Telecom Churn Case Study

Importing the Dataset

- ***This telecom dataset has 99999 rows and 226 columns.***

Handling missing values

• Let us conside the column `date_of_last_rech_data` indicating the date of the last recharge made in any given month for mobile internet. Here it can deduced if the `total_rech_data` and the `max_rech_data` also has missing values, the missing values in all the columns mentioned can be considered as meaningful missing.

•Hence imputing 0 as their values.

•Meaningfull missing in this case represents the the customer has not done any recharge for mobile interenet.

Handling the missing values for the attributes `count_rech_2g_*`,`count_rech_3g_*` for month 6,7,8 and 9.

- *From the above tablular the column values of `total_rech_data` for each month from 6 to 9 respectively is the sum of the columns values of `count_rech_2g` for each month from 6 to 9 respectively and `count_rech_3g` for each month from 6 to 9 respectively, which derives to a multicollinearity issue. In order to reduce the multicollinearity, we can drop the columns `count_rech_2g` for each month from 6 to 9 respectively and `count_rech_3g` for each month from 6 to 9 respectively.*

Handling the missing values for the attributes `arpu_3g_*`,`arpu_2g_*` for month 6,7,8 and 9

- From the above correlation table between attributes `arpu_2g_*` and `arpu_3g_*` for each month from 6 to 9 respectively is highly correlated to the attribute `av_rech_amt_data_*` for each month from 6 to 9 respectively.

Considering the high correlation between them, it is safer to drop the attributes `arpu_2g_*` and `arpu_3g_*`.

Handling the missing values for the attributes `av_rech_amt_data_*` for month 6,7,8 and 9

- From the above tabular it is deduced that the missing values for the column `av_rech_amt_data_*` for each month from 6 to 9 can be replaced as 0 if the `total_rech_data_*` for each month from 6 to 9 respectively is 0. i.e. if the total recharge done is 0 then the average recharge amount shall also be 0.

- ***Since the columns used to determine the High Value Customer is clear of null values, we can filter the overall data and then handle the remaining missing values for each column.***

Filtering the High Value Customer from Good Phase

- **`The 70th quantile value to determine the High Value Customer is: 478.0.`**

- The total number of customers is now limited to ~30k who lies under the High Value customer criteria basen upon which the model is built.

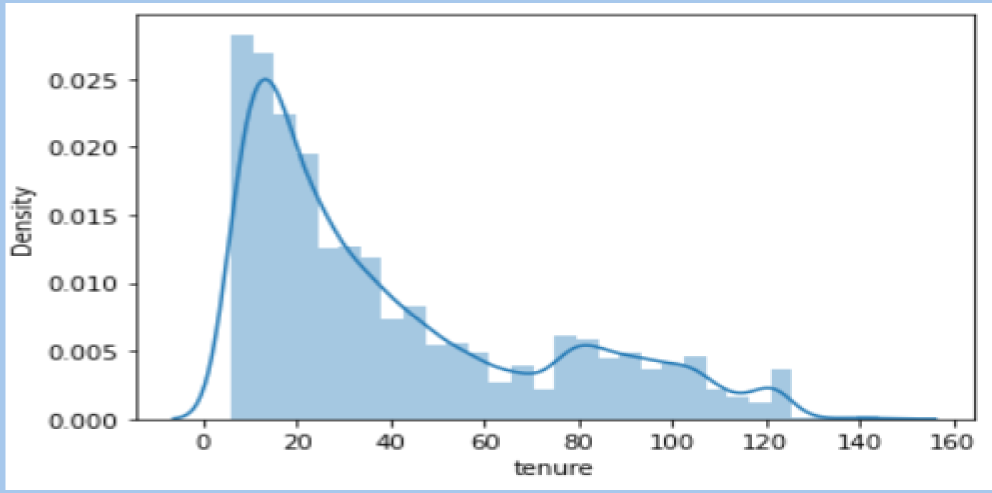
Defining Churn variable

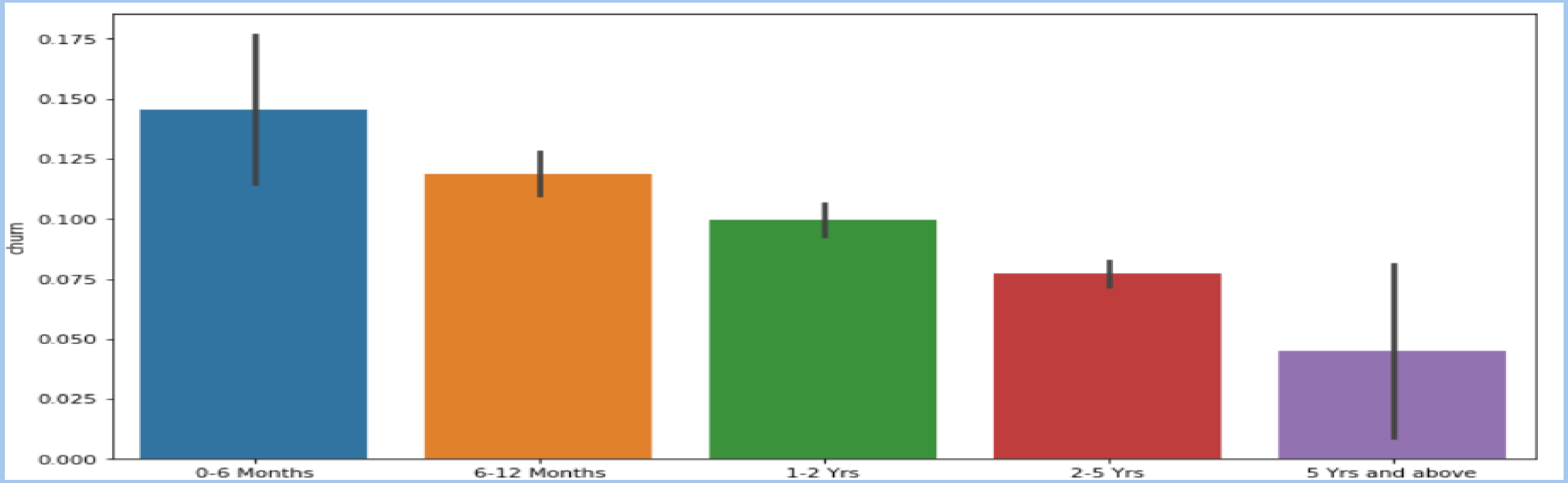
For that, we need to find the derive churn variable using `total_ic_mou_9`,`total_og_mou_9`, `vol_2g_mb_9` and `vol_3g_mb_9` attributes

```
0     91.863605
1      8.136395
Name: churn, dtype: float64
```

- ***As we can see that 91% of the customers do not churn, there is a possibility of class imbalance.***

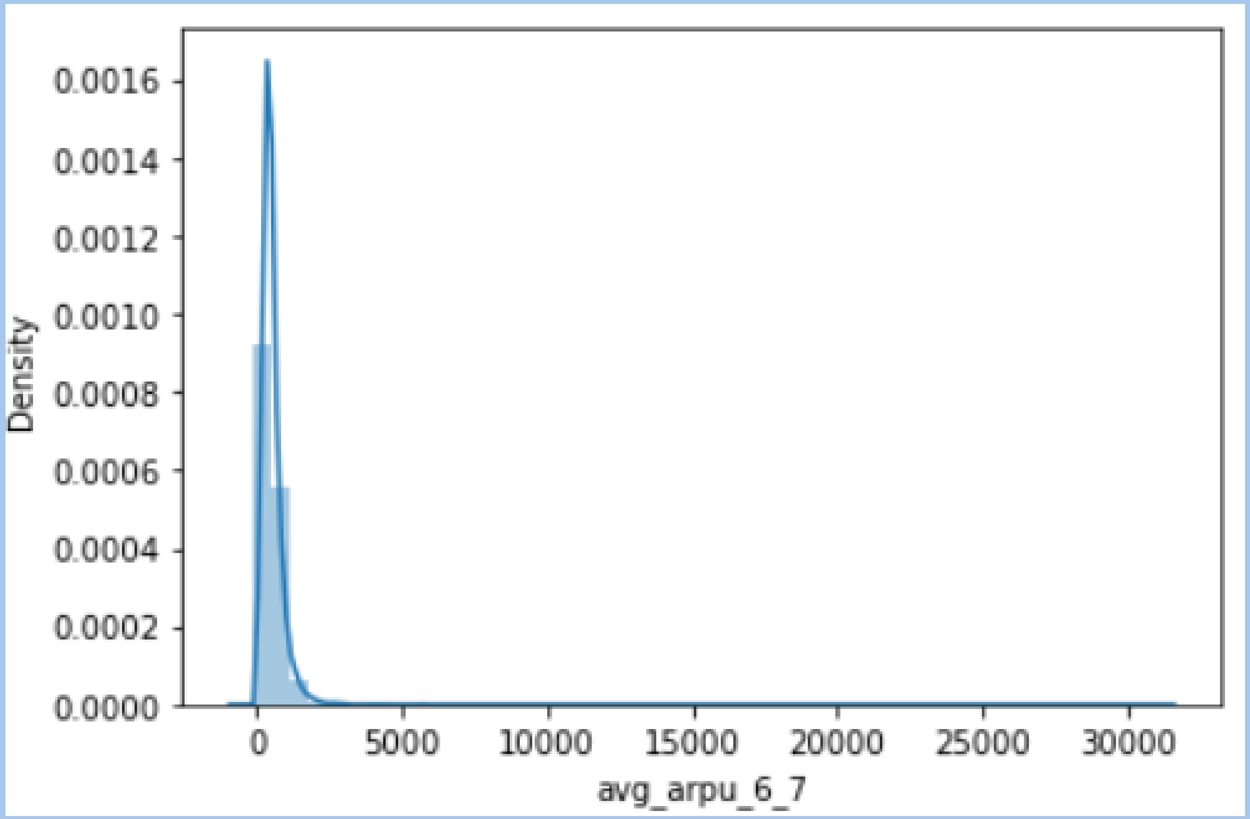
Deriving new variables to understand the data

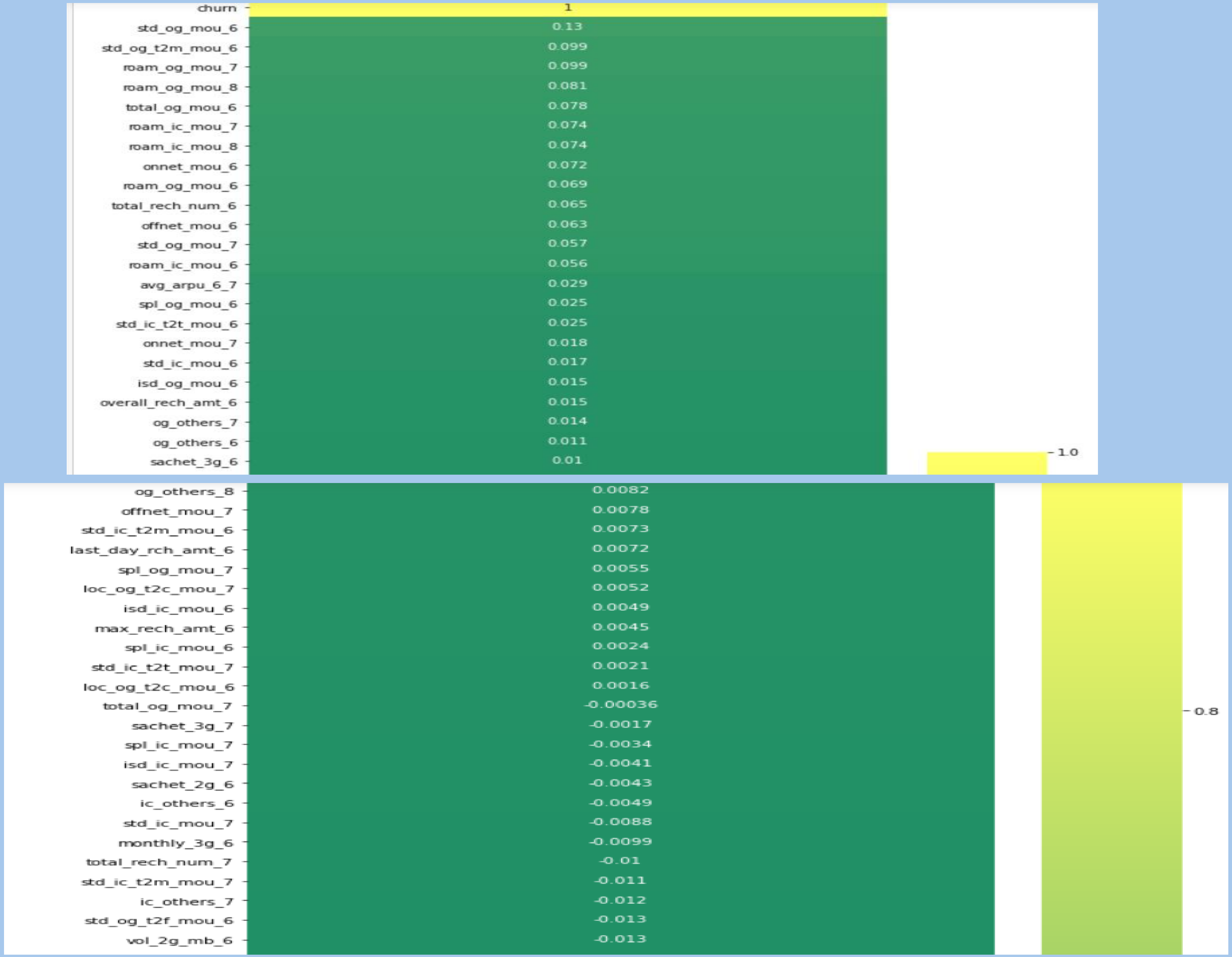


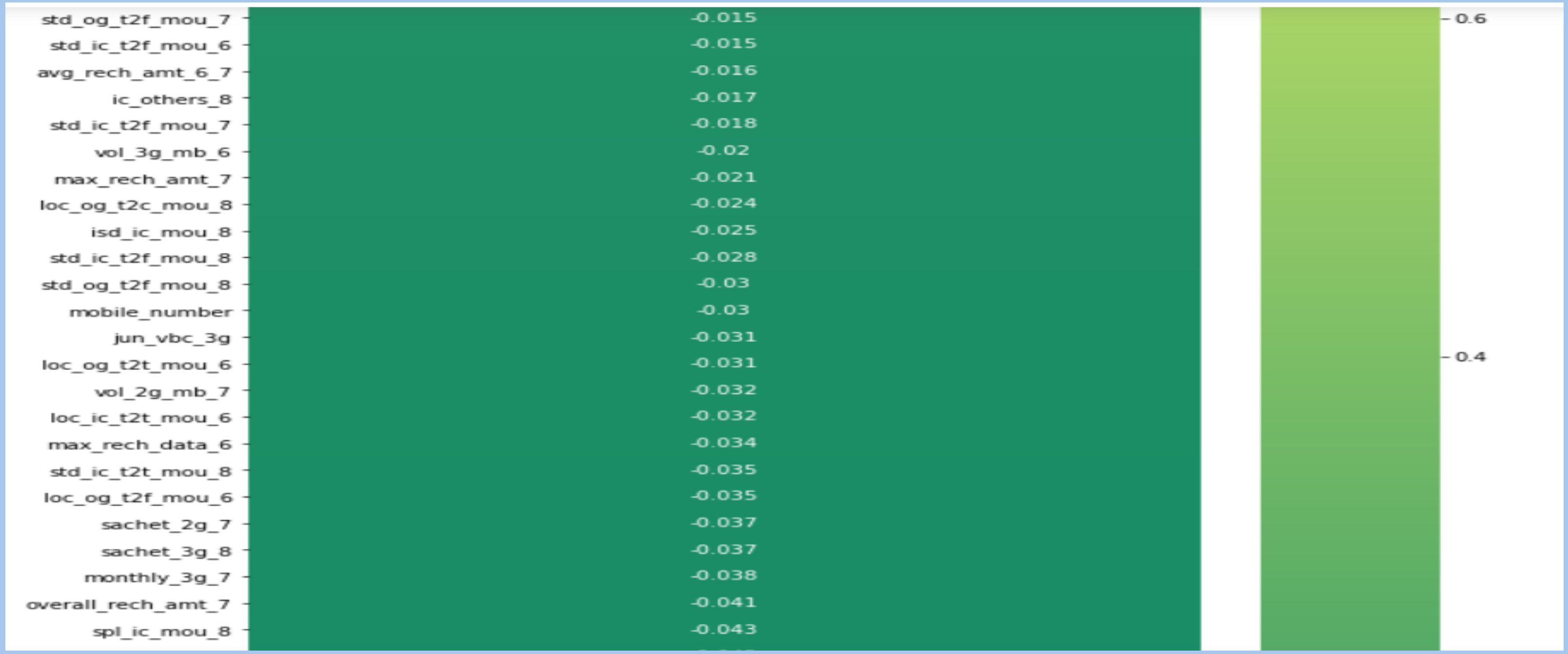
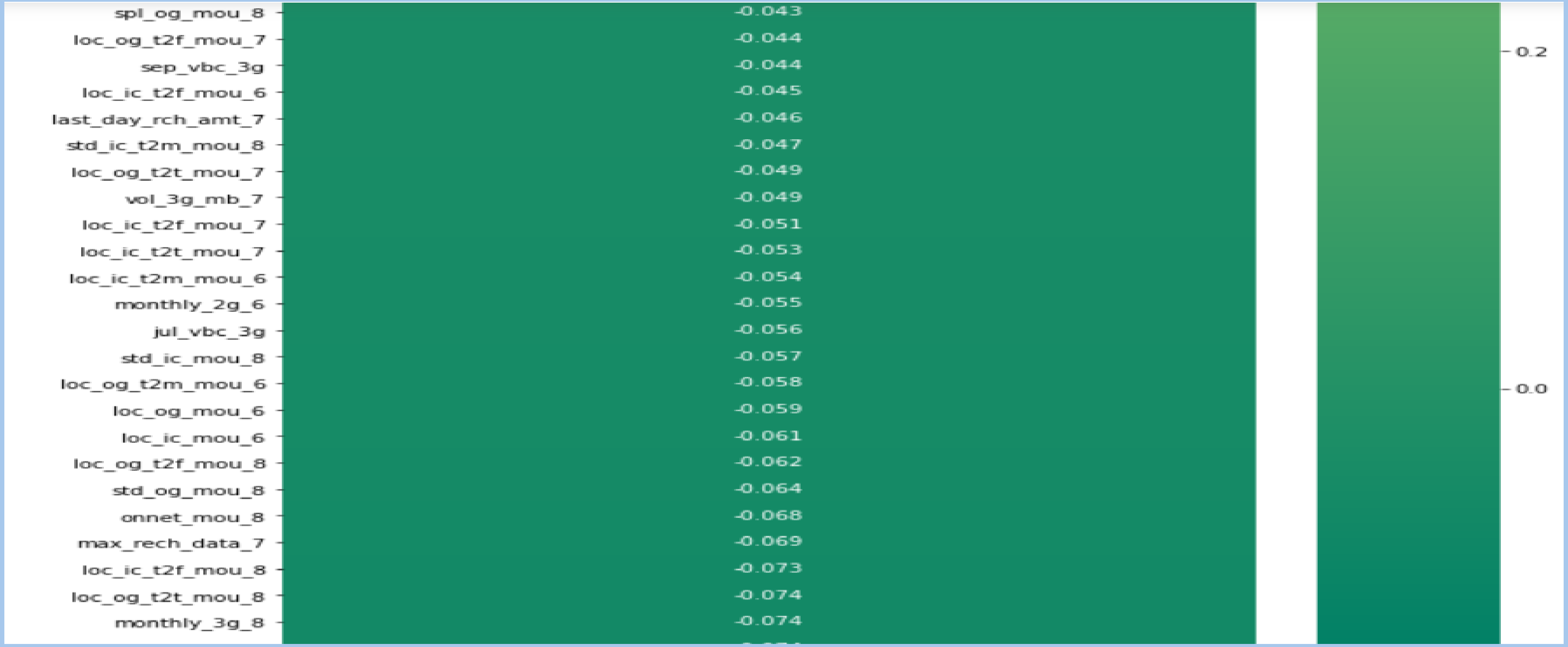


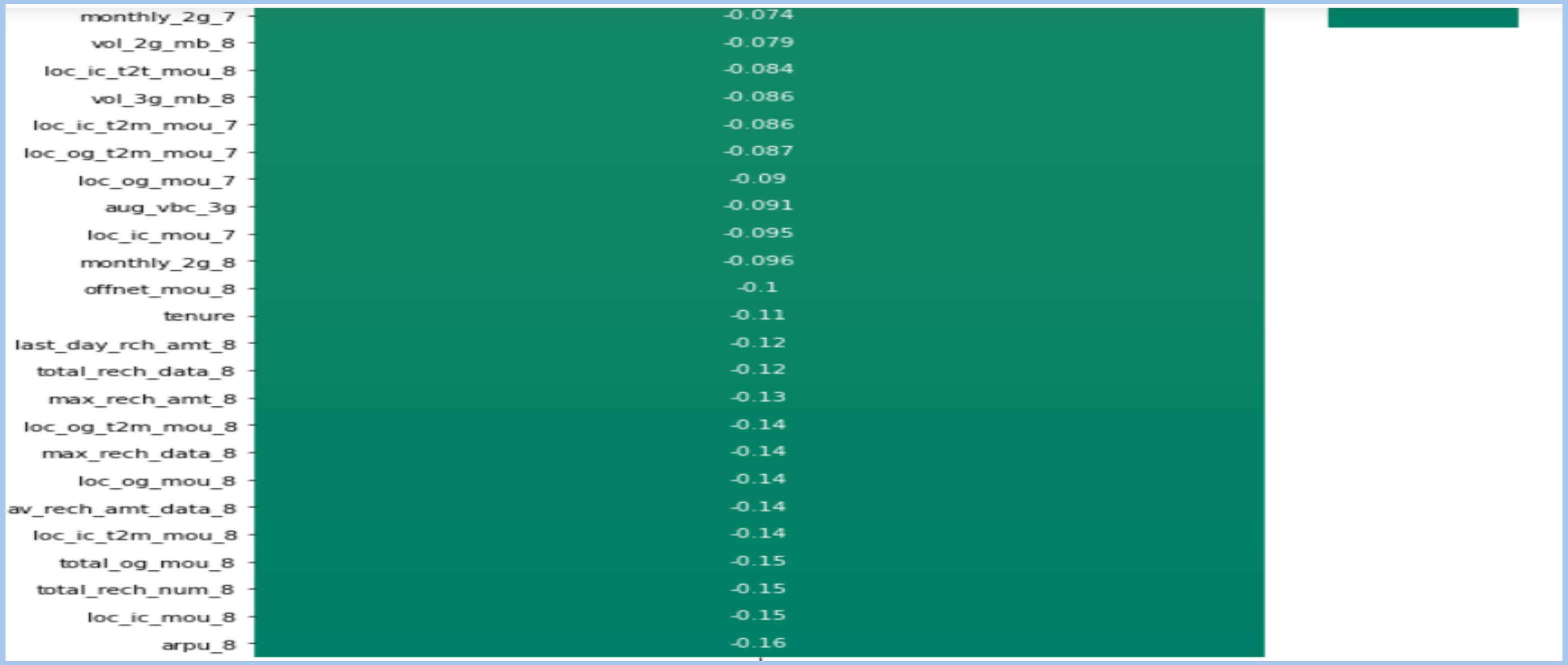
- **It can be seen that the maximum churn rate happens within 0-6 month, but it gradually decreases as the customer retains in the network.**

The average revenue per user is good phase of customer is given by arpu_6 and arpu_7. since we have two seperate averages, lets take an average to these two and drop the other columns.



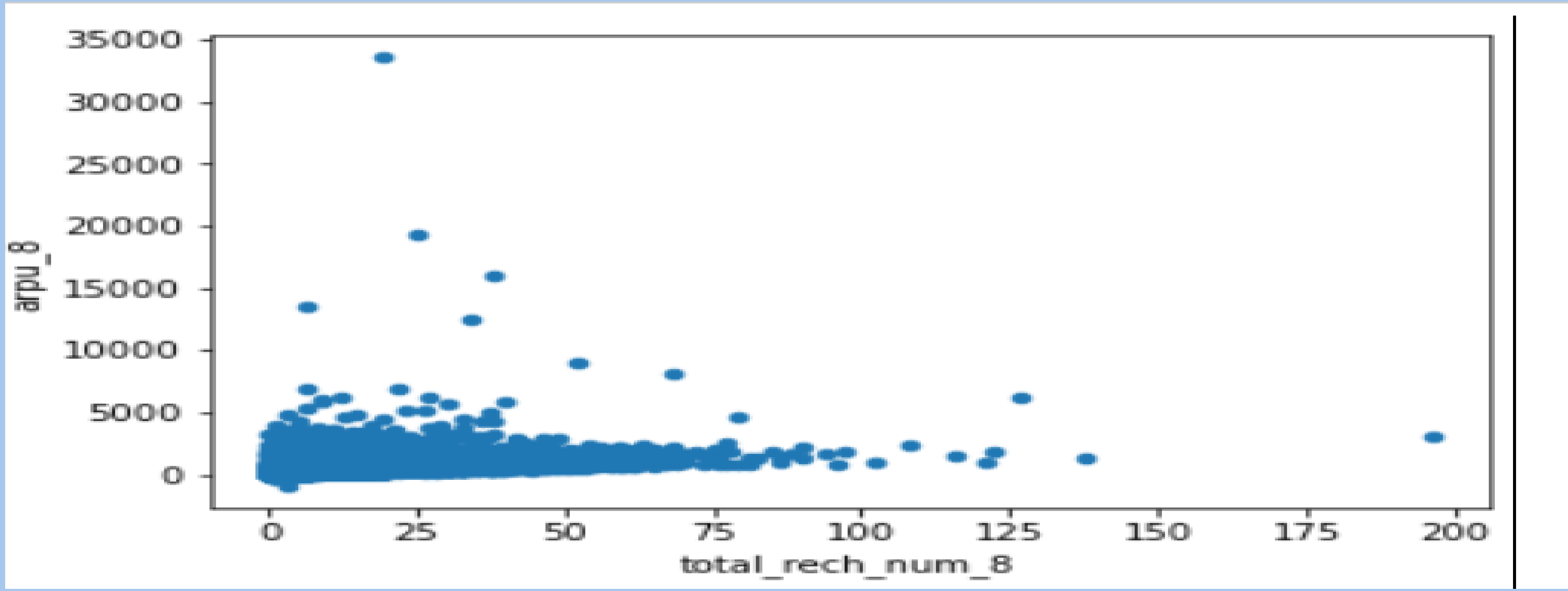

churn 1
std_og_mou_6 0.13
std_og_t2m_mou_6 0.099
roam_og_mou_7 0.099
roam_og_mou_8 0.081
total_og_mou_6 0.078
roam_ic_mou_7 0.074
roam_ic_mou_8 0.074
onnet_mou_6 0.072
roam_og_mou_6 0.069
total_rech_num_6 0.065
offnet_mou_6 0.063
std_og_mou_7 0.057
roam_ic_mou_6 0.056
avg_arpu_6_7 0.029
spl_og_mou_6 0.025
std_ic_t2t_mou_6 0.025
onnet_mou_7 0.018
std_ic_mou_6 0.017
isd_og_mou_6 0.015
overall_rech_amt_6 0.015
og_others_7 0.014
og_others_6 0.011
sachet_3g_6 0.01
1.0
og_others_8 0.0082
offnet_mou_7 0.0078
std_ic_t2m_mou_6 0.0073
last_day_rch_amt_6 0.0072
spl_og_mou_7 0.0055
loc_og_t2c_mou_7 0.0052
isd_ic_mou_6 0.0049
max_rech_amt_6 0.0045
spl_ic_mou_6 0.0024
std_ic_t2t_mou_7 0.0021
loc_og_t2c_mou_6 0.0016
total_og_mou_7 -0.00036
sachet_3g_7 -0.0017
spl_ic_mou_7 -0.0034
isd_ic_mou_7 -0.0041
sachet_2g_6 -0.0043
ic_others_6 -0.0049
std_ic_mou_7 -0.0088
monthly_3g_6 -0.0099
total_rech_num_7 -0.01
std_ic_t2m_mou_7 -0.011
ic_others_7 -0.012
std_og_t2f_mou_6 -0.013
vol_2g_mb_6 -0.013
0.8

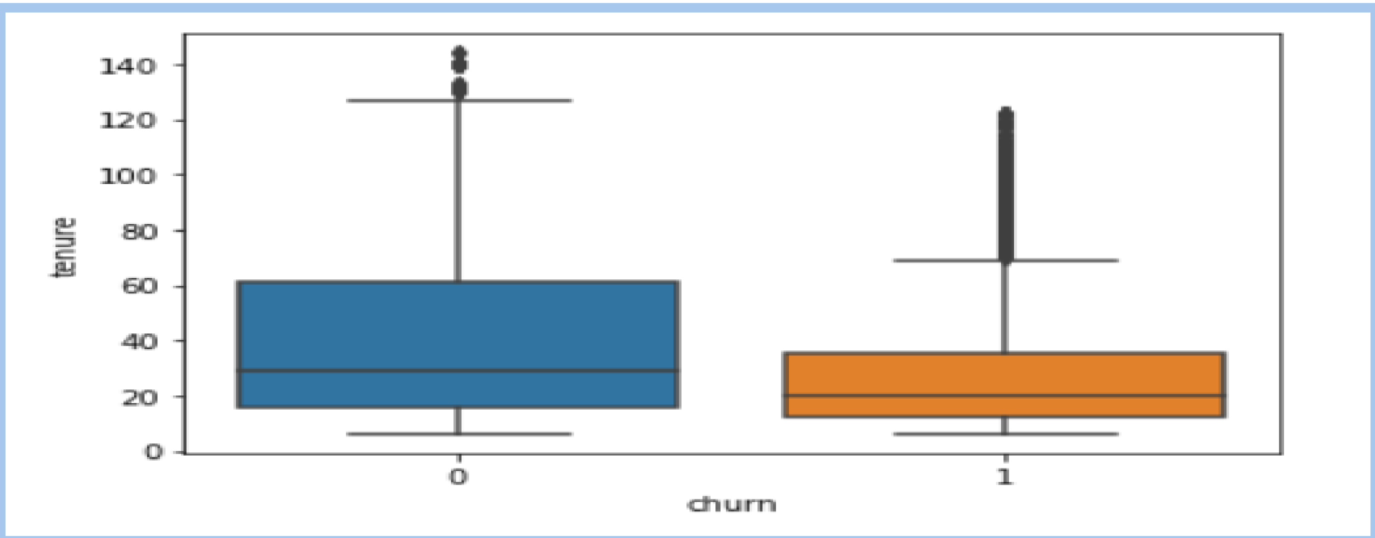

std_og_t2f_mou_7 -0.015
std_ic_t2f_mou_6 -0.015
avg_rech_amt_6_7 -0.016
ic_others_8 -0.017
std_ic_t2f_mou_7 -0.018
vol_3g_mb_6 -0.02
max_rech_amt_7 -0.021
loc_og_t2c_mou_8 -0.024
isd_ic_mou_8 -0.025
std_ic_t2f_mou_8 -0.028
std_og_t2f_mou_8 -0.03
mobile_number -0.03
jun_vbc_3g -0.031
loc_og_t2t_mou_6 -0.031
vol_2g_mb_7 -0.032
loc_ic_t2t_mou_6 -0.032
max_rech_data_6 -0.034
std_ic_t2t_mou_8 -0.035
loc_og_t2f_mou_6 -0.035
sachet_2g_7 -0.037
sachet_3g_8 -0.037
monthly_3g_7 -0.038
overall_rech_amt_7 -0.041
spl_ic_mou_8 -0.043
0.6
0.4

spl_og_mou_8 -0.043
loc_og_t2f_mou_7 -0.044
sep_vbc_3g -0.044
loc_ic_t2f_mou_6 -0.045
last_day_rch_amt_7 -0.046
std_ic_t2m_mou_8 -0.047
loc_og_t2t_mou_7 -0.049
vol_3g_mb_7 -0.049
loc_ic_t2f_mou_7 -0.051
loc_ic_t2t_mou_7 -0.053
loc_ic_t2m_mou_6 -0.054
monthly_2g_6 -0.055
jul_vbc_3g -0.056
std_ic_mou_8 -0.057
loc_og_t2m_mou_6 -0.058
loc_og_mou_6 -0.059
loc_ic_mou_6 -0.061
loc_og_t2f_mou_8 -0.062
std_og_mou_8 -0.064
onnet_mou_8 -0.068
max_rech_data_7 -0.069
loc_ic_t2f_mou_8 -0.073
loc_og_t2t_mou_8 -0.074
monthly_3g_8 -0.074
0.2
0.0



- •Avg Outgoing Calls & calls on romaning for 6 & 7th months are positively correlated with churn.
- •Avg Revenue, No. Of Recharge for 8th month has negative correlation with churn.

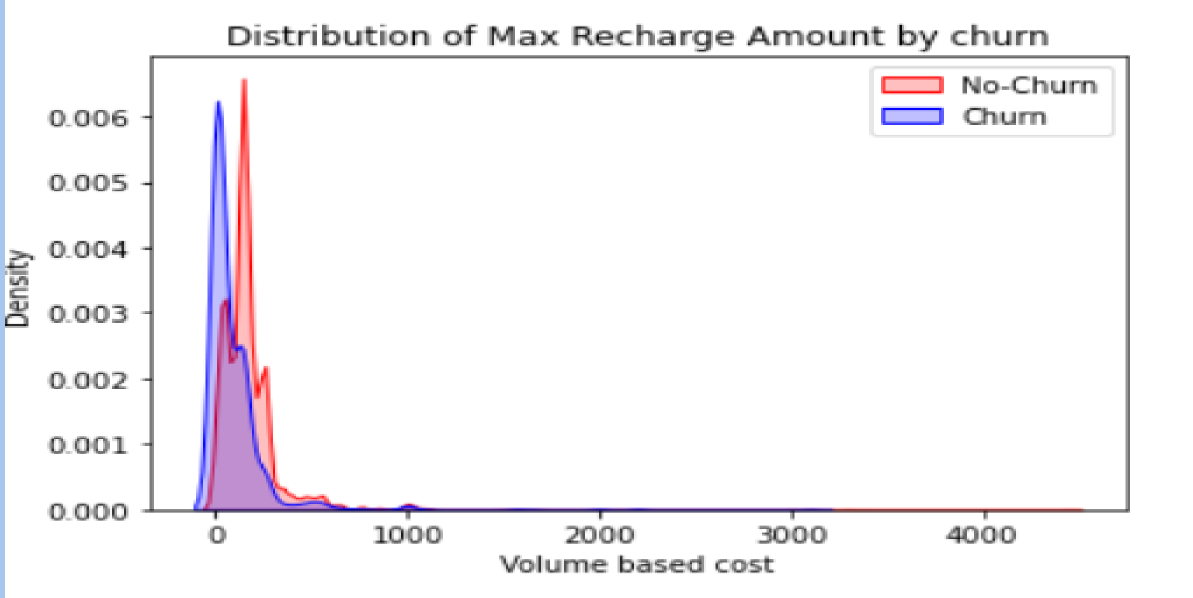
total_rech_num_8', 'arpu_8'

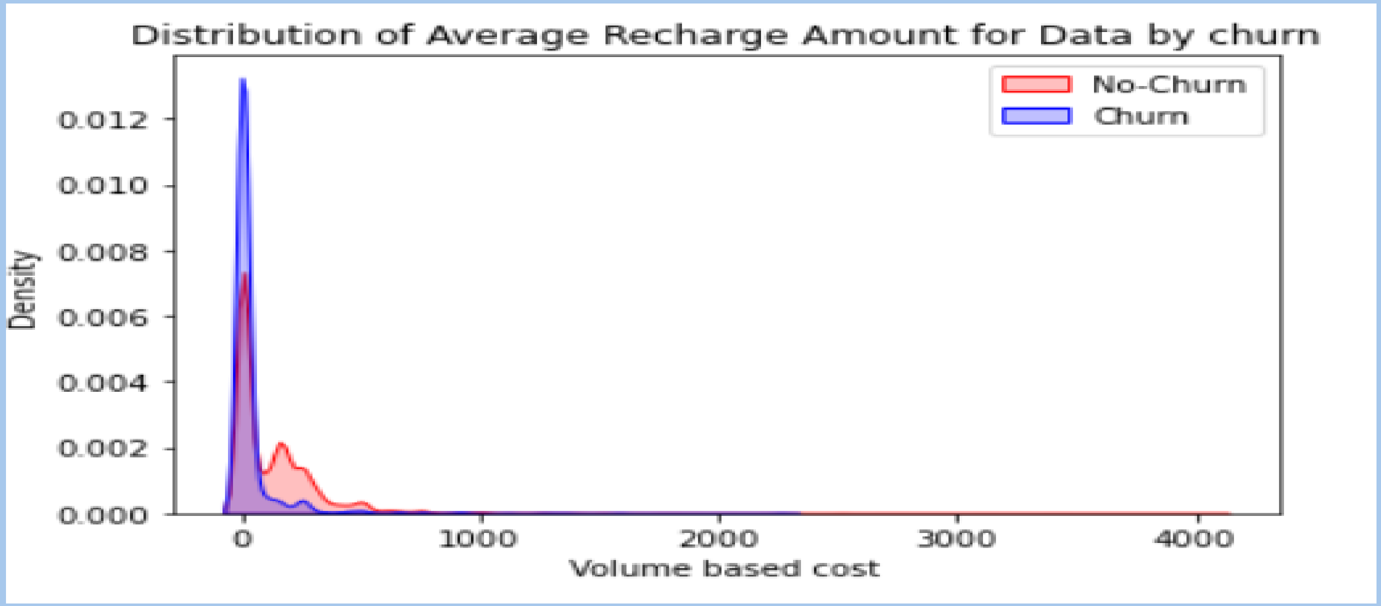




- **From the above plot , its clear tenured customers do no churn and they keep availing telecom services.**

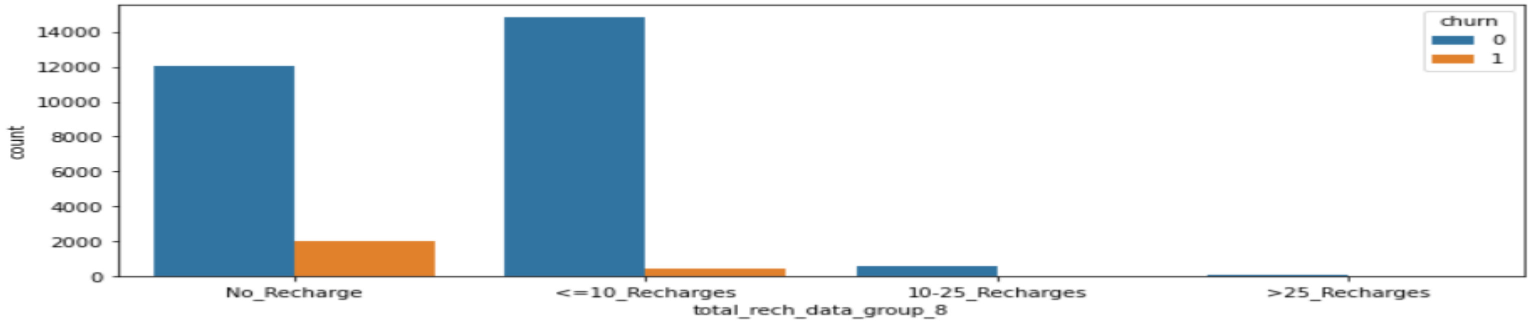
Plot between churn vs max rechare amount.





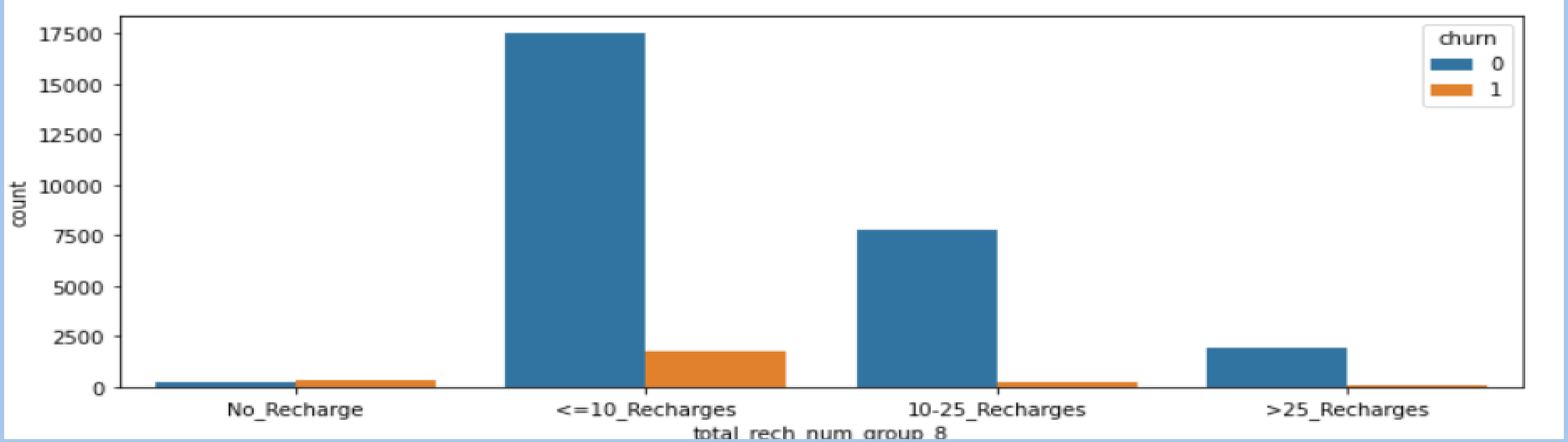
Distribution of total_rech_data_8 variable

```
<=10_Recharges     15307
No_Recharge        14048
10-25_Recharges      608
>25_Recharges         38
Name: total_rech_data_group_8, dtype: int64
```



Distribution of total_rech_num_8 variable

```
<=10_Recharges     19349
10-25_Recharges     8073
>25_Recharges       1996
No_Recharge          583
Name: total_rech_num_group_8, dtype: int64
```



- As the number of recharge rate increases, the churn rate decreases clearly.

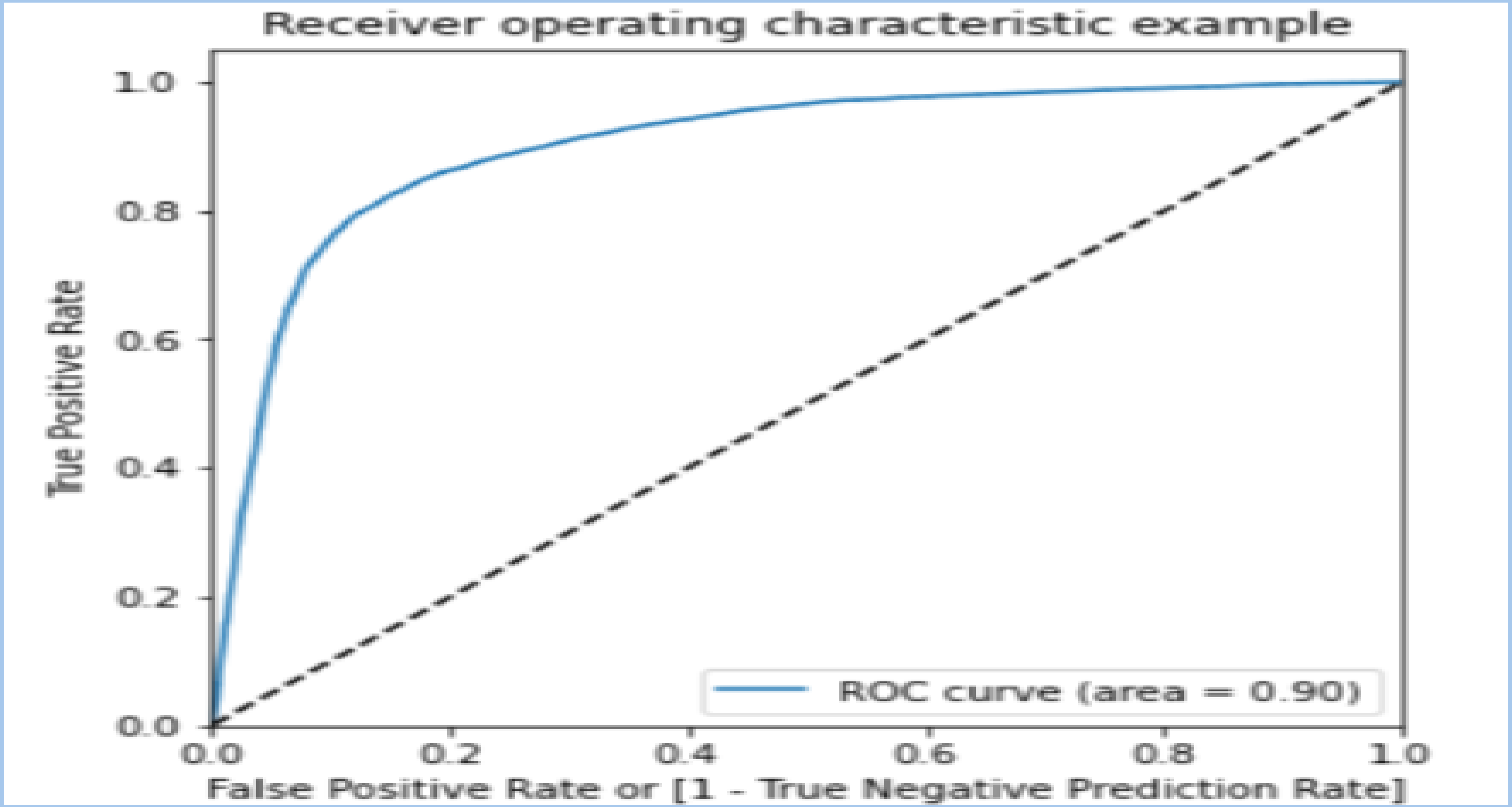
Data Imbalance Handling

- **Using SMOTE method, we can balance the data w.r.t. churn variable and proceed further**

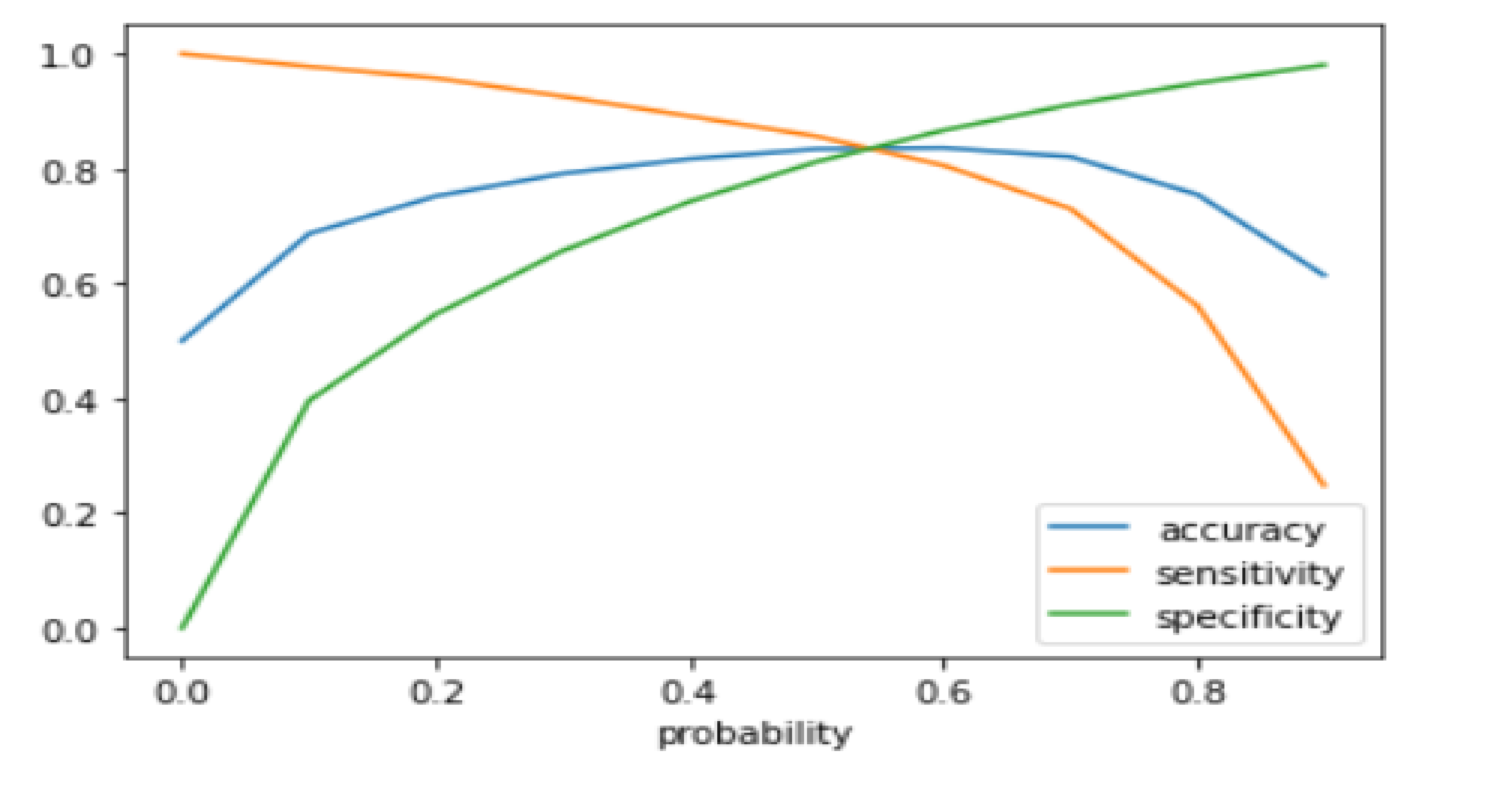
Logistic Regression

- **Logistic Regression using Feature Selection (RFE method).**
- ***Assessing the model with StatsModels***
- ***Creating a dataframe with the actual churn flag and the predicted probabilities***
- **Metrics beyond simply accuracy**

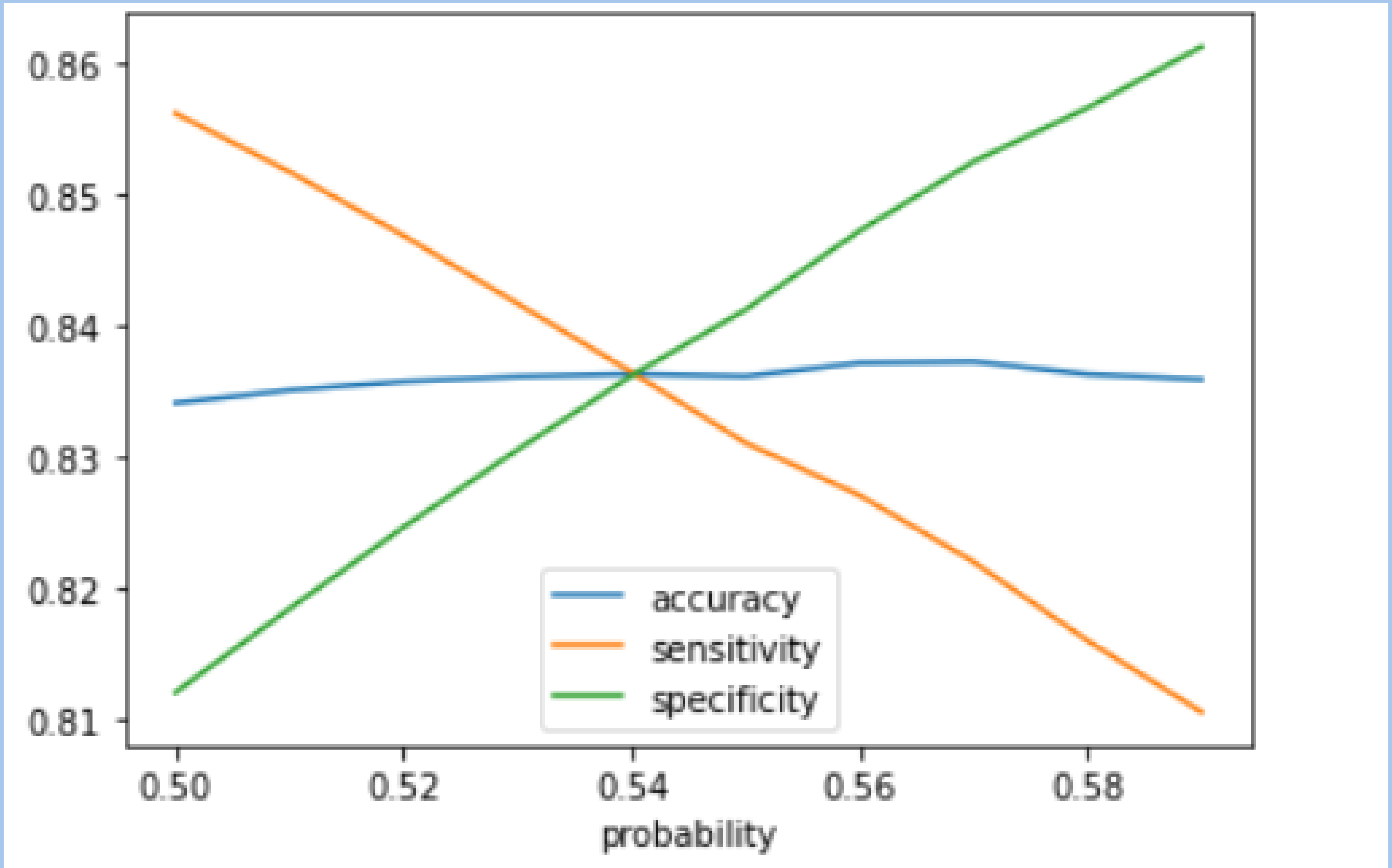
Plotting the ROC Curve



Finding Optimal Cutoff Point

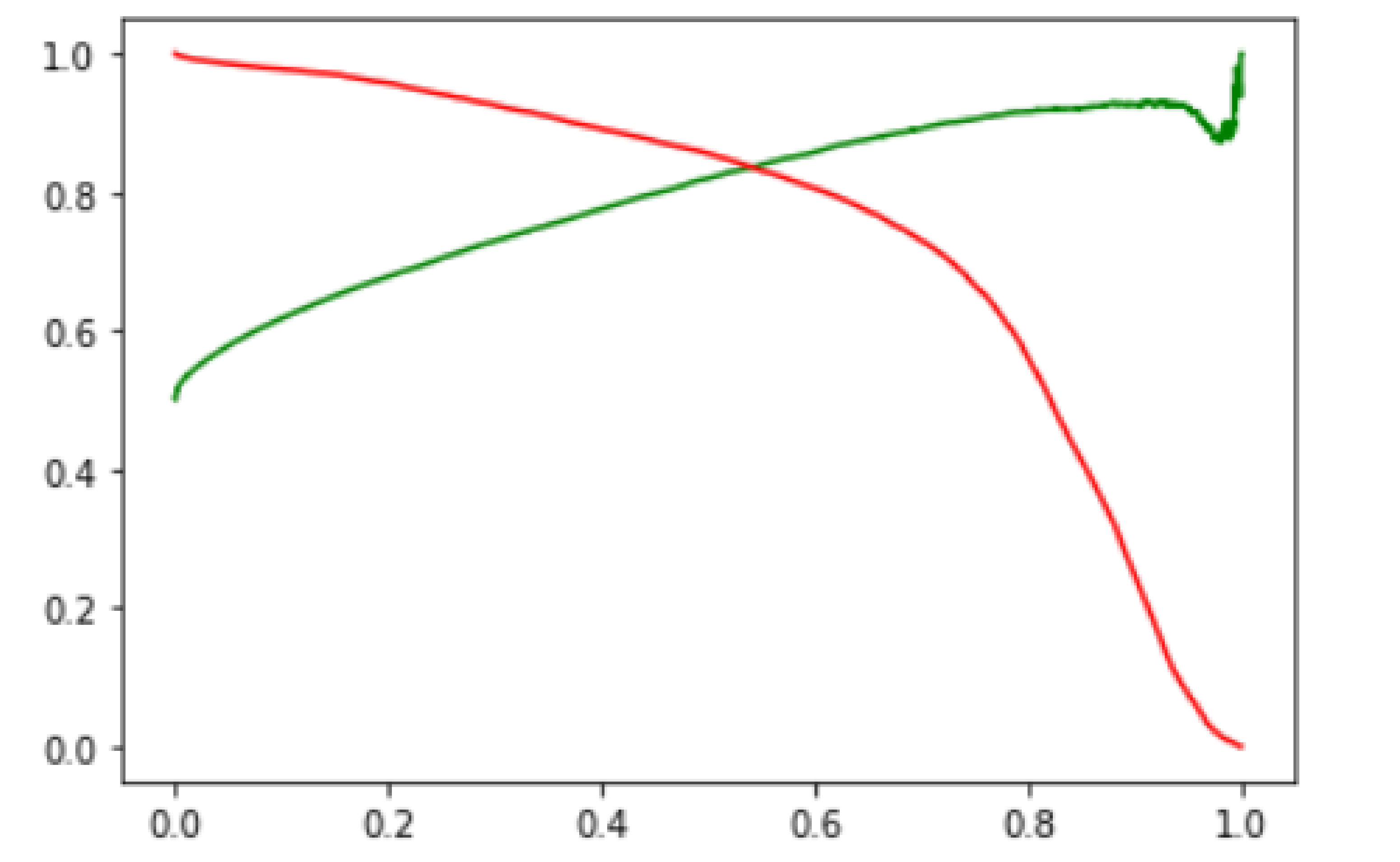


- **Initially we selected the optimm point of classification as 0.5.**

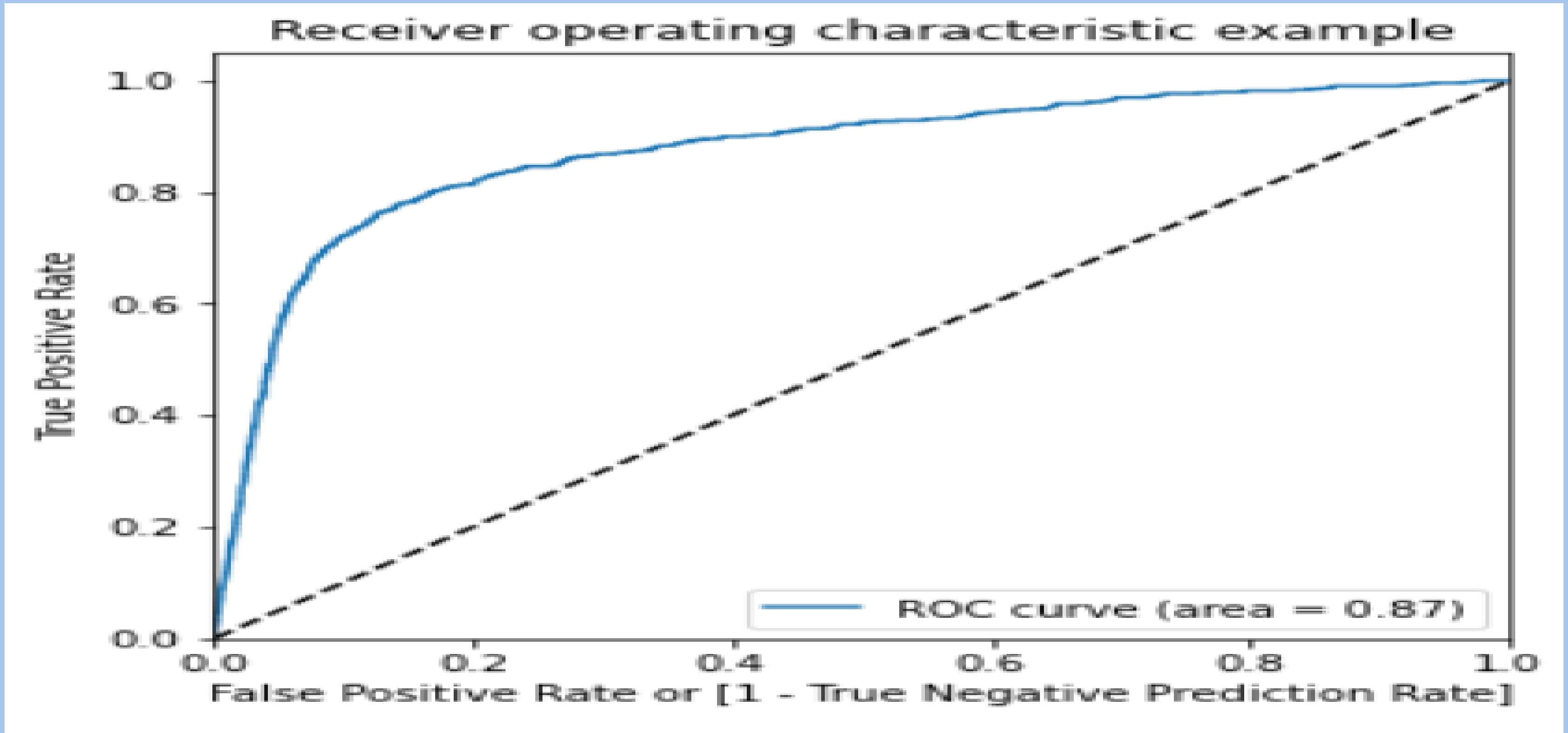


- **From the above graph we can conclude, the optimal cutoff point in the probability to define the predicted churn variabe converges at `0.54`**

Precision and recall tradeoff



Explaining the results



- **The AUC score for train dataset is 0.90 and the test dataset is 0.87.**
- **This model can be considered as a good model.**

Performing Logistic Regression