close to 1.



Metrics on test data

From confusion metrics

- Accuracy:- 0.8456957227937195
- Sensitivity:- 0.7616580310880829
- Specificity:- 0.8487284966342558

Model Summary

Train set

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

Test set

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

We observed 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 tells 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 Recommendations

Top predictors

From the results of model-3, we could observe that the most of the top variables have negative coefficients. That means, the variables are inversely correlated with the churn probability.

E.g.:- 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.

➢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.
➢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.