TELECOM CHURN CASE STUDY –Assignment

Submission by : T. Naresh Kumar, NEIL RAGHAV, Neelam KUMARI– DS-C54 March 2023

Business Objective

Maximize: Company's profit by retaining customer.
Minimize: Customer churn by identifying the key cause of the problem.
Business Constraint:

Provide offers and discount and improve the service quality without compromising with profit.

This is a classification project since the variable to be predicted is binary (churn or not churn). The objective here is to predict churn probability, conditioned on the customer features.

Data Inputs

1. Data + Dictionary+ Telecom+ Churn+ Case+ Study
2. Telecom_churn_data.csv.

Data Cleaning & Manipulation of Data

1. Null values
2. Filtering Unwanted columns
3. Missing values
4. Sorting the data
5. Fixing the data type
6. Feature scaling
7. Model building

Removed the 70% of Null values

Tagging the CHURNERS

Now tagged the churned customers (churn=1, else 0) based on the fourth month as follows: Those who have not made any calls (either incoming or outgoing) AND have not used mobile internet even once in the churn phase.
We'll use total_ic_mou_9, total_og_mou_9, vol_2g_mb_9, vol_3g_mb_9 columns to tag the curners. For churners there will not be any voice and data usage

```
DF.head()
```

	arpu_6	arpu_7	arpu_8	arpu_9	onnet_mou_6	onnet_mou_7	onnet_mou_8	onnet_mou_9	offnet_mou_6	offnet_mou_7	offnet_mou_8	offnet_mou_9	rc
7	1,069.18	1,349.85	3,171.48	500.00	57.84	54.68	52.29	0.00	453.43	567.16	325.91	0.00	
8	378.72	492.22	137.36	166.79	413.69	351.03	35.08	33.46	94.66	80.63	136.48	108.71	
13	492.85	205.67	593.26	322.73	501.76	108.39	534.24	244.81	413.31	119.28	482.46	214.06	
16	430.98	299.87	187.89	206.49	50.51	74.01	70.61	31.34	296.29	229.74	162.76	224.39	
17	690.01	18.98	25.50	257.58	1,185.91	9.28	7.79	558.51	61.64	0.00	5.54	87.89	

```
DF['Churn'].value_counts()
```

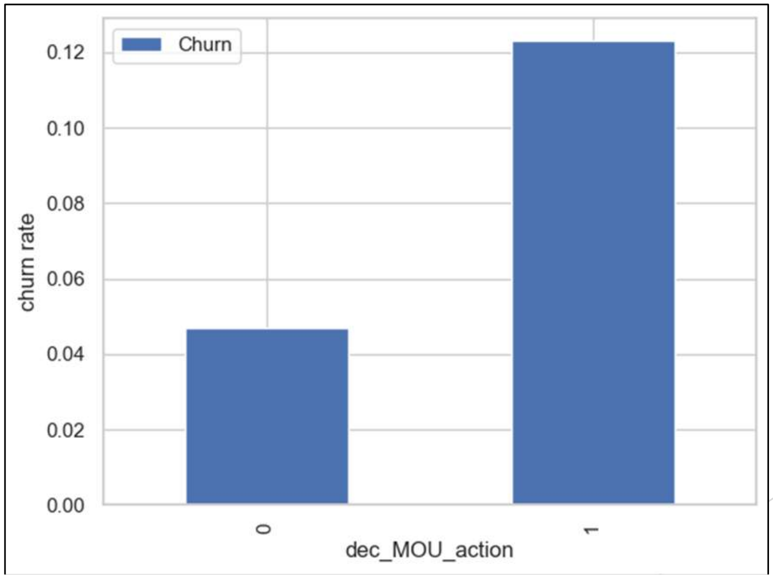
```
0    27418
1     2593
Name: Churn, dtype: int64
```

Outlier Treatment

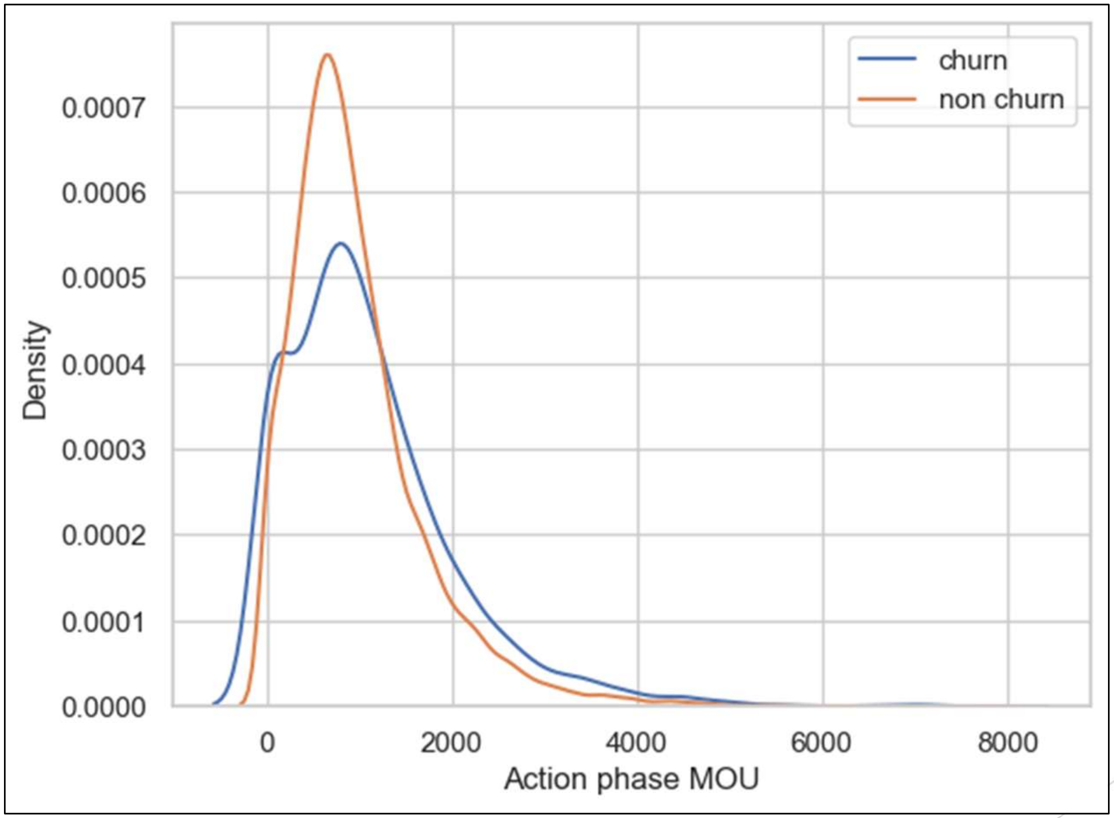
```
Index(['arpu_6', 'arpu_7', 'arpu_8', 'onnet_mou_6', 'onnet_mou_7',
       'onnet_mou_8', 'offnet_mou_6', 'offnet_mou_7', 'offnet_mou_8',
       'roam_ic_mou_6',
       ...
       'monthly_3g_7', 'monthly_3g_8', 'sachet_3g_6', 'sachet_3g_7',
       'sachet_3g_8', 'aon', 'aug_vbc_3g', 'jul_vbc_3g', 'jun_vbc_3g',
       'avg_rech_6_7'],
      dtype='object', length=134)
```

Univariate Analysis

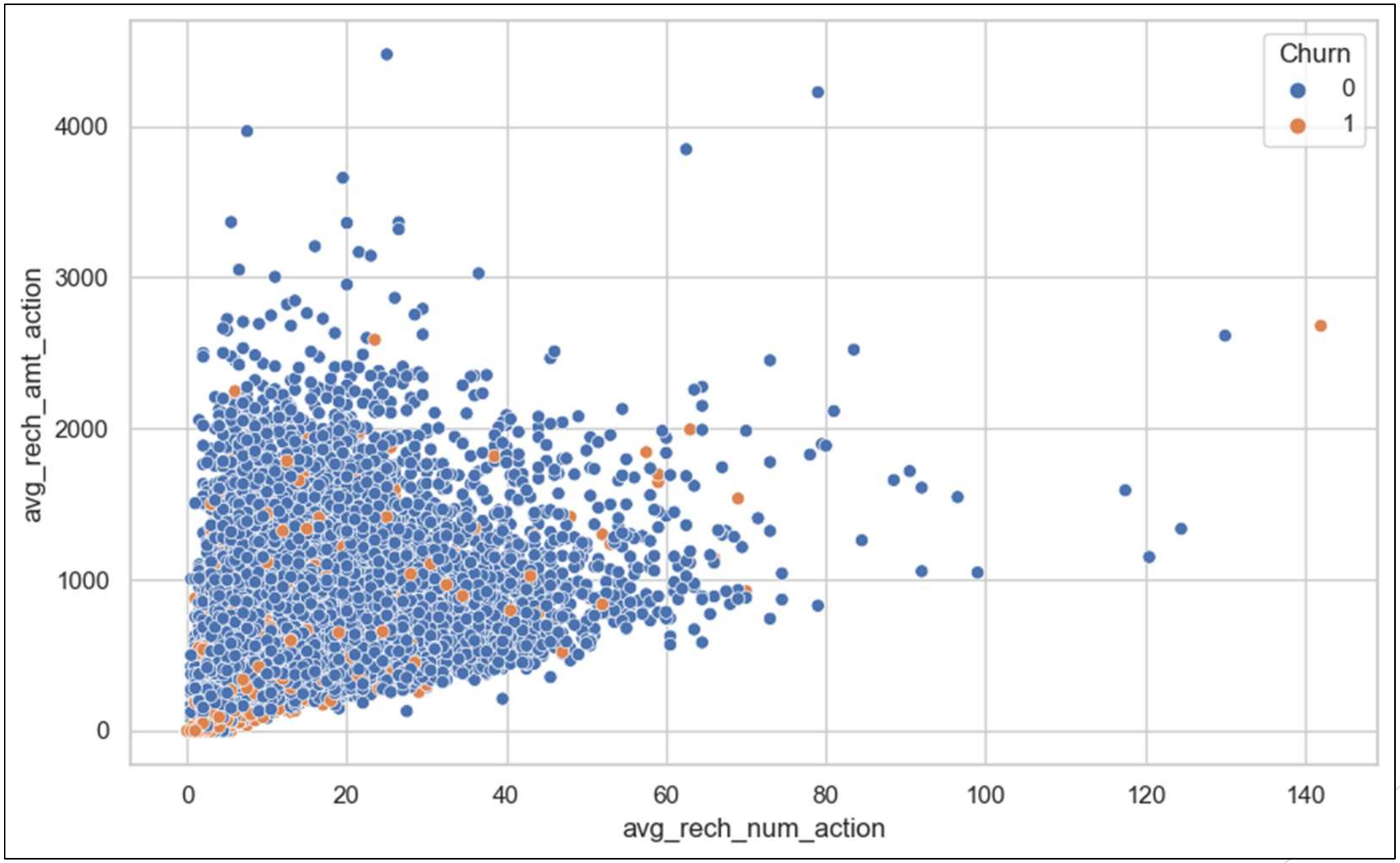
Churn rate on the basis whether the customer decreased her/his MOU in action month



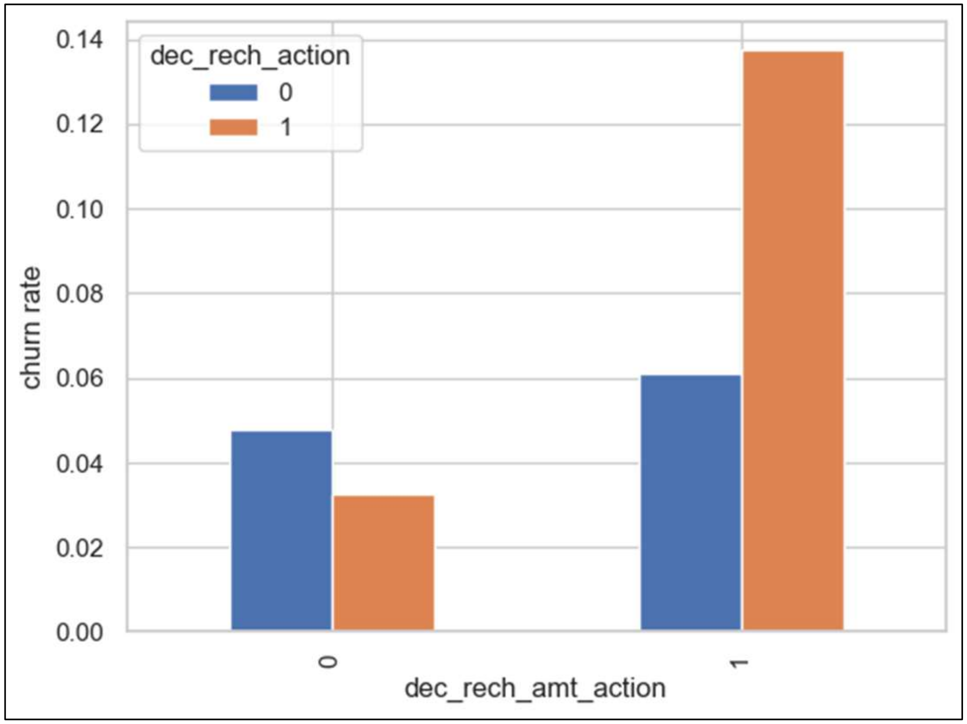
Analysis of the minutes of usage MOU (churn and not churn) in the action phase



Bi-Variate Analysis

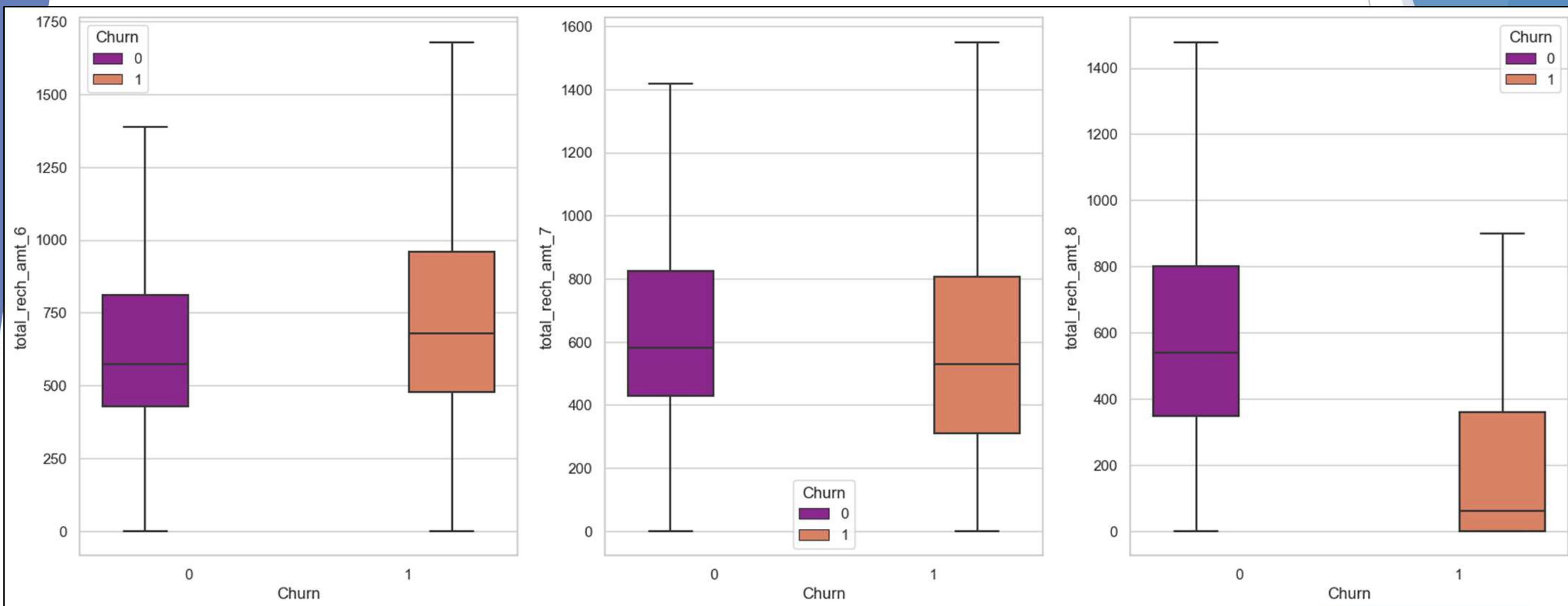


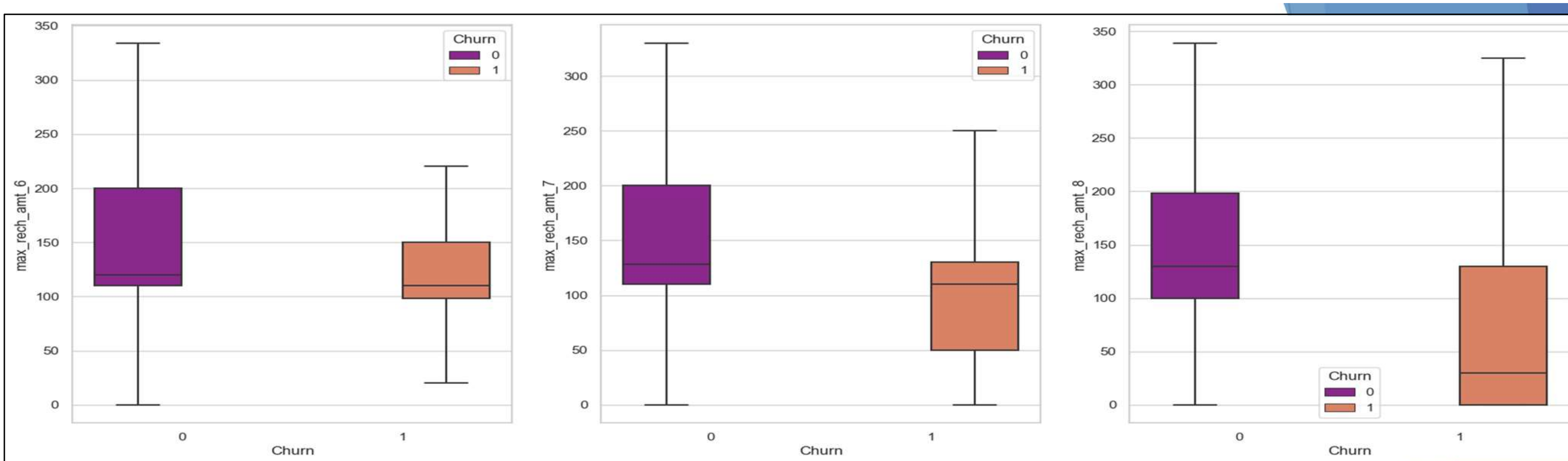
Analyzing churn rate WRT the decreasing recharge amount and number of recharge during the action phase



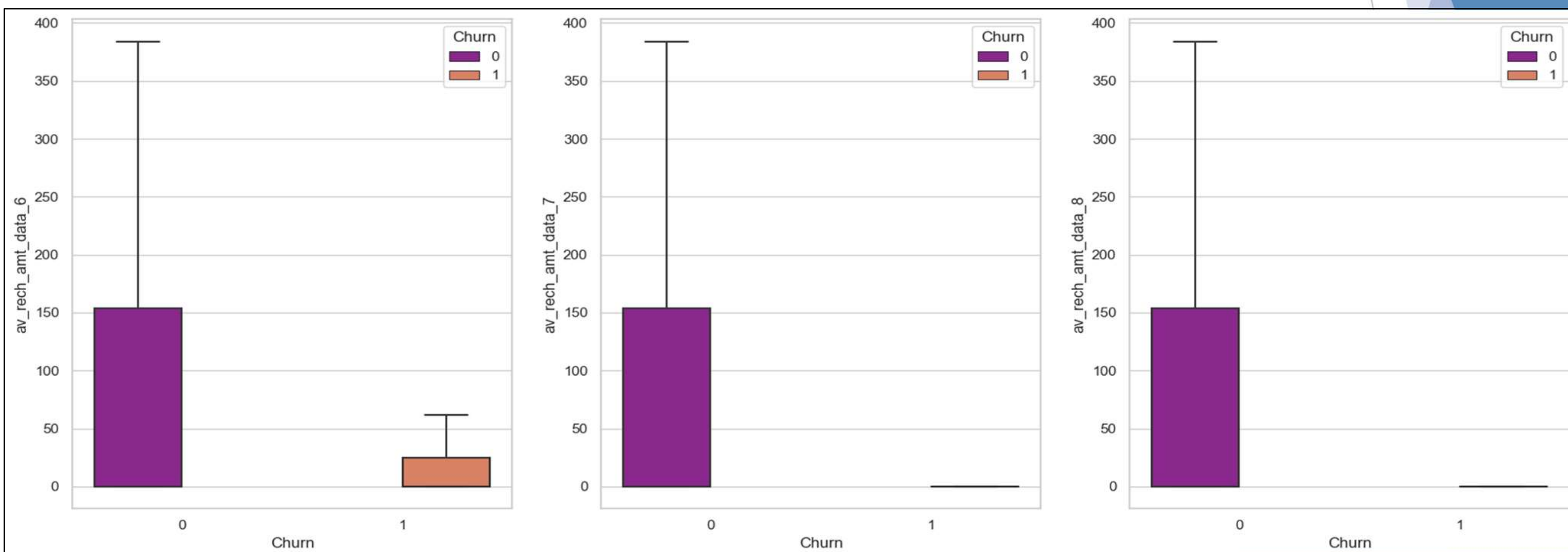
We can see from the above plot, that the churn rate is higher for the customers, whose recharge amount as well as number of recharge have decreased in the action phase when compared to the good phase.

From the Below plots we can see clearly that the recharge amounts (Total & Maximum) started to fall in the month 8 i.e near to the churn phase.

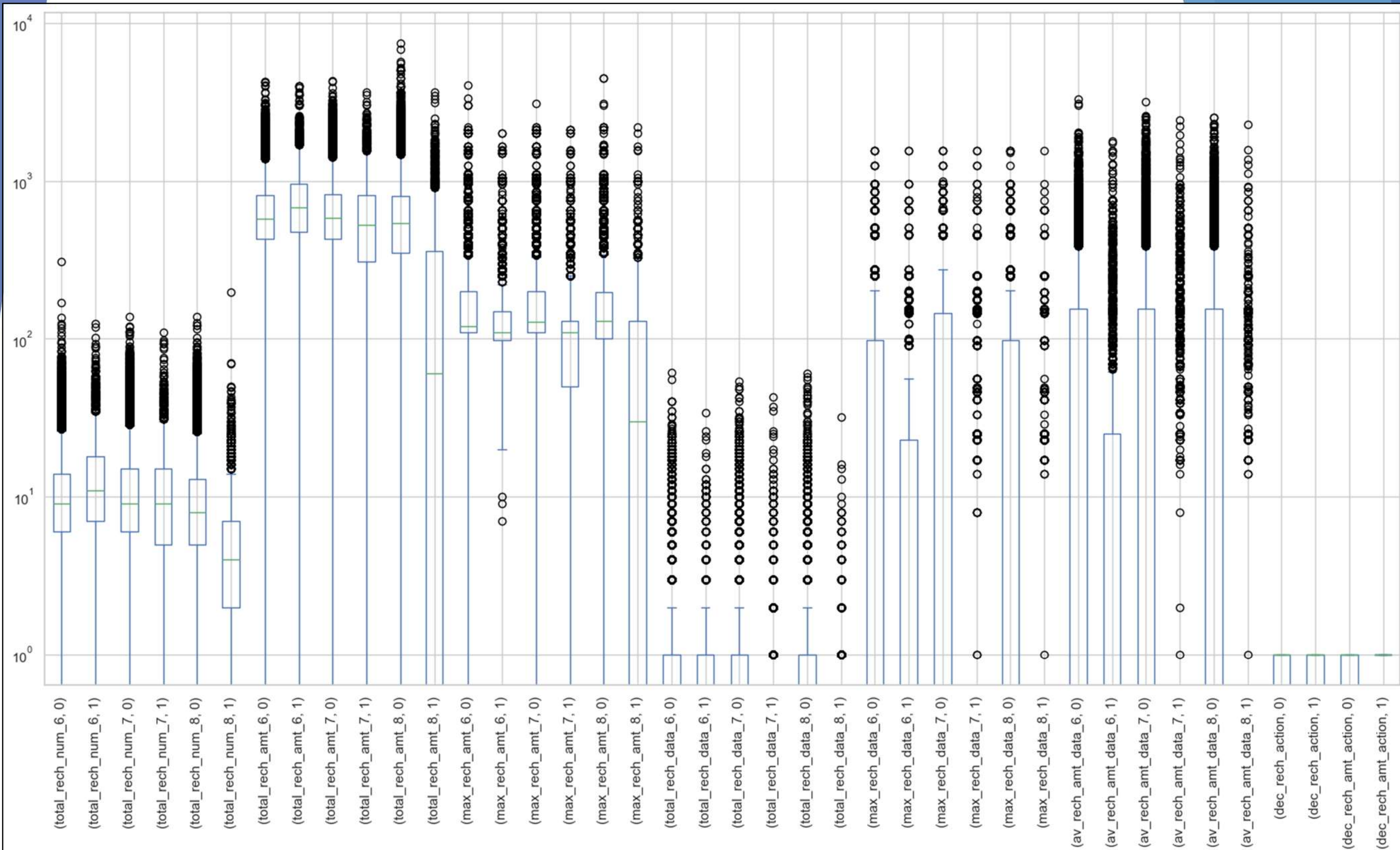




max_rech_amt_6
max_rech_amt_7
max_rech_amt_8
Churn
0
1



av_rech_amt_data_6
av_rech_amt_data_7
av_rech_amt_data_8
Churn
0
1



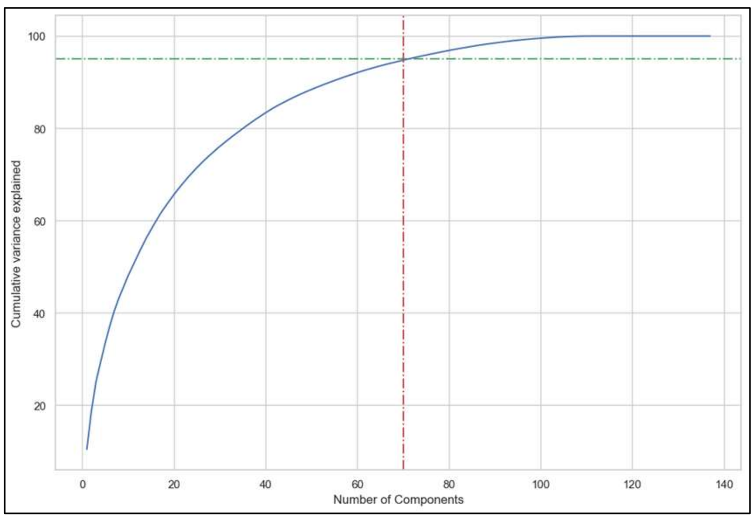
10^4
10^3
10^2
10^1
10^0
(total_rech_num_6, 0)
(total_rech_num_6, 1)
(total_rech_num_7, 0)
(total_rech_num_7, 1)
(total_rech_num_8, 0)
(total_rech_num_8, 1)
(total_rech_amt_6, 0)
(total_rech_amt_6, 1)
(total_rech_amt_7, 0)
(total_rech_amt_7, 1)
(total_rech_amt_8, 0)
(total_rech_amt_8, 1)
(max_rech_amt_6, 0)
(max_rech_amt_6, 1)
(max_rech_amt_7, 0)
(max_rech_amt_7, 1)
(max_rech_amt_8, 0)
(max_rech_amt_8, 1)
(total_rech_data_6, 0)
(total_rech_data_6, 1)
(total_rech_data_7, 0)
(total_rech_data_7, 1)
(total_rech_data_8, 0)
(total_rech_data_8, 1)
(max_rech_data_6, 0)
(max_rech_data_6, 1)
(max_rech_data_7, 0)
(max_rech_data_7, 1)
(max_rech_data_8, 0)
(max_rech_data_8, 1)
(av_rech_amt_data_6, 0)
(av_rech_amt_data_6, 1)
(av_rech_amt_data_7, 0)
(av_rech_amt_data_7, 1)
(av_rech_amt_data_8, 0)
(av_rech_amt_data_8, 1)
(dec_rech_action, 0)
(dec_rech_action, 1)
(dec_rech_amt_action, 0)
(dec_rech_amt_action, 1)

Scaling numeric features

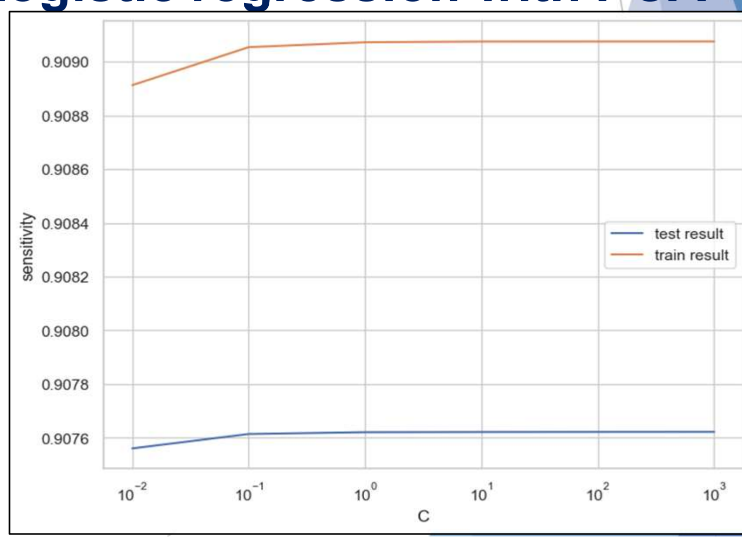
During EDA we have observed few outliers in numeric features. So, using Robust Scaling using median and quantile values instead of Standard Scaling using mean and standard deviation.

	arpu_6	arpu_7	arpu_8	onnet_mou_6	onnet_mou_7	onnet_mou_8	offnet_mou_6	offnet_mou_7	offnet_mou_8	roam_ic_mou_6	roam_ic_mou_7	roam_
0	1,409.37	1,052.63	1,674.24	453.28	343.38	589.58	826.99	811.99	815.96	70.83	39.78	
1	388.90	533.34	675.71	13.28	11.94	48.51	201.43	230.93	277.83	0.00	0.00	
2	19.42	597.25	709.65	3.68	1,031.28	1,018.29	24.89	927.86	1,043.43	0.00	0.00	
3	874.33	925.35	969.89	574.06	363.44	382.78	1,131.76	1,137.78	1,049.96	0.00	0.00	
4	464.52	433.63	422.34	118.33	147.34	176.88	80.99	58.54	22.44	155.34	578.74	

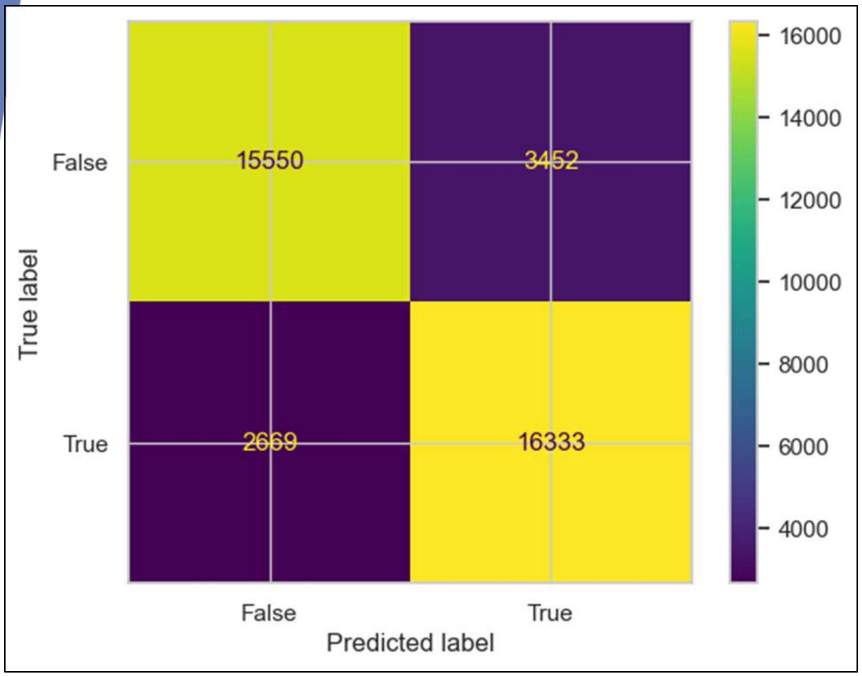
Model building with PCA



Logistic regression with PCA

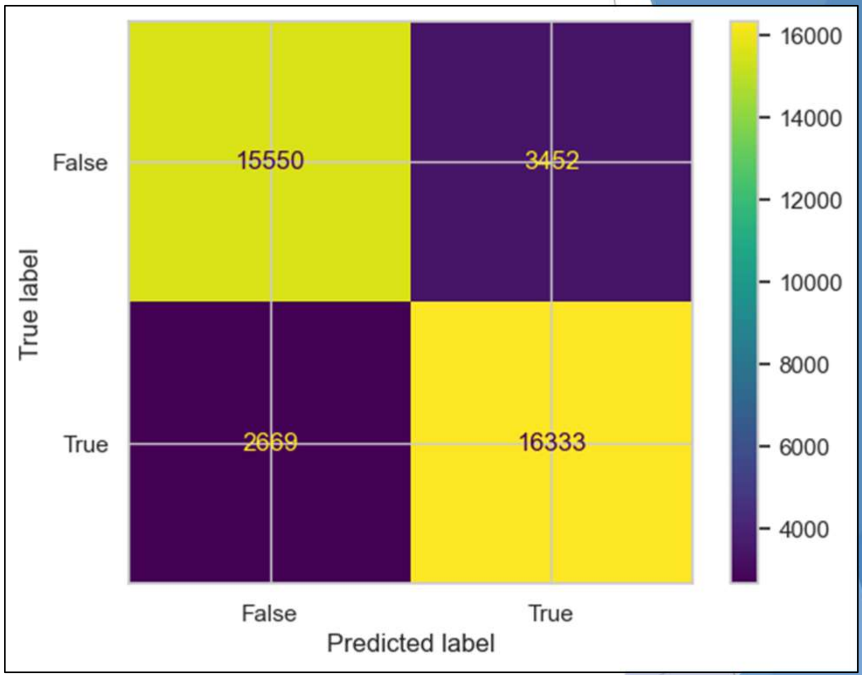


Prediction on the train set



Accuracy:- 0.84
Sensitivity:- 0.86
Specificity:- 0.82
Recall:- 0.86 AUC: 0.91

Prediction on the test set



Accuracy:- 0.82
Sensitivity:- 0.86
Specificity:- 0.82
Recall:- 0.86 AUC:- 0.89

Decision tree with PCA

	mean_fit_time	std_fit_time	mean_score_time	std_score_time	param_max_depth	param_min_samples_leaf	param_min_samples_split	params
0	1.75	0.02	0.01	0.00	5	50	50	{'max_depth': 5, 'min_samples_leaf': 50, 'min_samples_split': 50}
1	1.33	0.08	0.00	0.00	5	50	100	{'max_depth': 5, 'min_samples_leaf': 50, 'min_samples_split': 100}
2	1.29	0.02	0.00	0.00	5	100	50	{'max_depth': 5, 'min_samples_leaf': 100, 'min_samples_split': 50}
3	1.30	0.03	0.00	0.00	5	100	100	{'max_depth': 5, 'min_samples_leaf': 100

Prediction on the train set

Accuracy:- 0.87 Sensitivity:- 0.87 Specificity:- 0.87 Recall:- 0.87

Area under curve is: 0.87

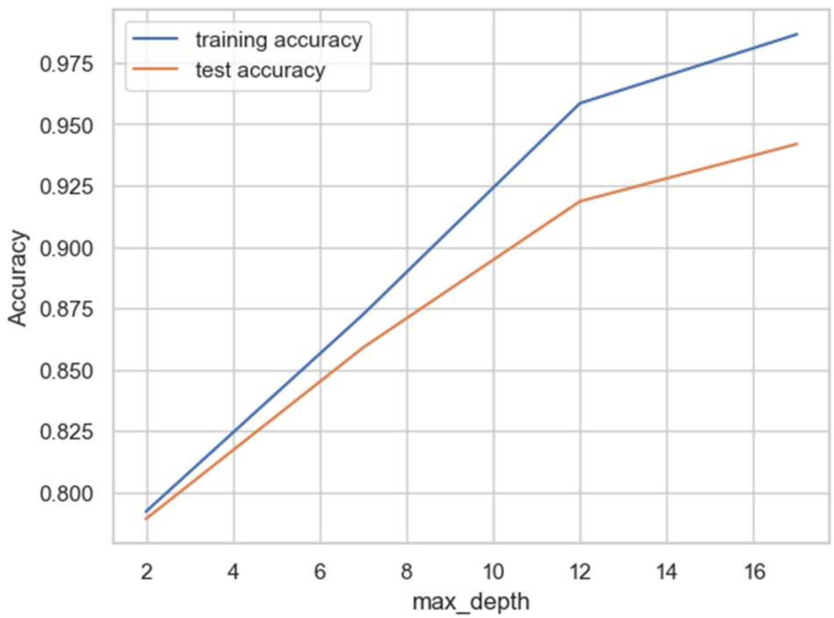
Prediction on the test set

Accuracy:- 0.84 Sensitivity:- 0.87 Specificity:- 0.87 Recall:- 0.87 Area

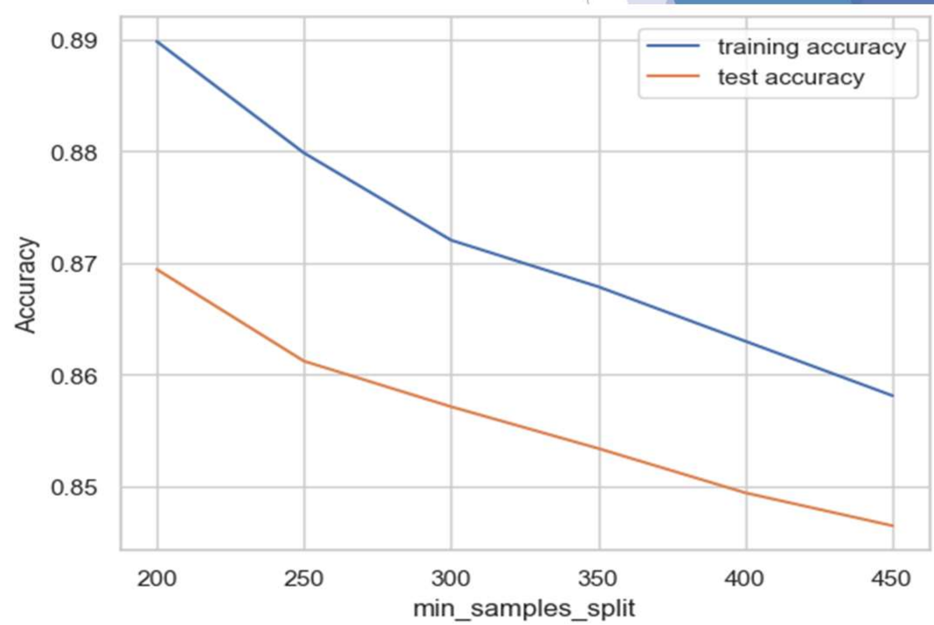
under curve is: 0.77

We can see from the model performance that the Sensitivity and Specificity remains same while evaluating the model on the test set and Train Set. However, the accuracy dropped a little in TEST set but still it is quite good in the test set.

Random forest with PCA



We see that as we increase the value of max_depth, both train and test scores increase till a point. The ensemble tries to overfit as we increase the max_depth.
Thus, controlling the depth of the constituent trees will help reduce overfitting in the forest.



Finally we find the optimal hyperparameters using GridSearchCV.

---0.866--- OOB Score tells how accurate will be our model, calculated the OOB score based on the train data set. Now, next we will also see the predictions and other metrics.

```
confusion matrix
[[7228 915] [ 222 545]]
Accuracy:- 0.87
sensitivity 0.71
specificity 0.89
AUC: 0.88
```

Logistic Regression without PCA

1.As we see there are Many features with high p-values and hence those are insignificant for our model.
2.Also, there are few features with negative coefficients as well.

Generalized Linear Model Regression Results

Dep. Variable:	Churn	No. Observations:	38004
Model:	GLM	Df Residuals:	37869
Model Family:	Binomial	Df Model:	134
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	nan
Date:	Sun, 05 Nov 2023	Deviance:	27114.
Time:	12:27:34	Pearson chi2:	2.72e+05
No. Iterations:	100	Pseudo R-squ. (CS):	nan
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.6388	0.052	12.347	0.000	0.537	0.740

Feature selection using RFE

Model-1

Generalized Linear Model Regression Results

Dep. Variable:	Churn	No. Observations:	38004
Model:	GLM	Df Residuals:	37988
Model Family:	Binomial	Df Model:	15
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	nan
Date:	Sun, 05 Nov 2023	Deviance:	30556.
Time:	12:38:03	Pearson chi2:	1.94e+08
No. Iterations:	13	Pseudo R-squ. (CS):	nan
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.4739	0.049	9.576	0.000	0.377	0.571
arpu_6	0.5393	0.023	23.797	0.000	0.495	0.584
onnet_mou_8	1.3359	0.065	20.554	0.000	1.209	1.463
std_og_t2m_mou_8	1.1846	0.059	20.141	0.000	1.069	1.300
og_others_8	-6.8052	3.118	-2.183	0.029	-12.916	-0.694
total_og_mou_8	-2.5263	0.099	-25.572	0.000	-2.720	-2.333
loc_ic_t2m_mou_7	0.8600	0.040	21.648	0.000	0.782	0.938
loc_ic_t2m_mou_8	-1.0917	0.092	-11.881	0.000	-1.272	-0.912
loc_ic_mou_8	-0.4331	0.096	-4.534	0.000	-0.620	-0.246
total_ic_mou_8	-0.7308	0.063	-11.671	0.000	-0.854	-0.608

```
VIF_CALC(X_train[rfe_cols])
```

	Features	VIF
4	total_og_mou_8	13.79
7	loc_ic_mou_8	8.71
1	onnet_mou_8	6.79
6	loc_ic_t2m_mou_8	5.79
8	total_ic_mou_8	5.70
2	std_og_t2m_mou_8	5.49
14	dec_avg_revenuePC_action	3.45
13	dec_rech_action	3.40
5	loc_ic_t2m_mou_7	2.29
10	total_rech_num_8	1.83
9	total_rech_num_6	1.67
0	arpu_6	1.39
11	last_day_rch_amt_8	1.25
12	max_rech_data_8	1.15
3	og_others_8	1.00

Removing column total_og_mou_8, which is insignificant as it has very high p-value

Model-2

```
10 log_no_pca_2.summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	Churn	No. Observations:	38004
Model:	GLM	Df Residuals:	37989
Model Family:	Binomial	Df Model:	14
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	nan
Date:	Sun, 05 Nov 2023	Deviance:	31405.
Time:	12:46:22	Pearson chi2:	5.01e+08
No. Iterations:	13	Pseudo R-squ. (CS):	nan
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.4162	0.051	8.099	0.000	0.315	0.517

```
3 VIF_CALC(X_train[log_cols])
```

	Features	VIF
6	loc_ic_mou_8	8.69
7	total_ic_mou_8	5.70
5	loc_ic_t2m_mou_8	5.58
13	dec_avg_revenuePC_action	3.44
12	dec_rech_action	3.39
4	loc_ic_t2m_mou_7	2.29
9	total_rech_num_8	1.78
8	total_rech_num_6	1.65
0	arpu_6	1.37
1	onnet_mou_8	1.27
10	last_day_rch_amt_8	1.23
2	std_og_t2m_mou_8	1.20
11	max_rech_data_8	1.15
3	og_others_8	1.00

Removing column loc_ic_mou_8, which is insignificant as it has very high p-value and high VIF

Model-3

Dep. Variable:	Churn	No. Observations:	38004
Model:	GLM	Df Residuals:	37990
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	nan
Date:	Sun, 05 Nov 2023	Deviance:	31440.
Time:	12:46:56	Pearson chi2:	5.41e+08
No. Iterations:	13	Pseudo R-squ. (CS):	nan
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.4246	0.051	8.292	0.000	0.324	0.525

	Features	VIF
5	loc_ic_t2m_mou_8	4.14
12	dec_avg_revenuePC_action	3.44
11	dec_rech_action	3.39
6	total_ic_mou_8	2.93
4	loc_ic_t2m_mou_7	2.28
8	total_rech_num_8	1.78
7	total_rech_num_6	1.65
0	arpu_6	1.37
1	onnet_mou_8	1.27
9	last_day_rch_amt_8	1.23
2	std_og_t2m_mou_8	1.19
10	max_rech_data_8	1.15

Removing column dec_avg_revenue PC_action, which is insignificant as it has very high p-value

Model-4

Dep. Variable:	Churn	No. Observations:	38004
Model:	GLM	Df Residuals:	37991
Model Family:	Binomial	Df Model:	12
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	nan
Date:	Sun, 05 Nov 2023	Deviance:	31750.
Time:	12:47:33	Pearson chi2:	2.26e+08
No. Iterations:	13	Pseudo R-squ. (CS):	nan
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1547	0.048	3.223	0.001	0.061	0.249

	Features	VIF
5	loc_ic_t2m_mou_8	4.14
6	total_ic_mou_8	2.92
4	loc_ic_t2m_mou_7	2.28
8	total_rech_num_8	1.77
7	total_rech_num_6	1.64
0	arpu_6	1.33
1	onnet_mou_8	1.26
9	last_day_rch_amt_8	1.21
2	std_og_t2m_mou_8	1.18
10	max_rech_data_8	1.15
11	dec_rech_action	1.14
3	og_others_8	1.00

Removing total_ic_mou_8 as it still has high p-Value

Model-5

Dep. Variable:	Churn	No. Observations:	38004
Model:	GLM	Df Residuals:	37992
Model Family:	Binomial	Df Model:	11
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	nan
Date:	Sun, 05 Nov 2023	Deviance:	32241.
Time:	12:48:11	Pearson chi2:	3.99e+08
No. Iterations:	13	Pseudo R-squ. (CS):	nan
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]

	Features	VIF
5	loc_ic_t2m_mou_8	2.46
4	loc_ic_t2m_mou_7	2.28
7	total_rech_num_8	1.76
6	total_rech_num_6	1.64
0	arpu_6	1.33
1	onnet_mou_8	1.24
8	last_day_rch_amt_8	1.21
2	std_og_t2m_mou_8	1.18
9	max_rech_data_8	1.14
10	dec_rech_action	1.14
3	og_others_8	1.00

Removing total_rech_num_8 due to high VIF

Model-6

Dep. Variable:	Churn	No. Observations:	38004
Model:	GLM	Df Residuals:	37993
Model Family:	Binomial	Df Model:	10
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	nan
Date:	Sun, 05 Nov 2023	Deviance:	33459.
Time:	12:49:04	Pearson chi2:	8.12e+06
No. Iterations:	13	Pseudo R-squ. (CS):	nan
Covariance Type:	nonrobust		

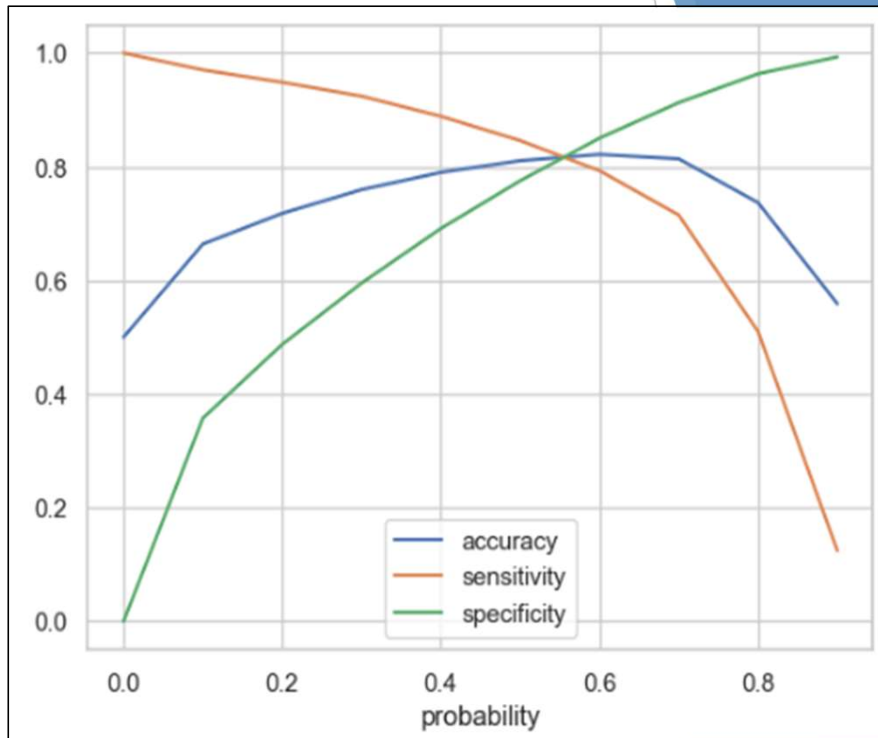
	coef	std err	z	P>|z|	[0.025	0.975]

	Features	VIF
5	loc_ic_t2m_mou_8	2.37
4	loc_ic_t2m_mou_7	2.28
6	total_rech_num_6	1.32
0	arpu_6	1.27
7	last_day_rch_amt_8	1.20
8	max_rech_data_8	1.13
9	dec_rech_action	1.09
1	onnet_mou_8	1.05
2	std_og_t2m_mou_8	1.04
3	og_others_8	1.00

Here we see the p-values are in the Acceptable Range also the VIF's of all the values are also below 5 which is a good and acceptable range. Hence Model-6 will be the final Model

Calculation of the accuracy sensitivity and specificity for various probability cutoffs.

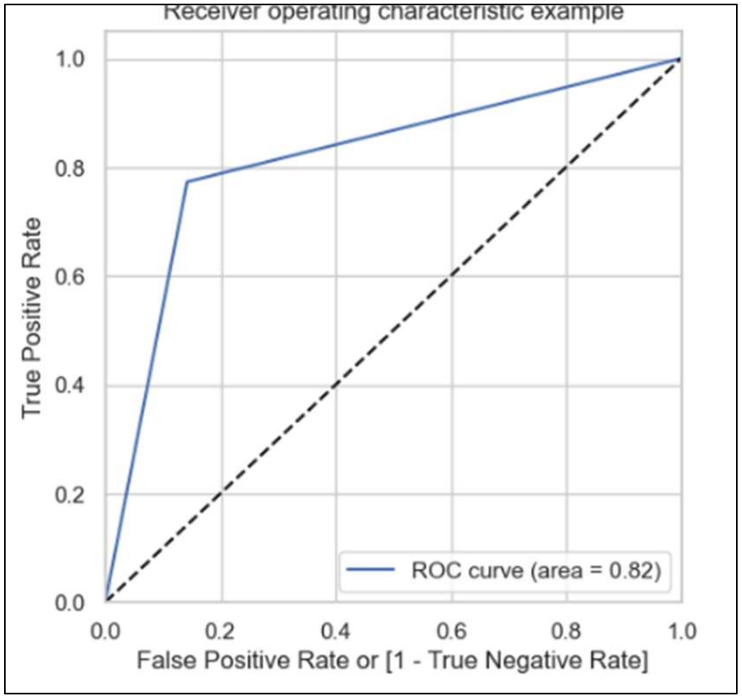
	probability	accuracy	sensitivity	specificity
0.00	0.00	0.50	1.00	0.00
0.10	0.10	0.66	0.97	0.36
0.20	0.20	0.72	0.95	0.49
0.30	0.30	0.76	0.92	0.60
0.40	0.40	0.79	0.89	0.69
0.50	0.50	0.81	0.85	0.78
0.60	0.60	0.82	0.79	0.85
0.70	0.70	0.81	0.71	0.91
0.80	0.80	0.74	0.51	0.96
0.90	0.90	0.56	0.12	0.99



Analysis of the above curve
Accuracy - Becomes stable around 0.6 approx.
Sensitivity - Decreases with the increased probability.
Specificity - Increases with the increasing probability.
Hence we consider cutoff point to be 0.6

EVALUATION METRICS

As we can see we from the above ROC plot we get AUC of 0.82.



Model summary (Logistic Regression Without PCA)

- **Train set**
- Accuracy:- 0.82
- Sensitivity:- 0.79
- Specificity:- 0.85
- Recall:- 0.79
- **Test set**
- Accuracy:- 0.85
- Sensitivity:- 0.79
- Specificity:- 0.85
- Recall:- 0.79

Conclusion

In conclusion and regarding the upcoming strategy, our exploratory data analysis (EDA) has unveiled a notable decline in recharges, call usage, and data usage during the 8th month, which corresponds to the Action Phase. Key findings include the following significant features:

- loc_og_t2m_mou_7
- total_og_mou_6
- loc_og_t2t_mou_7
- roam_ic_mou_7
- onnet_mou_7
- arpu_7
- loc_og_t2c_mou_7
- onnet_mou_8
- roam_og_mou_8
- arpu_6

Of particular importance is the Average Revenue Per User (ARPU) in the 7th month, which plays a critical role in determining churn. A sudden drop in ARPU may indicate that a customer is contemplating churn, necessitating appropriate action.

The most influential factors contributing to customer churn are local outgoing minutes of usage. Roaming minutes of usage (both incoming and outgoing) also significantly affect churn, as does the total outgoing minutes of usage.

To address these findings, the following strategies can be implemented:

•A sudden decline in local outgoing minutes of usage could be attributed to subpar customer service, network issues, or inappropriate customer plans. Efforts should be directed towards improving network quality and enhancing customer satisfaction.

•Based on usage patterns, recent recharges, and on-net usage, regular feedback calls should be conducted to gauge customer satisfaction and understand their concerns and expectations. Appropriate measures should be taken to mitigate churn risks.

•Introducing attractive offers to customers experiencing a sudden decrease in their expenditure on calls and data during the Action Phase can entice them to stay.

•Tailored plans should be offered to such customers to retain them and prevent churn.

•Promotional offers can also be an effective tool in retaining customers and reducing churn.

THANK YOU