TELECOM CHURN CASE STUDY

PROBLEM STATEMENT

- In the dynamic landscape of the telecom industry, customers have the freedom to choose among various service providers, leading to active switching between operators. This competitive environment contributes to an annual churn rate ranging from 15-25% within the telecommunications sector. Recognizing that the cost of acquiring new customers far exceeds that of retaining existing ones, customer retention has evolved into a paramount objective, surpassing customer acquisition in significance.

- For incumbent operators, the foremost business goal is to retain high-profit customers, acknowledging their substantial contribution to revenue.

- In response to the challenge of customer churn, telecom companies are compelled to proactively identify customers at a heightened risk of churn. This necessitates a keen understanding of the primary indicators that signal potential churn, enabling companies to implement effective retention strategies.

STEPS INVOLVED :

- Data Reading
- Data Understanding
- Data Cleaning
- Exploratory Data Analysis(EDA)
- Data Preparation
- Model Building
- Conclusion

DATA UNDERSTANDING

```
tele_data.head()
```

	mobile_number	circle_id	loc_og_t2o_mou	std_og_t2o_mou	loc_ic_t2o_mou	last_date_of_month_6	last_date_of_month_7	la
0	7000842753	109	0.0	0.0	0.0	6/30/2014	7/31/2014	
1	7001865778	109	0.0	0.0	0.0	6/30/2014	7/31/2014	
2	7001625959	109	0.0	0.0	0.0	6/30/2014	7/31/2014	
3	7001204172	109	0.0	0.0	0.0	6/30/2014	7/31/2014	
4	7000142493	109	0.0	0.0	0.0	6/30/2014	7/31/2014	

```
tele_data.shape
```

```
(99999, 226)
```

```
tele_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99999 entries, 0 to 99998
Columns: 226 entries, mobile_number to sep_vbc_3g
dtypes: float64(179), int64(35), object(12)
memory usage: 172.4+ MB
```

```
tele_data.describe()
```

	mobile_number	circle_id	loc_og_t2o_mou	std_og_t2o_mou	loc_ic_t2o_mou	arpu_6	arpu_7	arpu_8
count	9.999900e+04	99999.0	98981.0	98981.0	98981.0	99999.000000	99999.000000	99999.000000
mean	7.001207e+09	109.0	0.0	0.0	0.0	282.987358	278.536648	279.154731
std	6.956694e+05	0.0	0.0	0.0	0.0	328.439770	338.156291	344.474791
min	7.000000e+09	109.0	0.0	0.0	0.0	-2258.709000	-2014.045000	-945.808000
25%	7.000606e+09	109.0	0.0	0.0	0.0	93.411500	86.980500	84.126000
50%	7.001205e+09	109.0	0.0	0.0	0.0	197.704000	191.640000	192.080000
75%	7.001812e+09	109.0	0.0	0.0	0.0	371.060000	365.344500	369.370500
max	7.002411e+09	109.0	0.0	0.0	0.0	27731.088000	35145.834000	33543.624000

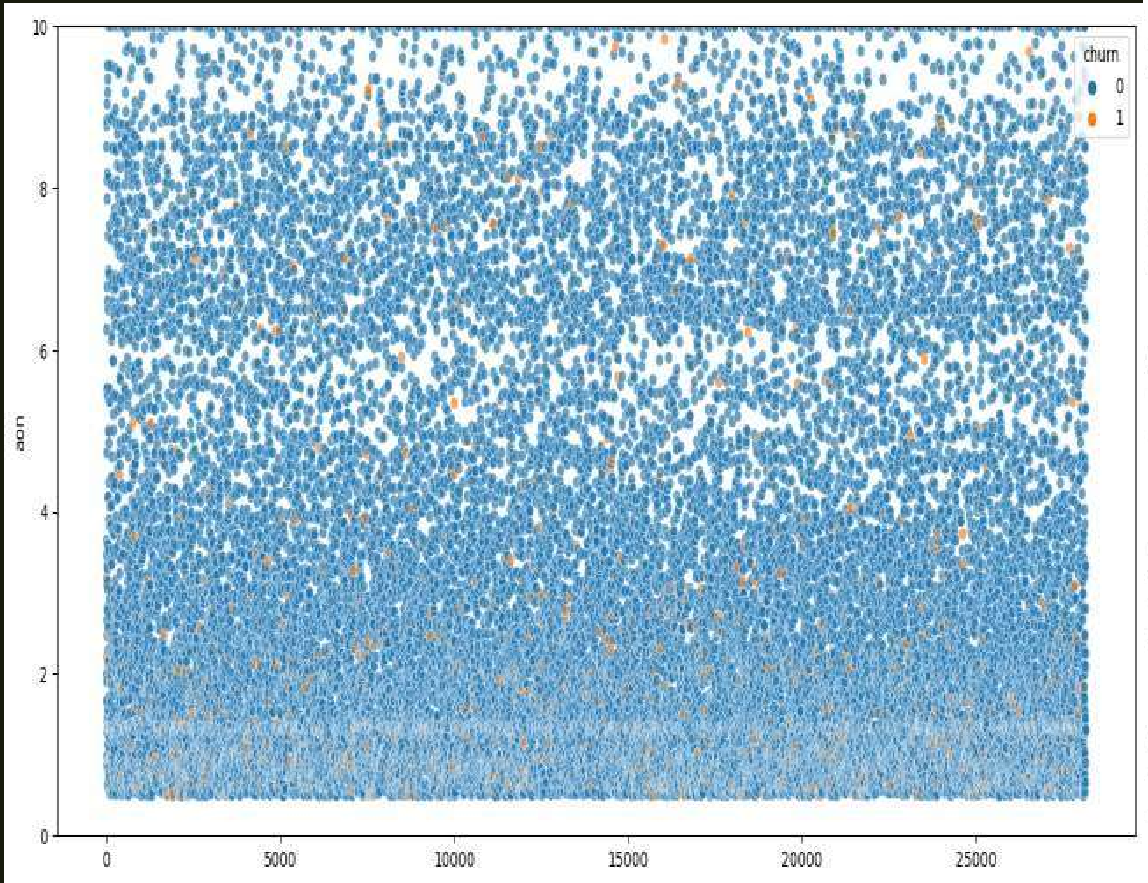
DATA CLEANING

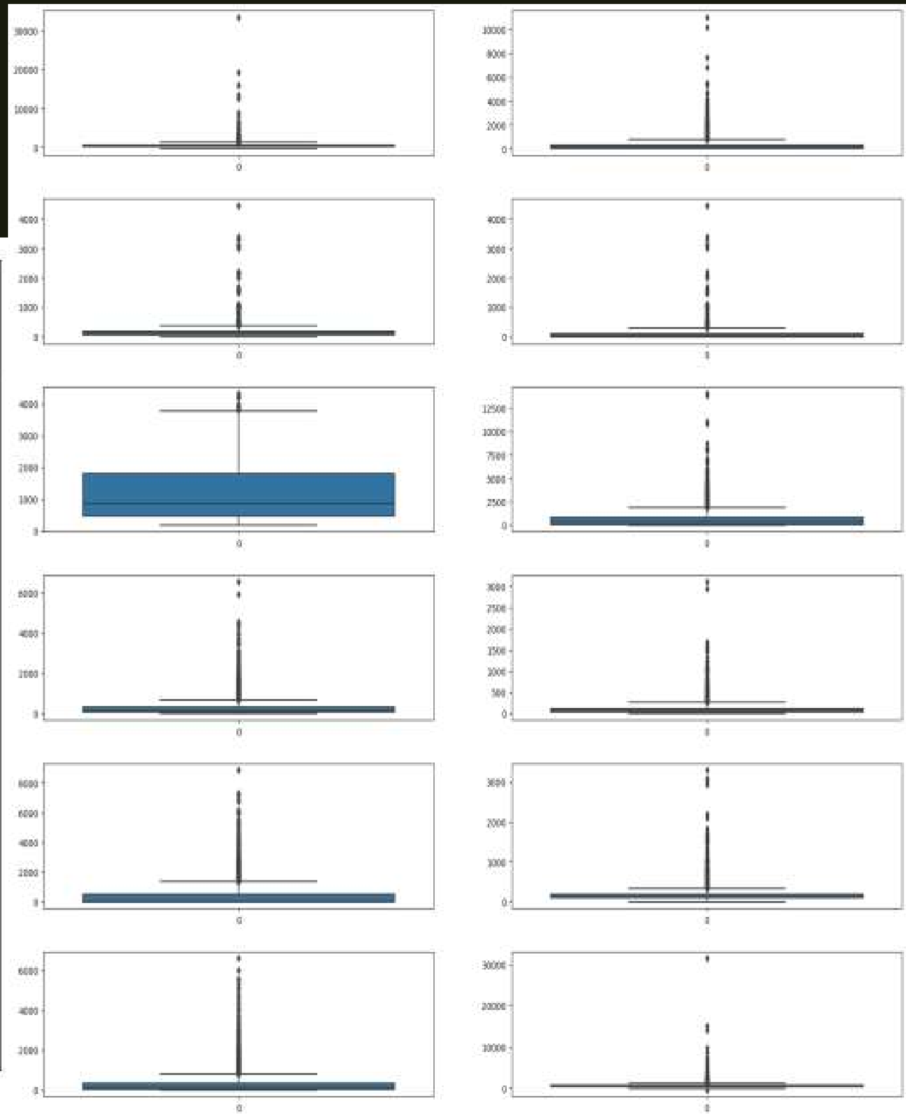
- Drop columns with a missing value percentage exceeding 40%.
- Exclude the columns with high correlation.
- Deriving New Features from the exisitng features by Generate a total MOU by combining offnet and onnet, and then dropping the individual columns.
- After dropping all the irrelevant or redundant columns the shape of the data remains to be :

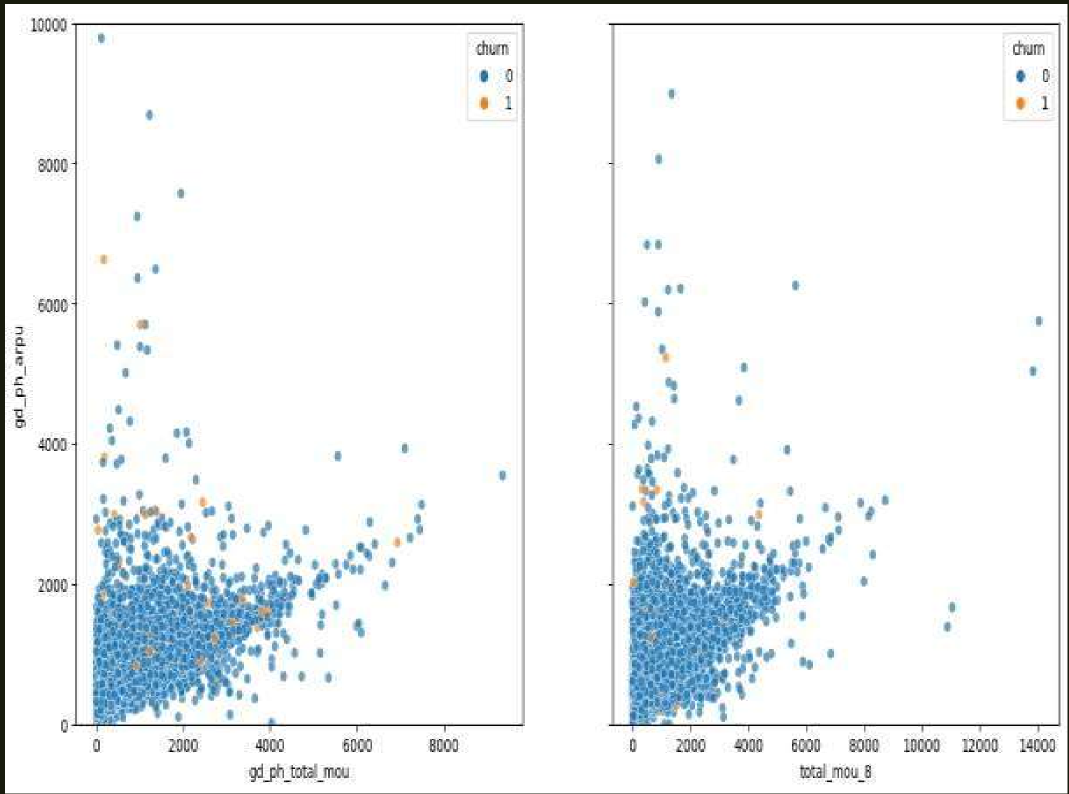
```
churn_data.shape
```

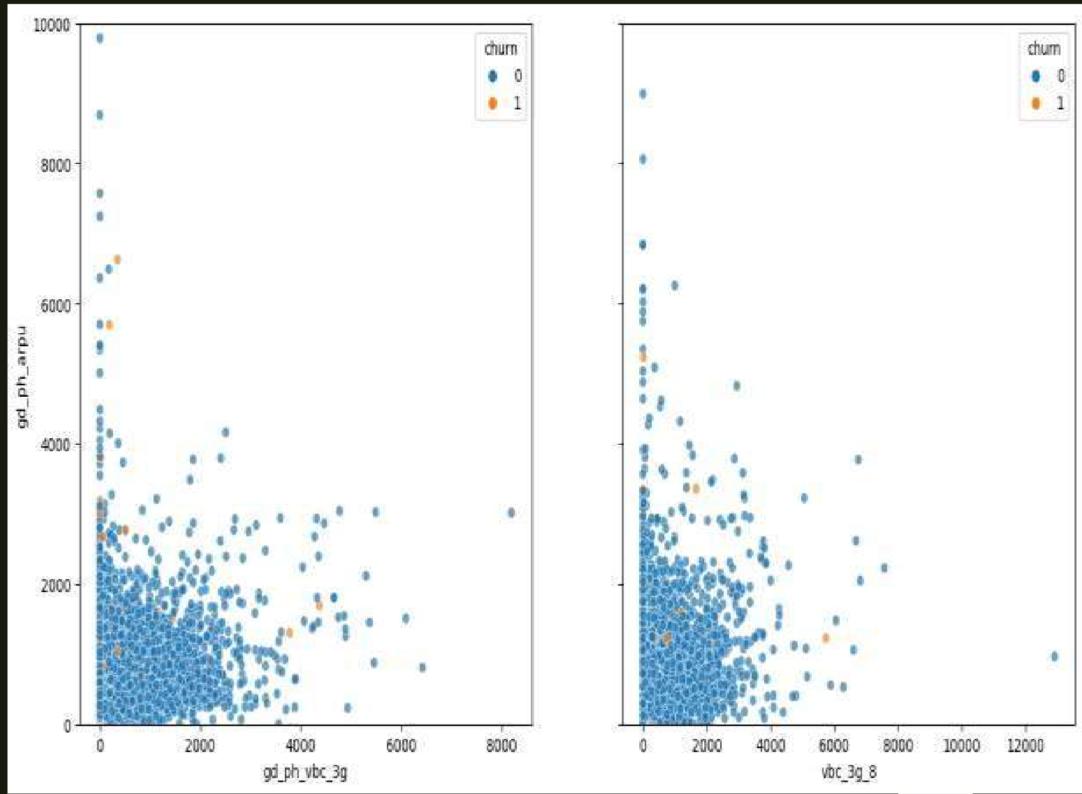
```
(28163, 56)
```

EDA:



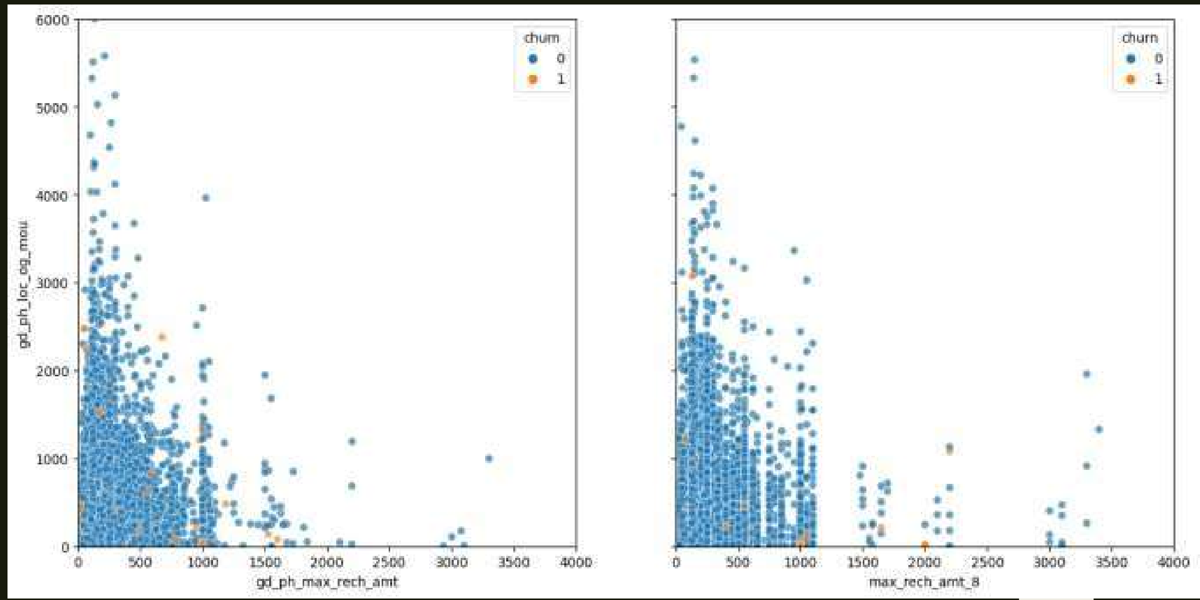


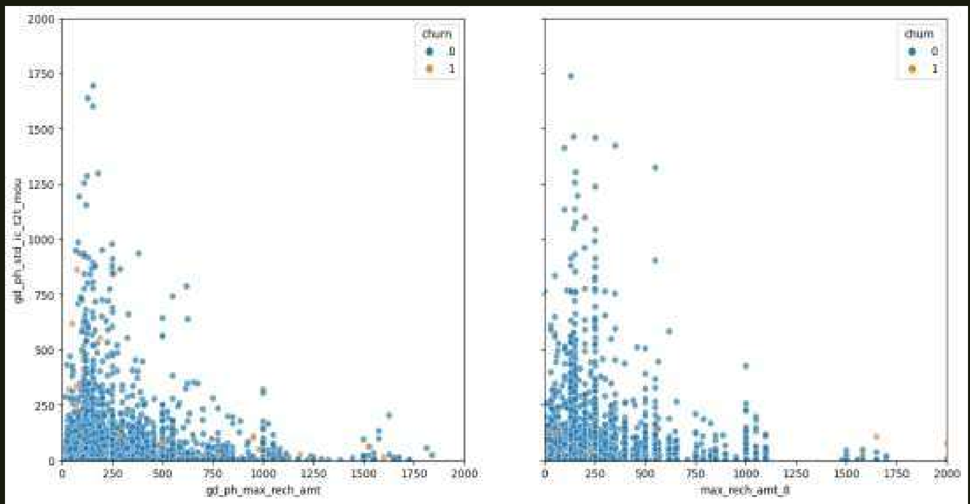


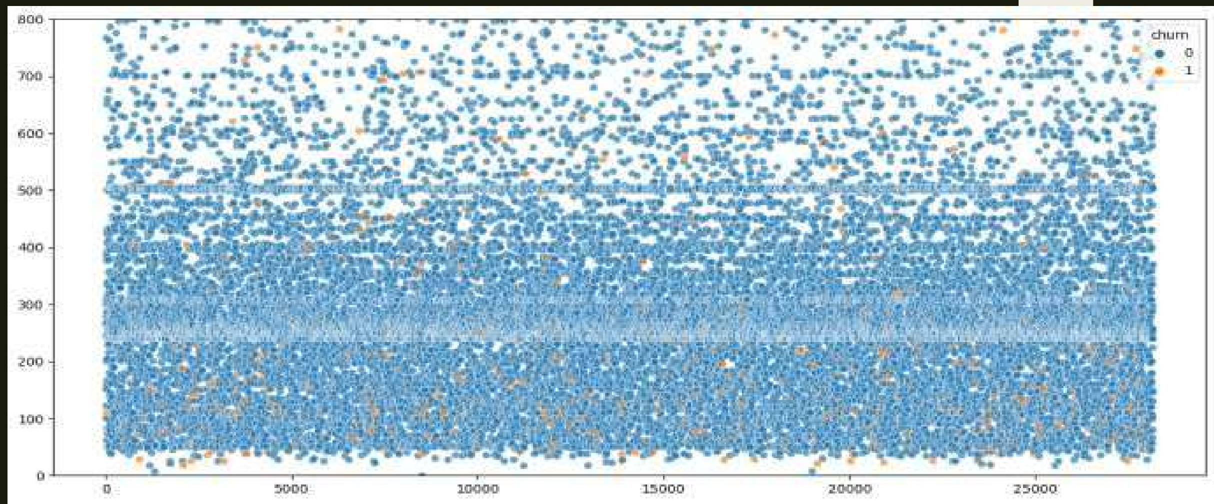


- It is evident that MOU has substantially decreased for churners in the action phase (8th month), resulting in a decline in the generated revenue. It's intriguing to note that even though the MOU is between 0-2000, the revenue is highest in that range, indicating that these users likely had other services contributing to the overall revenue.
- It's noticeable that users who utilized a minimal amount of VBC data but still generated high revenue eventually churned. Once more, we observe that the revenue tends to be higher on the lower consumption side.

- Users who recharged with higher amounts tended to use the service for local purposes less compared to users who recharged with lesser amounts. Intuitively, individuals whose maximum recharge amount and local outgoing usage were very low, even in the good phase, exhibited a higher churn rate.
- we can see that users who had the max recharge amount less tha 200 churned more
- Users who experienced higher maximum recharge amounts and yet had low incoming call minutes of usage during the good phase tended to exhibit a higher churn rate







MODEL BUILDING:

Logistic Regression :
Using RFE for feature Selection

MODEL 1:

	Features	VIF
17	gd_ph_total_mou	91.27
19	gd_ph_std_og_mou	82.36
16	gd_ph_loc_og_mou	24.76
15	total_mou_8	4.40
2	loc_og_mou_8	4.33
0	arpu_8	4.28
3	loc_ic_mou_8	3.76
22	gd_ph_loc_ic_mou	3.40
6	total_rech_num_8	3.30
23	gd_ph_roam_og_mou	3.01
7	max_rech_amt_8	2.65
21	gd_ph_total_rech_num	2.45
8	last_day_rch_amt_8	1.98
12	monthly_3g_8	1.93
9	vol_2g_mb_8	1.84

- Here gd_ph_total_mou has high VIF , so we will drop it and before dropping the accuracy : 0.7965

MODEL 2:

	Features	VIF
15	total_mou_8	4.39
2	loc_og_mou_8	4.33
0	arpu_8	4.17
3	loc_ic_mou_8	3.75
19	gd_ph_loc_ic_mou	3.40
23	gd_ph_loc_og_mou	3.33
6	total_rech_num_8	3.28
7	max_rech_amt_8	2.65
16	gd_ph_std_og_mou	2.45
22	gd_ph_total_rech_num	2.44
8	last_day_rch_amt_8	1.98
12	monthly_3g_8	1.92
9	vol_2g_mb_8	1.84
21	gd_ph_monthly_3g	1.82
13	sachet_3g_8	1.70

- Here, total_mou_8 Has high VIF, we will drop it and the accuracy is: 0.7962

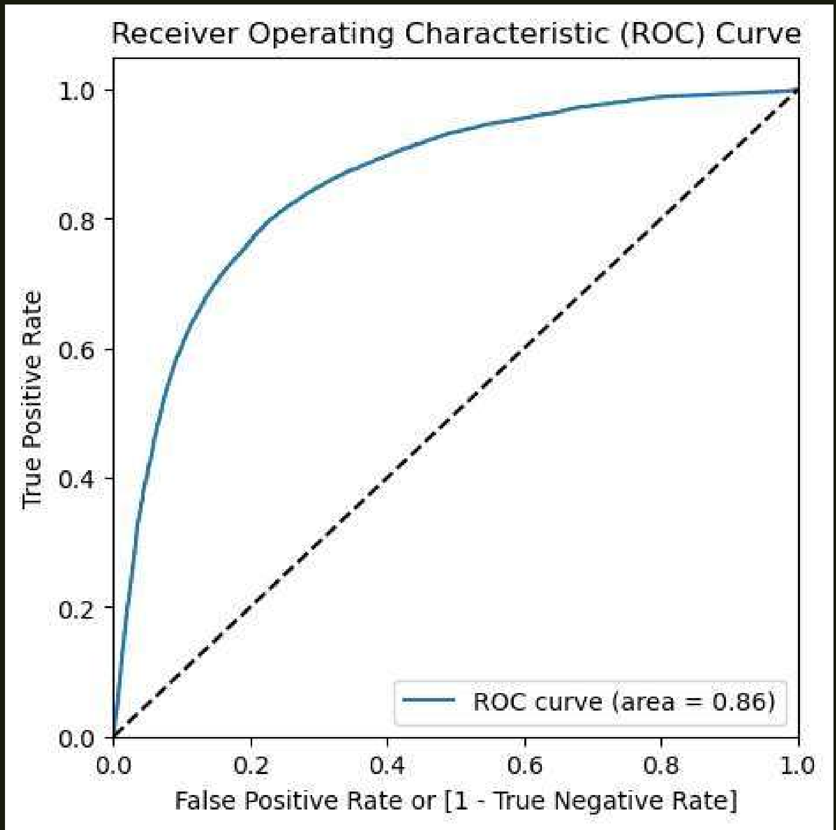
MODEL 3:

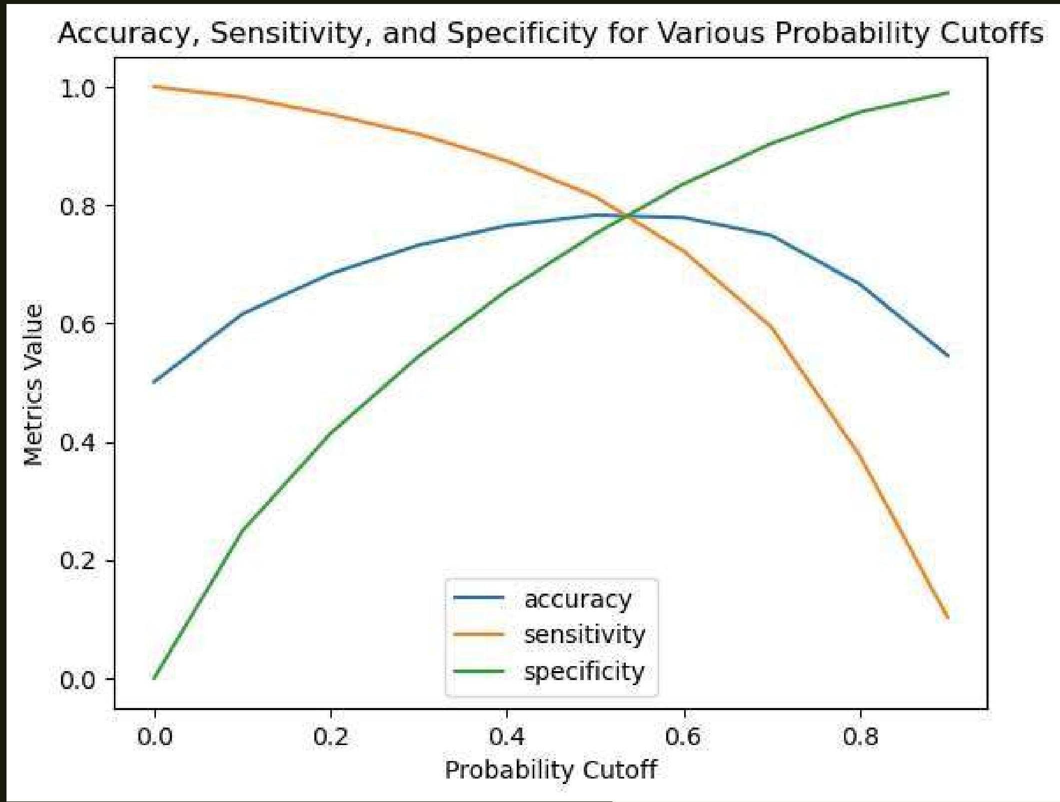
3	loc_ic_mou_8	3.75
18	gd_ph_loc_ic_mou	3.40
0	arpu_8	3.35
22	gd_ph_loc_og_mou	3.32
6	total_rech_num_8	2.87
7	max_rech_amt_8	2.62
21	gd_ph_total_rech_num	2.22
8	last_day_rch_amt_8	1.98
12	monthly_3g_8	1.89
9	vol_2g_mb_8	1.84
20	gd_ph_monthly_3g	1.82
13	sachet_3g_8	1.68
17	gd_ph_sachet_3g	1.64
19	gd_ph_vol_2g_mb	1.63

Here,loc_ic_mou_8 has high VIF.

Accuracy is :0.7916

- Since Accuracy is 0.78 and we have columns with less than 0.5 VIF, so we will stop it here and our optimal cutoff is 0.5.





- And our accuracy for test data is 0.78

DECISION TREE:

```
# Train Accuracy
y_train_pred = initial_dt.predict(X_train)
print(f'Train accuracy : {metrics.accuracy_score(y_train, y_train_pred)}')
```

Train accuracy : 0.8776070970612095

```
# Print the classification report on test data
print(metrics.classification_report(y_test, y_test_pred))
```

	precision	recall	f1-score	support
0	0.92	0.87	0.90	8215
1	0.88	0.92	0.90	8162
accuracy			0.90	16377
macro avg	0.90	0.90	0.90	16377
weighted avg	0.90	0.90	0.90	16377

RANDOM FOREST

Train accuracy : 0.9161018501557062

	precision	recall	f1-score	support
0	0.92	0.86	0.89	8215
1	0.87	0.92	0.90	8162
accuracy			0.89	16377
macro avg	0.89	0.89	0.89	16377
weighted avg	0.89	0.89	0.89	16377

- Since Random Forest gives us the highest Accuracy and recall, we will conider Random Forest as the final model.

CONCLUSION

- Considering our business problem of customer retention, prioritizing higher recall is crucial. Identifying potential churners accurately is more cost-effective than losing a customer and acquiring new ones.
- Upon comparing the trained models, it is evident that the tuned Random Forest model perform exceptionally well, achieving the highest accuracy 91%. In this context, we choose the Random Forest model due to its simplicity and comparable performance

Strategies to Manage Customer Churn

- The top 10 predictors are as follows:
  1. loc_og_mou_8
  2. total_rech_num_8
  3. monthly_3g_8
  4. monthly_2g_8
  5. gd_ph_loc_og_mou
  6. gd_ph_total_rech_num
  7. last_day_rch_amt_8
  8. std_ic_t2t_mou_8
  9. sachet_2g_8
  10. aon
- It's noticeable that the majority of these predictors are from the action phase, indicating a significant drop in engagement during that period.
- Further insights from EDA suggest additional strategies:
  1. Users with a maximum recharge amount less than 200, even in the good phase, should be tagged and re-evaluated periodically as they are more likely to churn.
  2. Users with less than 4 years of association with the network should be monitored regularly, as data shows that users with less than 4 years of association tend to churn more.
  3. While MOU is a major factor, special attention should be given to VBC data, especially if the user is not using a data pack.

BUSINESS INSIGHTS:

- The telecom company should prioritize roaming rates and consider offering attractive deals to customers using services from roaming zones.
- There is a need for a closer look at STD and ISD rates. If these rates are relatively high, providing special STD and ISD packages might help in retaining customers.
- Pay attention to users with incoming calls from fixed lines deviating 1.27 standard deviations below the average. This group is at a higher risk of churning, and strategies to engage and retain these users should be explored.

THANK YOU