TELECOM CHURN CASE STUDY

CASE STUDY PARTNER:

VIREN GANDHI

MAYANK JAIN

NAMRATA AHIRRAO

BATCH: DS C62

BATCH ID: 5674

STUDENT NAME: VIREN ASHOK GANDHI

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 costs5-10 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, you 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.

BUSINESS OBJECTIVES:

- 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.To do this task well, understanding the typical customer behavior during churn will be helpful.
- Highligting the main variables/factors influencing the Customer churn.
- Use various ML algorithms to build prediction models, evaluate the accuracy and performance of the models.
- Finding out the best model for our business case and providing executive summary.

Process and Methodologies followed to analysis:

- Data cleaning and data manipulation.

  1.Check and handle duplicate data.

  2.Check and handle NA values and missing values.

  3.Drop columns, if it contains large amount of missing values and not useful for the analysis.

  4.Imputation of the values, if necessary.

  5.Check and handle outliers in data.

- EDA

  1.Univariate data analysis: value count, distribution of variable etc.

  2.Bivariate data analysis: correlation coefficients and pattern between the variables etc.

  3. Using different plot like count/bar/pie/hitmap

Methodologies followed to analysis:

- Feature Scaling & Dummy Variables and encoding of the data.
- Splitting the data into Test and Train dataset
- Classification technique: logistic regression used for the model making and prediction.
- Evaluation the model by using different metrics like Specificity and sensitivity or precision and Recall
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

Methodologies followed to analysis:

Data Sourcing, clearing and preparation:

1. Read the data from sources
2. Convert the data in clean format so it is suitable for analysis
3. Remove duplicate data if any
4. Outlier treatment
5. EDA
6. Feature standardization

Feature scaling and splitting of data:

1. Feature scaling of numeric data
2. Splitting data in to train and test at ratio of 80:20

Model Building:

1. Feature selection using REF
2. Determine the optimal model using logistic regression
3. Calculation of various metrics like accuracy, sensitivity, specificity, precision and recall
4. Evaluation of model

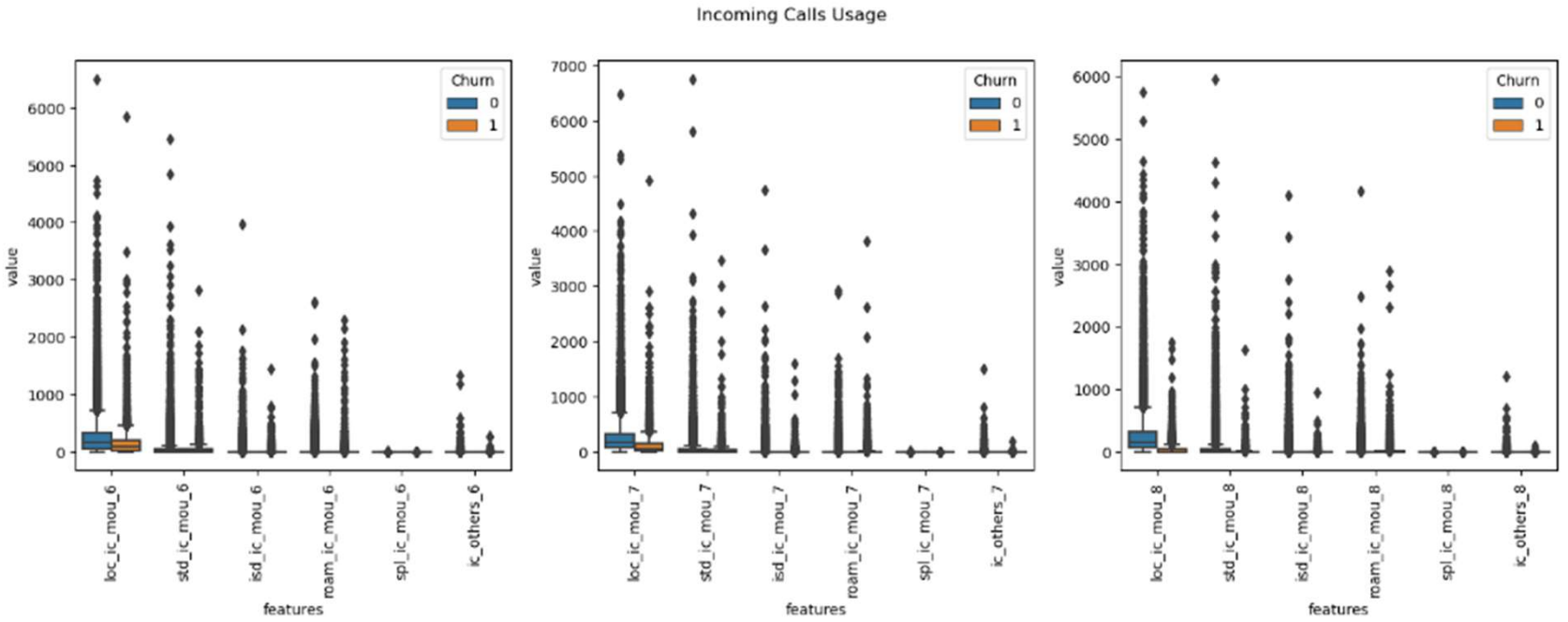
Results

1. Creating column avg_rech_amt_6_7 by summing up total recharge amount of month 6 and 7.Then taking the average of sum.
2. Evaluate final prediction on the test set using threshold limit

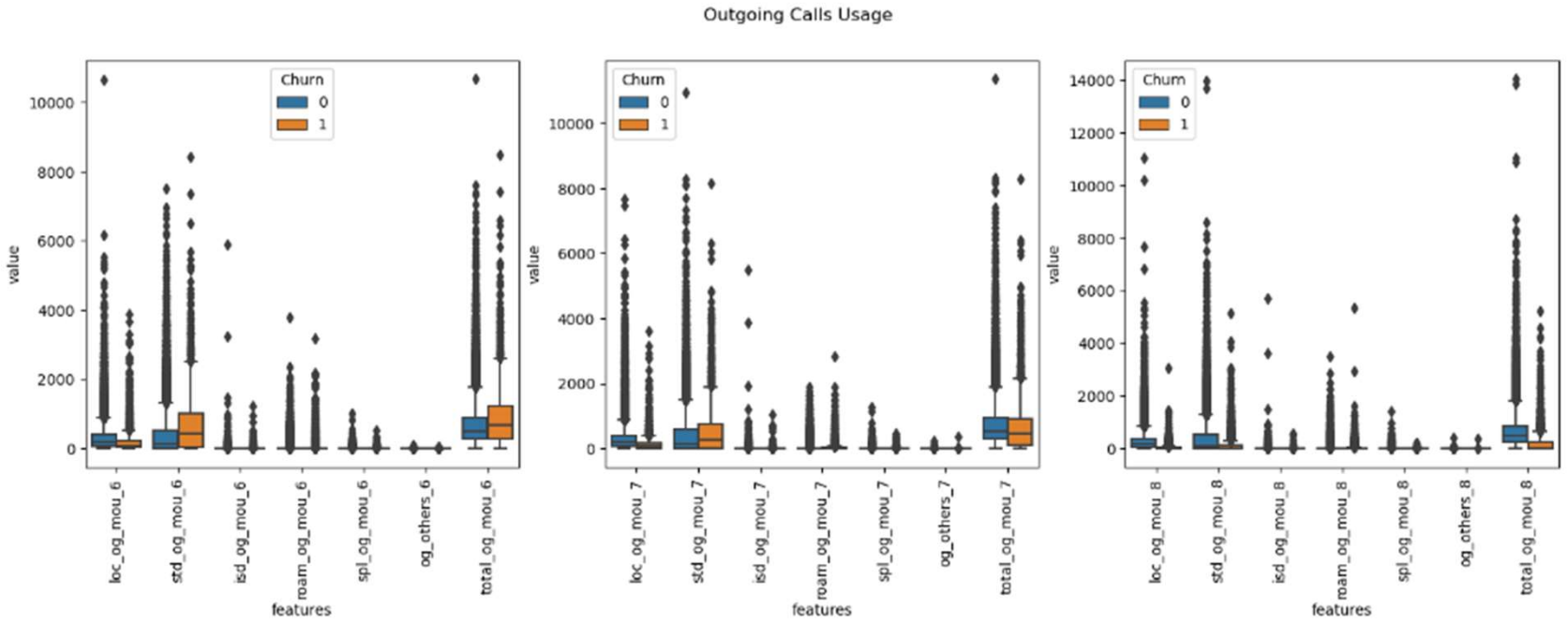
Read and understand the data:

1. There is total 99999 rows and 226 columns.
2. Out of 226 there are 179 float, 35 integer and 12 are objective variables.
3. Out of total business only high value customers gives around 80%.
4. There are outliers present mostly in ‘incoming calls usage ‘ and ‘outgoing calls usage.

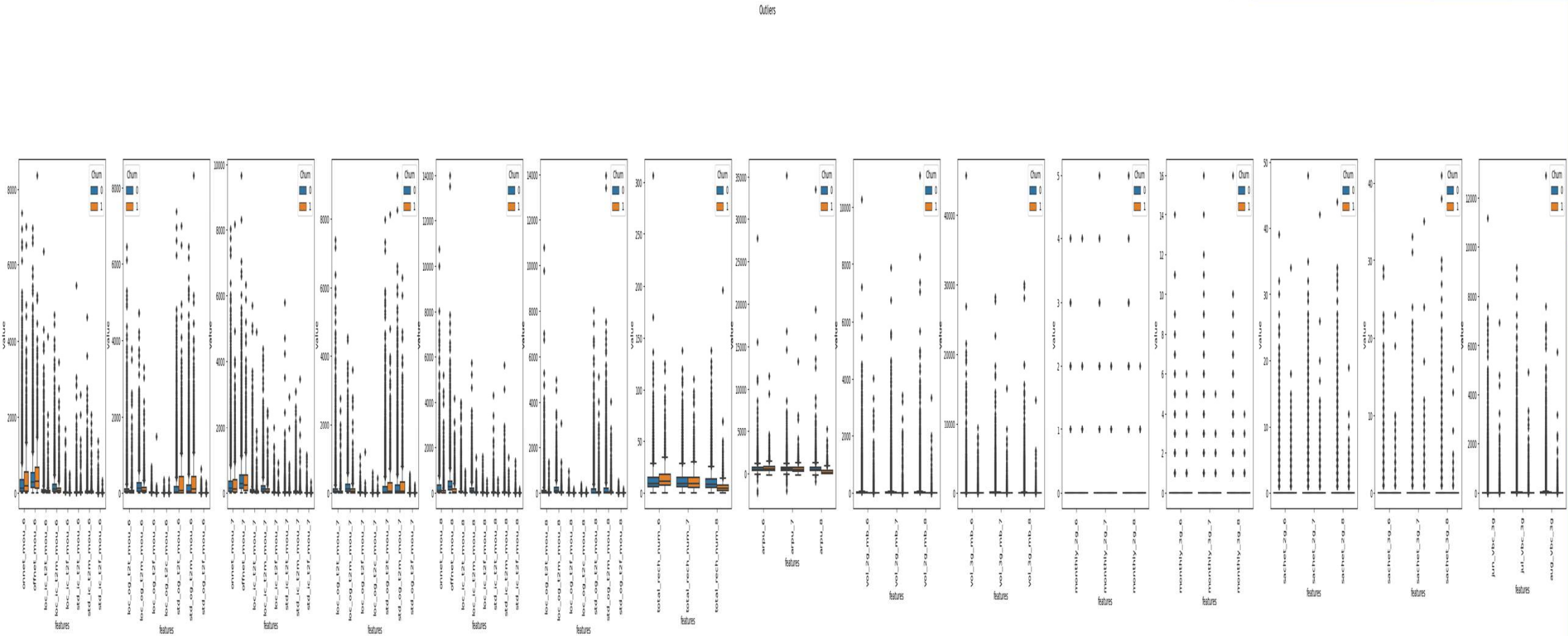
Exploratory Data Analysis-check for outliers-Incoming call



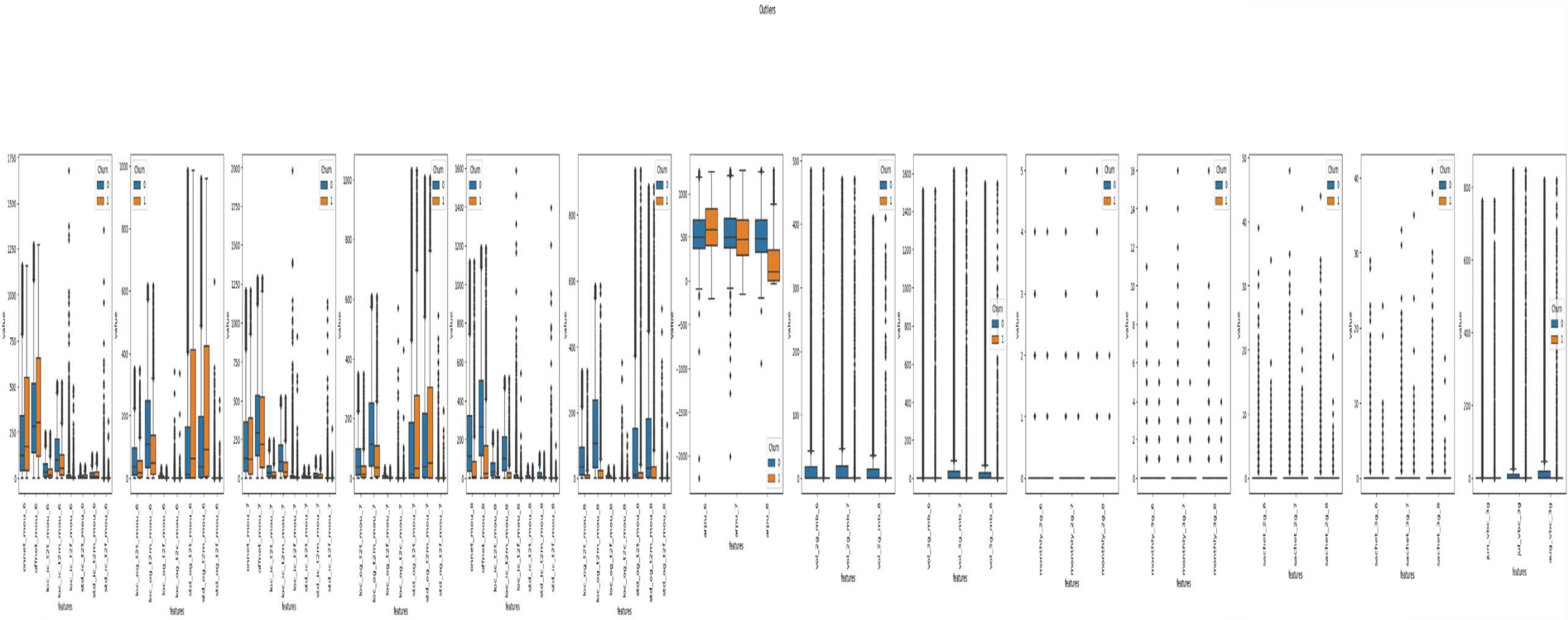
Exploratory Data Analysis-check for outliers-Outgoing call



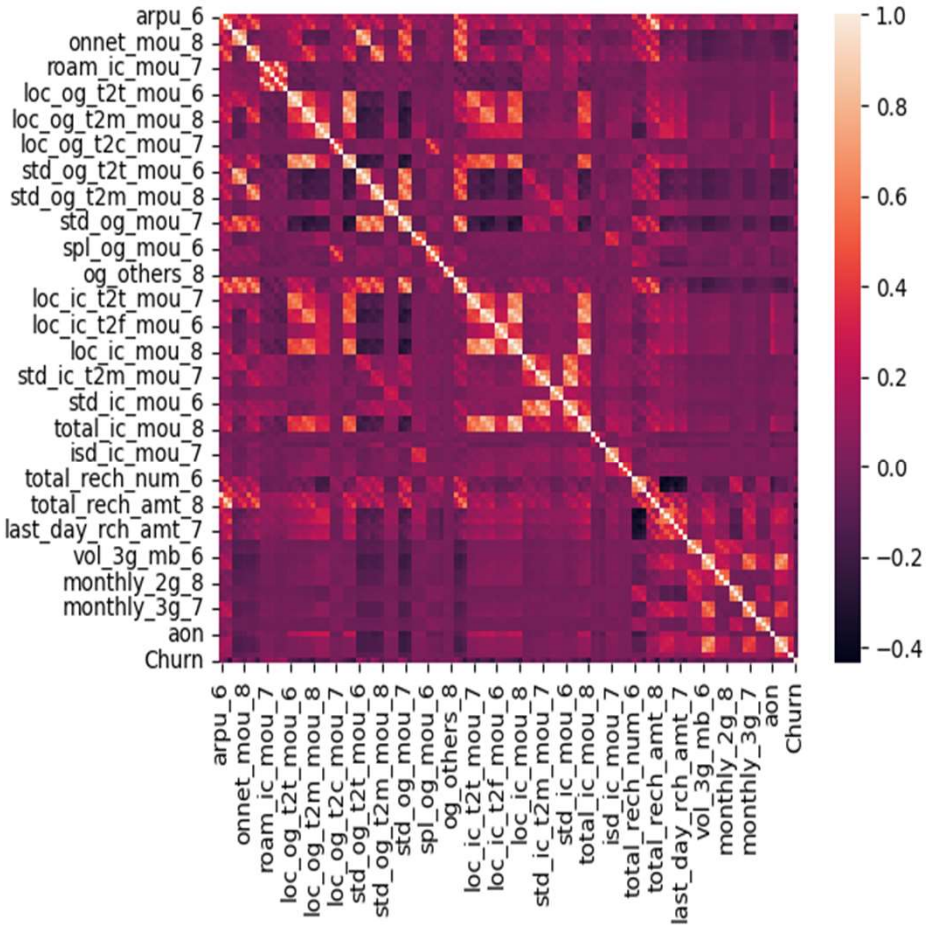
Exploratory Data Analysis-check for outliers

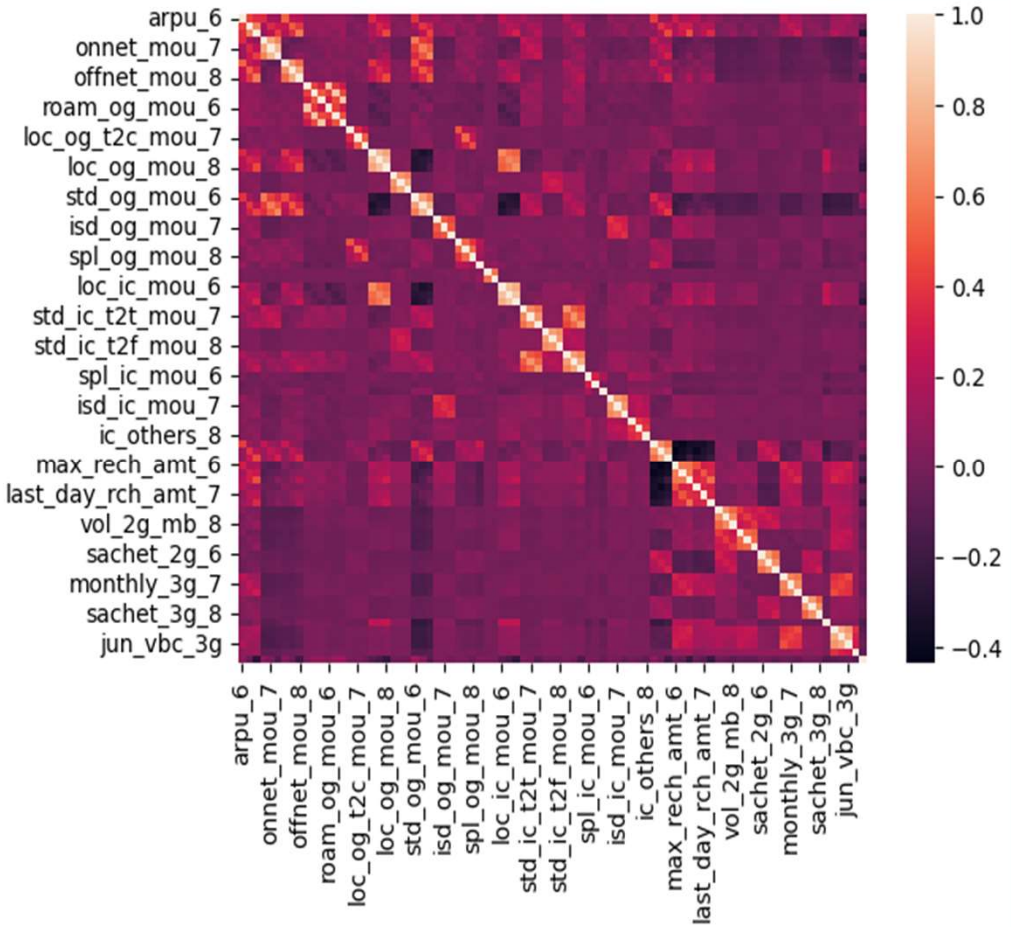


Exploratory Data Analysis-after working on outliers

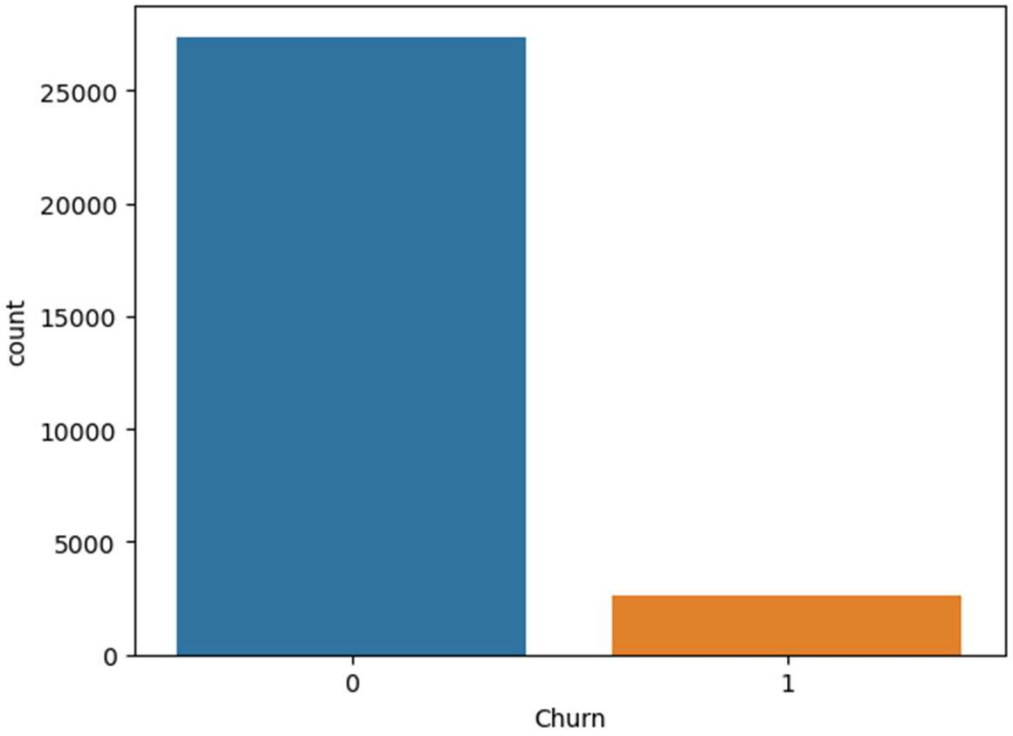


Handling Multicollinearity using Correlation- Before and after treatment





EDA Box plot for Class imbalance.



As we can see there is a clear case of class imbalance

So we need to handle the class imbalance.

TRAIN AND TEST DATA:

- Data set before split

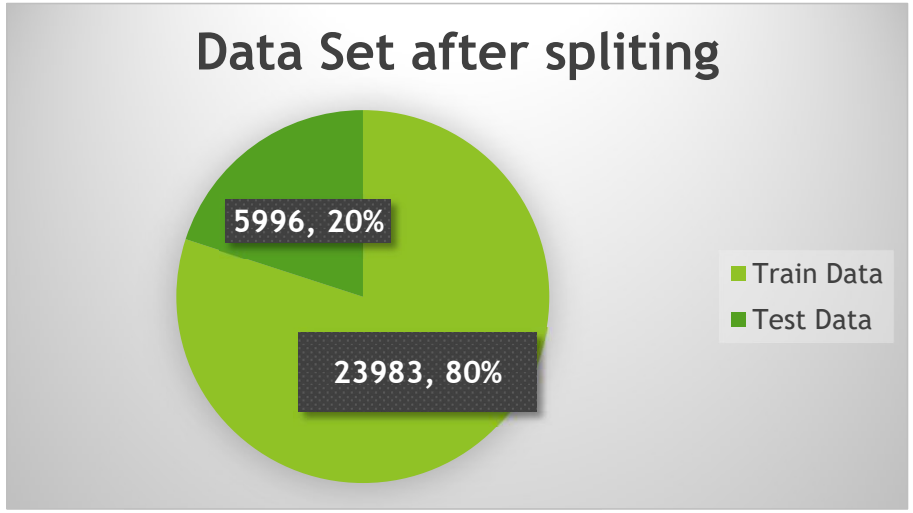
```
df2.shape
```

```
(29979, 87)
```

- Data Set after split in to 80:20

```
df_train.shape, df_test.shape
```

```
((23983, 87), (5996, 87))
```

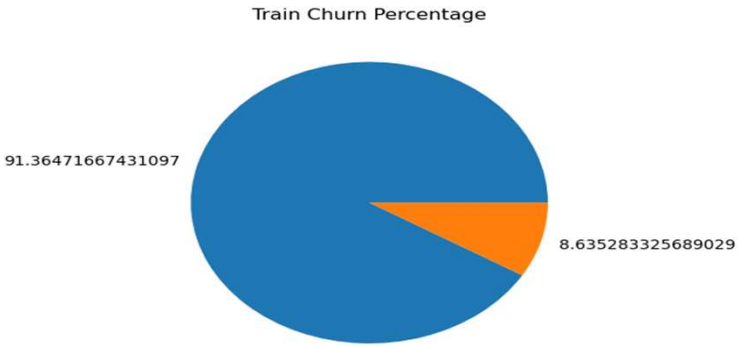


TRAIN AND TEST CHURN VALUE COUNTS:

- Train Churn Value Count

```
df_train.Churn.value_counts(normalize=True)
```

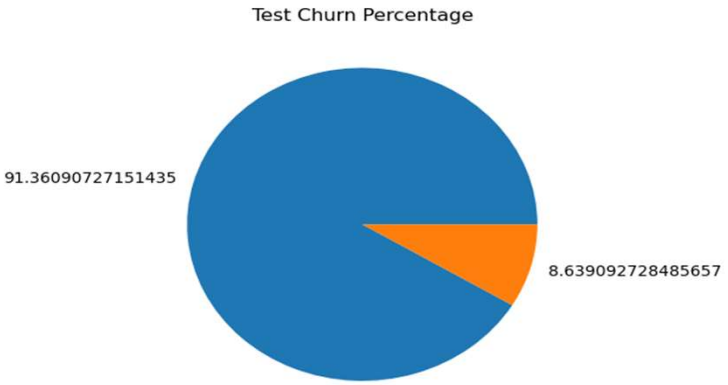
```
Churn
0    0.913647
1    0.086353
Name: proportion, dtype: float64
```



- Test Churn Value Count

```
df_test.Churn.value_counts(normalize=True)
```

```
Churn
0    0.913609
1    0.086391
Name: proportion, dtype: float64
```



Model Building Using Logistic Regression:

Generalized Linear Model Regression Results

Dep. Variable:	Churn	**No. Observations:**	43564
Model:	GLM	**Df Residuals:**	43550
Model Family:	Binomial	**Df Model:**	13
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-18776.
Date:	Mon, 08 Jul 2024	**Deviance:**	37553.
Time:	12:07:17	**Pearson chi2:**	4.10e+07
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.4080
Covariance Type:	nonrobust		

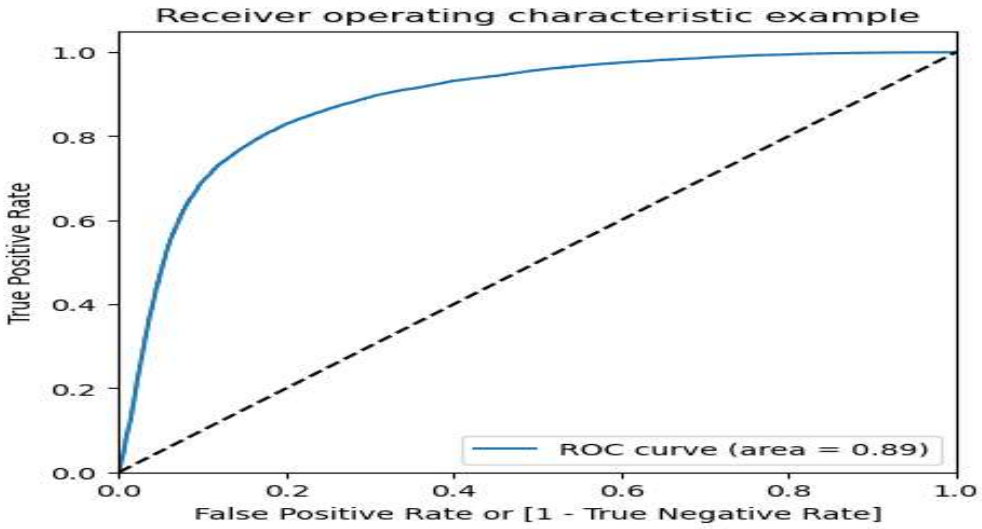
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.4728	0.024	-62.419	0.000	-1.519	-1.427
arpu_7	0.4805	0.016	30.204	0.000	0.449	0.512
roam_og_mou_8	0.3808	0.012	30.824	0.000	0.357	0.405
loc_og_mou_8	-1.0520	0.026	-39.703	0.000	-1.104	-1.000
loc_ic_mou_7	0.0656	0.019	3.496	0.000	0.029	0.102
std_ic_mou_8	-0.3979	0.017	-23.262	0.000	-0.431	-0.364
spl_ic_mou_8	-0.2773	0.021	-13.382	0.000	-0.318	-0.237
total_rech_num_8	-0.8243	0.017	-47.546	0.000	-0.858	-0.790
last_day_rch_amt_8	-0.7676	0.017	-45.554	0.000	-0.801	-0.735
monthly_2g_8	-0.5209	0.025	-20.925	0.000	-0.570	-0.472
sachet_2g_8	-0.3432	0.026	-13.352	0.000	-0.394	-0.293
monthly_3g_8	-0.4323	0.025	-17.078	0.000	-0.482	-0.383
aon	-0.4151	0.017	-25.149	0.000	-0.447	-0.383
sep_vbc_3g	-0.9159	0.053	-17.192	0.000	-1.020	-0.812

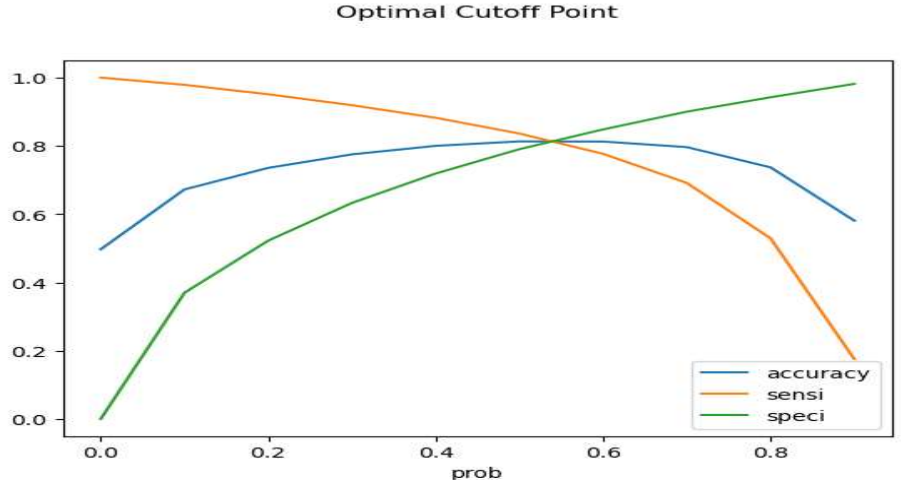
	Features	VIF
2	loc_og_mou_8	2.01
3	loc_ic_mou_7	1.64
7	last_day_rch_amt_8	1.45
6	total_rech_num_8	1.42
0	arpu_7	1.28
11	aon	1.21
4	std_ic_mou_8	1.16
9	sachet_2g_8	1.16
10	monthly_3g_8	1.14
1	roam_og_mou_8	1.09
12	sep_vbc_3g	1.09
8	monthly_2g_8	1.07
5	spl_ic_mou_8	1.03

- Fourth training model gives us good p value and scores.

```
Accuracy:  0.8131255164814984
F1 score:  0.8164415683975559
Recall:  0.8361814151117679
Precision:  0.7976122296136393
ROC_AUC_SCORE:  0.8132623030286843

 clasification report:
                precision    recall  f1-score   support

           0       0.83      0.79      0.81     21912
           1       0.80      0.84      0.82     21652

    accuracy                           0.81     43564
   macro avg       0.81      0.81      0.81     43564
weighted avg       0.81      0.81      0.81     43564


 confussion matrix:
 [[17318  4594]
 [ 3547 18105]]
```

MODEL EVALUATION-ROC,SENSITIVITY AND SPECIFICITY



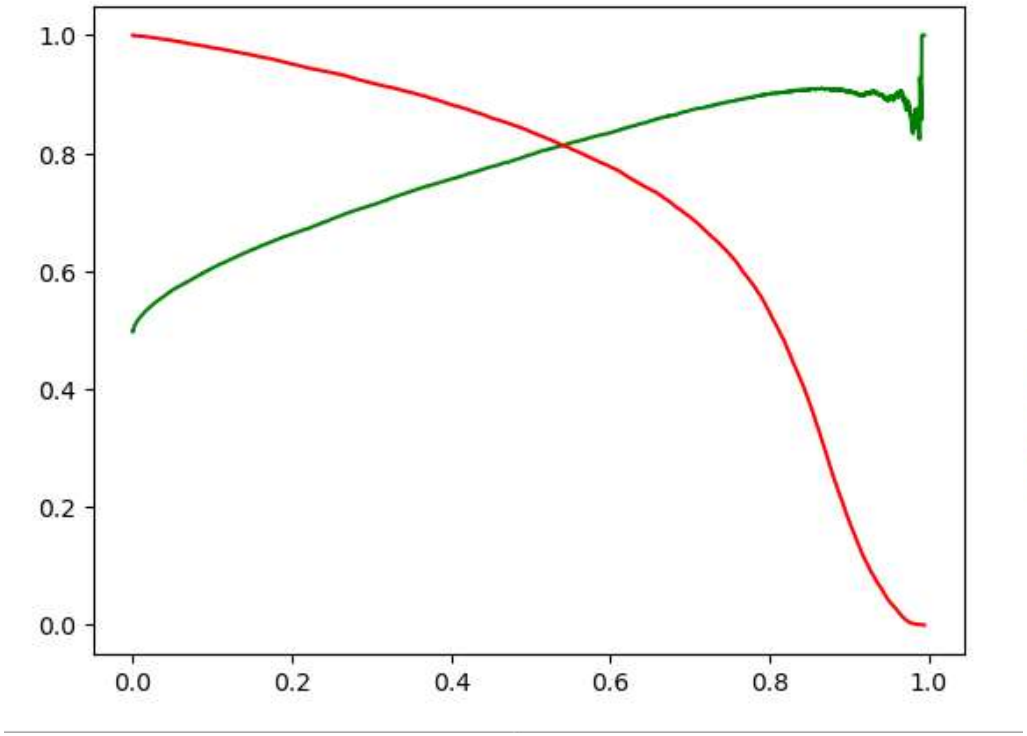


Confusion Matrix

17318	4594
3547	18105

Accuracy: 0.8131255164814984
F1 score: 0.8164415683975559
Recall: 0.8361814151117679
Precision: 0.7976122296136393
ROC_AUC_SCORE: 0.8132623030286843

PRECISION AND RECALL TRADE OFF



Confusion Matrix

4537	941
93	425

```
Accuracy:  0.8275517011340894
F1 score:  0.45116772823779194
Recall:  0.8204633204633205
Precision:  0.31112737920937045
ROC_AUC_SCORE:  0.8243426496438545

 clasification report:
              precision    recall  f1-score   support

           0       0.98      0.83      0.90      5478
           1       0.31      0.82      0.45       518

    accuracy                           0.83      5996
   macro avg       0.65      0.82      0.67      5996
weighted avg       0.92      0.83      0.86      5996
```

DECISSION TREE

- On Train Data

Accuracy: 0.8205470313542361
F1 score: 0.4532520325203252
Recall: 0.861003861003861
Precision: 0.30758620689655175
ROC_AUC_SCORE: 0.8388626460915617

clasification report:

	precision	recall	f1-score	support
0	0.98	0.82	0.89	5478
1	0.31	0.86	0.45	518
accuracy			0.82	5996
macro avg	0.65	0.84	0.67	5996
weighted avg	0.93	0.82	0.85	5996

confussion matrix:
[[4474 1004]
[72 446]]

- On Test Data

Accuracy: 0.8205470313542361
F1 score: 0.4532520325203252
Recall: 0.861003861003861
Precision: 0.30758620689655175
ROC_AUC_SCORE: 0.8388626460915617

clasification report:

	precision	recall	f1-score	support
0	0.98	0.82	0.89	5478
1	0.31	0.86	0.45	518
accuracy			0.82	5996
macro avg	0.65	0.84	0.67	5996
weighted avg	0.93	0.82	0.85	5996

confussion matrix:
[[4474 1004]
[72 446]]

most important features for our final model are all from action phase.

1. roam_og_mou_8
2. loc_ic_mou_8
3. roam_ic_mou_8
4. arpu_8
5. loc_og_mou_8
6. last_day_rch_amt_8
7. max_rech_amt_8
8. total_rech_num_8
9. std_ic_mou_8
10. onnet_mou_8

CONCULSION:

- The company after identifying customers in action phase can give offers for increasing local incoming and outgoing minutes of usage.
- Customers are more keen towards the local incoming calls over anything so, we can provide more free incoming calls and also we can reduce the outgoing calls charges for better connectivity.
- This can provide an advantage over other operators in the market
- The roaming charges can be made lesser by giving offers.
- More importantly we can provide free incoming calls on roaming.
- We can provide attractive offers and packages for the customers.

THANK YOU