Telecom Churn Case Study

Submitted by :
Nitin & Mayank

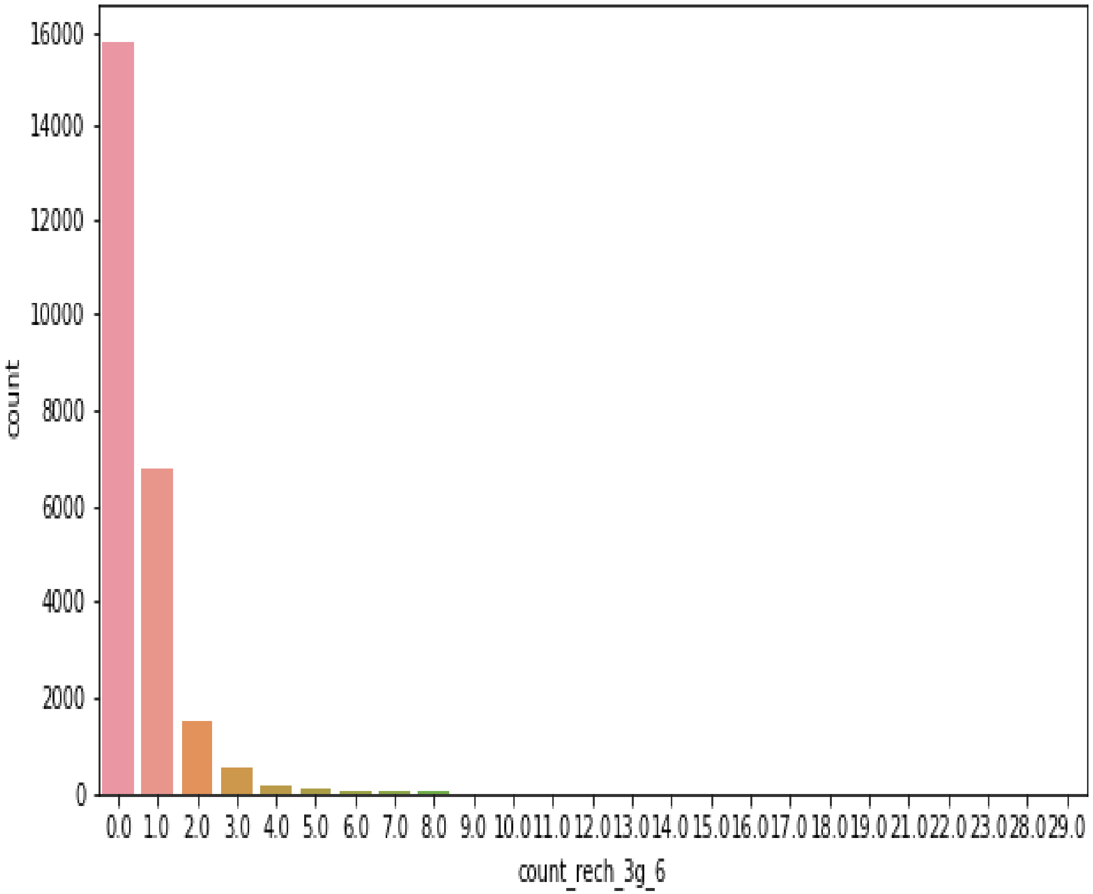
Steps For Solving the Problem :

- Steps to be followed:
- Reading and understanding the data
- Cleaning the data
- Short listing the High-value customers
- Treating the outliers
- Handling imbalance in the data
- Interpretable Model:
  - Model 1 - Logistic Regression
  - Model 2 - Decision Tree
- Inferences, Conclusions and Recommendations

Reading ,understanding & and Cleaning the data:

- Shape of the Dataframe: (99999, 226)
- 214 out of the 226 columns hold numeric information. There is a possibility that these columns use numbers to represent categories.
- Dropping the insignificant columns & One may observe that the top value of null percentages is around 75%. hence it makes sense to handle null values by first checking for columns whose null value percentages are 70< and <80.
- Data has to be inspected month wise, one column at a time.

Cleaning of the data contd. :

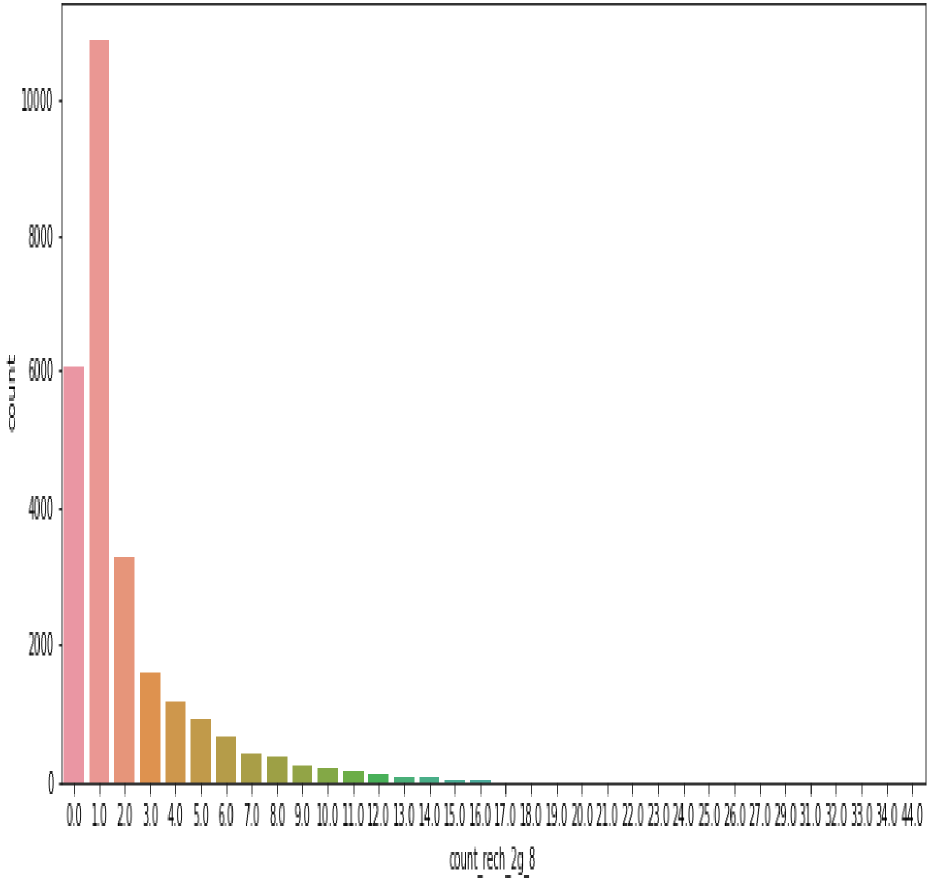


- It can be seen that the volume of 3g pack recharges are not significant in the month of June. It follows that null values in this column could be replaced with 0

Applying the similar procedures in data handling and interpretations for Data:

- Similar procedure applied in the previous slide will be followed for most of the columns .
- Due to the insignificance of the date column in predicting the churn rate, date fields are dropped and stored in temporary variable.
- Following the same rationale as column 'arpu_3g_6', imputing null values with median in this case too.

Data Cleaning contd.



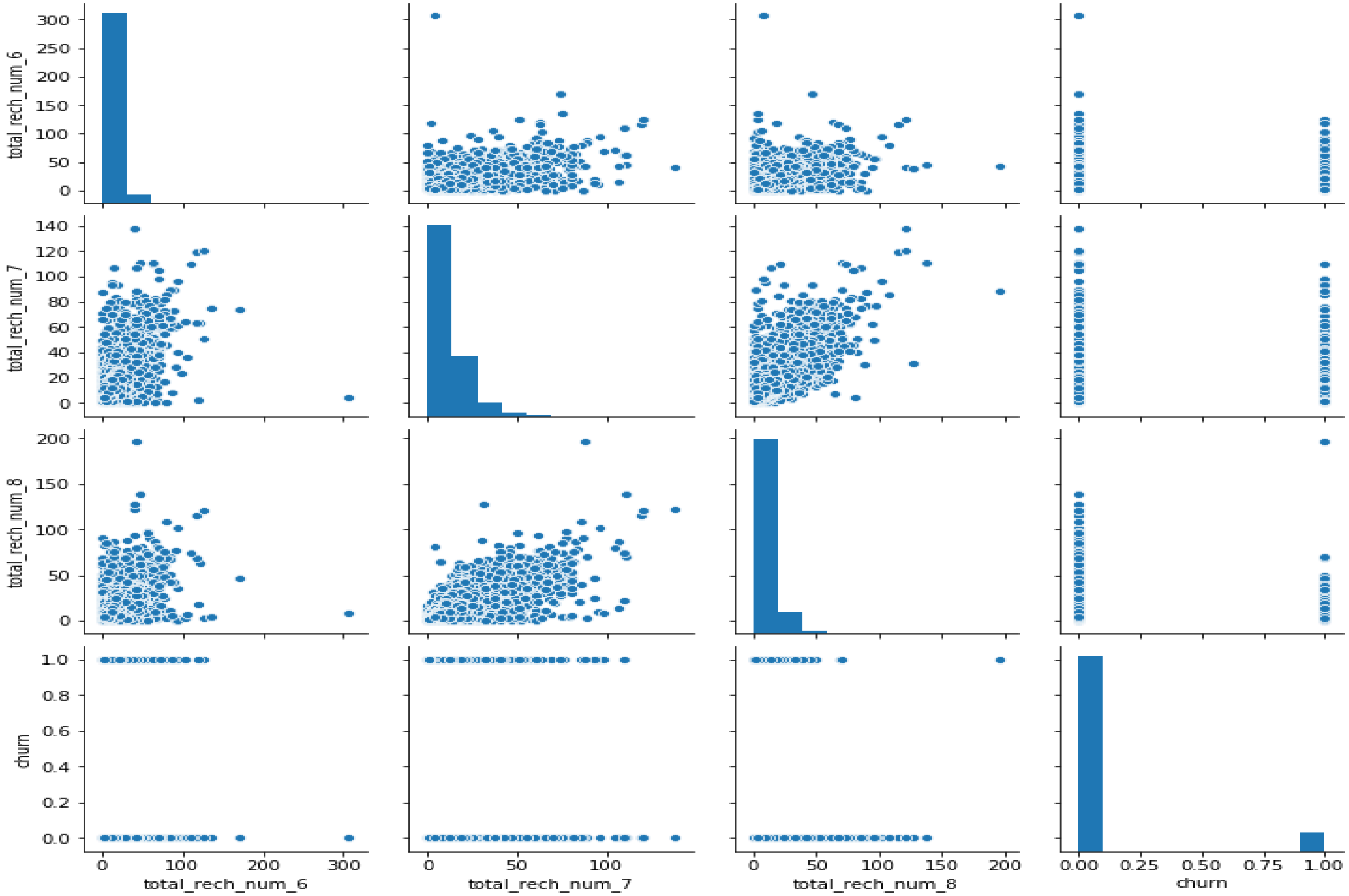
- Just as in June and July, from the graph, it is evident that not a lot of customer did more than 5 recharges for 2g pack in August as well
- A null in this coulmn could mean that no 2g recharge was done by the customer. Therefore, the nulls can be replaced by 0 as it was done for June and July.

Filter high-value customers:

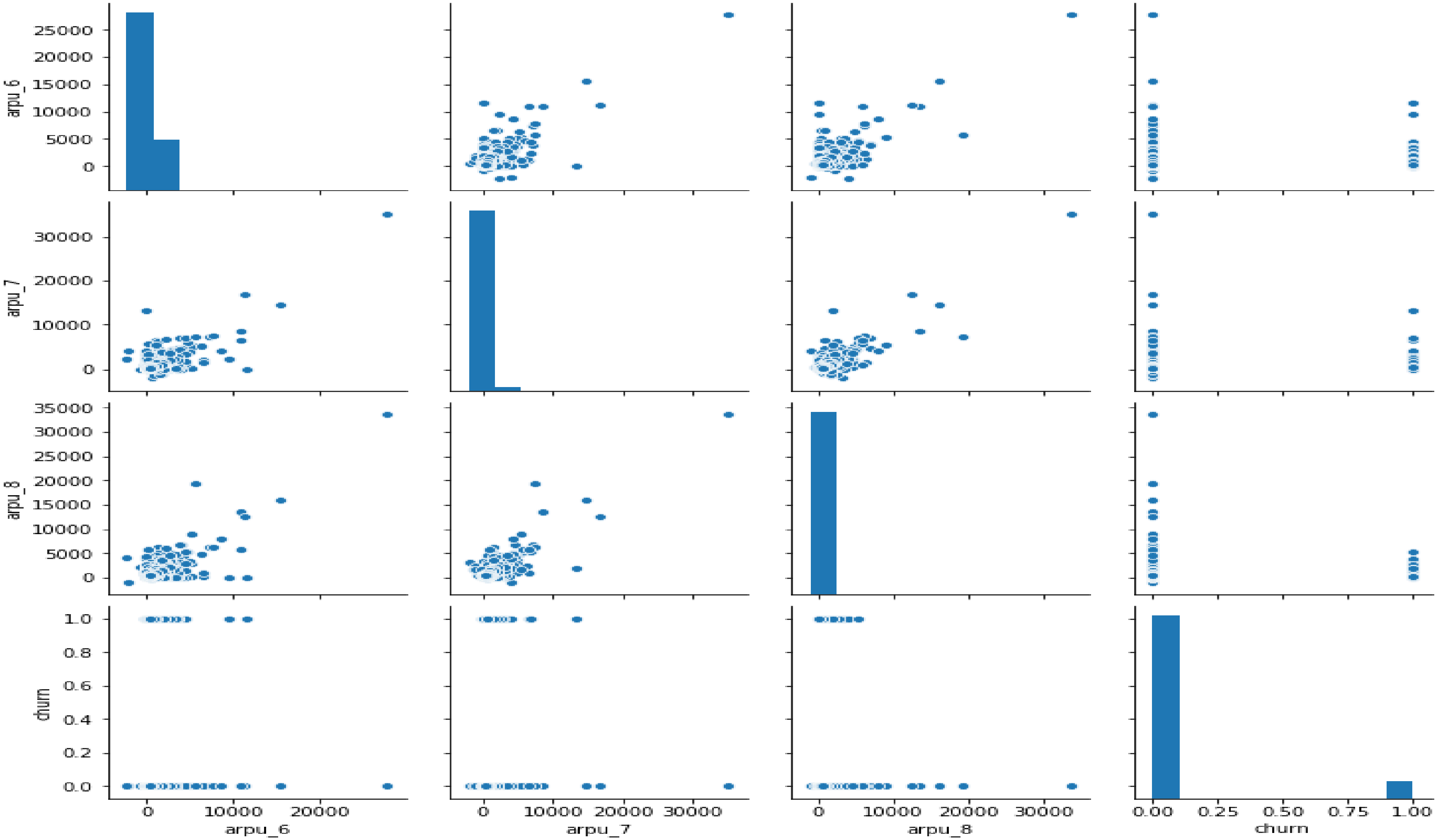
- Problem Statement - "As mentioned above, you need to predict churn only for high-value customers. Define high-value customers as follows: Those who have recharged with an amount more than or equal to X, where X is the 70th percentile of the average recharge amount in the first two months (the good phase).“

- **Tag churners and remove attributes of the churn phase**
- Now tag 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. The attributes you need to use to tag churners are:
- total_ic_mou_9
- total_og_mou_9
- vol_2g_mb_9
- vol_3g_mb_9

BiVariate Analysis on the HVC Data Set- A



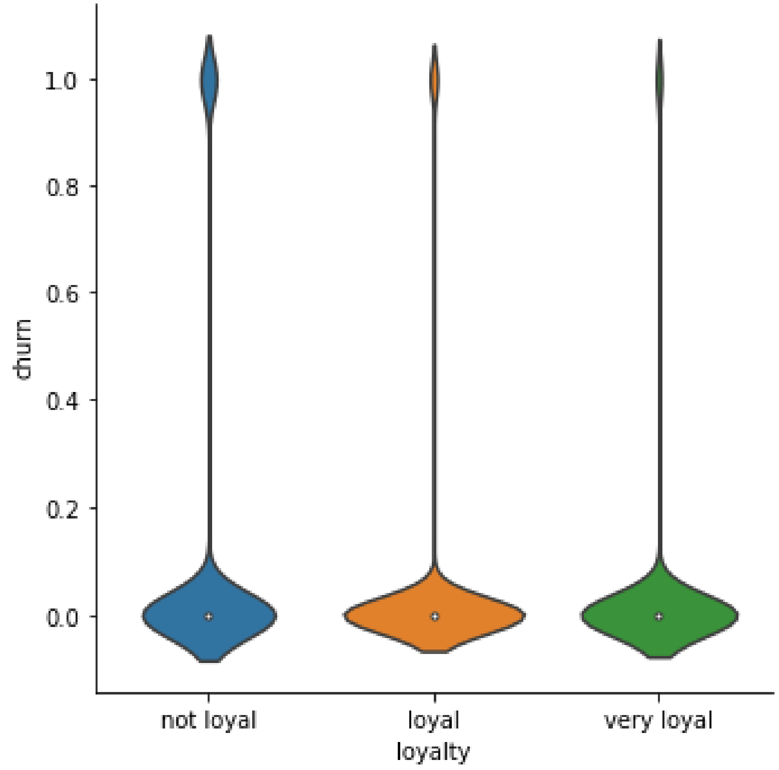
BiVariate Analysis on the HVC Data Set- B



Inference of the Bivariate Analysis:

- Conventional sense says that the more a person recharges, the less likely they are switch service providers. However, the plots above don't seem to give a clear indication of that. It is possibe that other factors too are at play. **(A)**
- One would expect a customer to not churn if the contribute more towards the revenue of the telco. If such is the case, the plots above fail to explain why some customers who do not contribute as much revenue as some other, stick to their current service provider. **(B)**
- One possible explanation is that perhaps they are not heavy users of mobile services and use it very sparingly. Therefore, the good and bad aspects (pros and cons) of the telco are very diminished for these users for whom these services are very utilitarian. **(B)**

Customers Cat plot :



- It can be seen that 'Very Loyal' customers have very less churn and 'Not Loyal' customers churn a lot.

Steps followed:

- **Treating Outliers**
- **Handling Imbalance in Data**

- Because of the 8.64 percent churn in the dataset, the results from this would be highly skewed owing to its imbalance.

- Such class imbalance can be handled by creating synthetic records of the minor class for churn = 1.

- **Scaling the data :** Standard Scalar

Logistic Regression:

- For evaluating the classification we'll be using the logistic regression model 1st.
- RFE : Using the RFE as assessment gives, us the most important features which can then be checked further by building a StatsModels model using the same.
- In the output after 2nd and 3rd iterations, a lot of p-values are greater than 0.05. Deleting the feature with the largest p-value and rebuilding the model, repeatedly until all the coefficients are significant.

Predicting the train and test results:

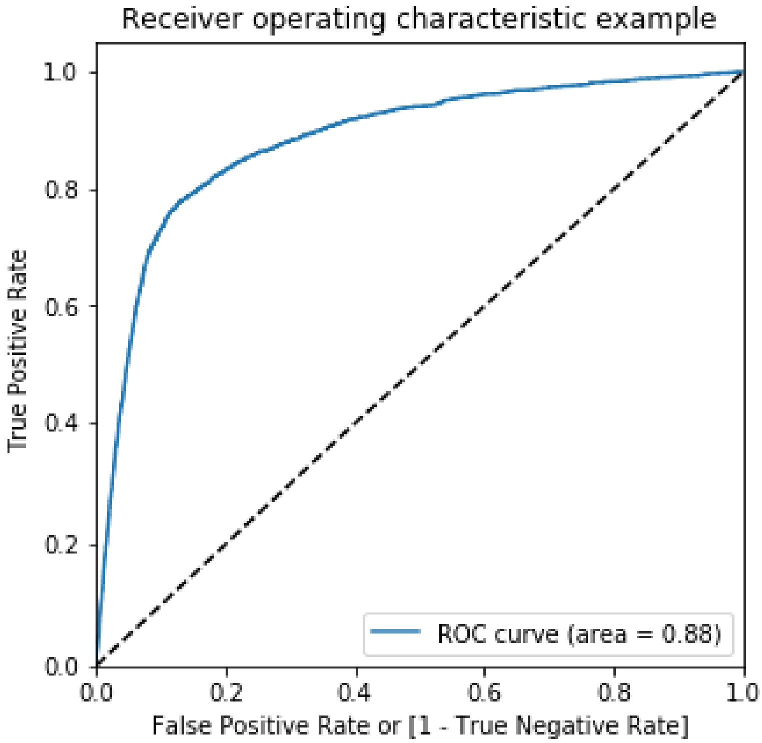
- Calculating all the metrics :

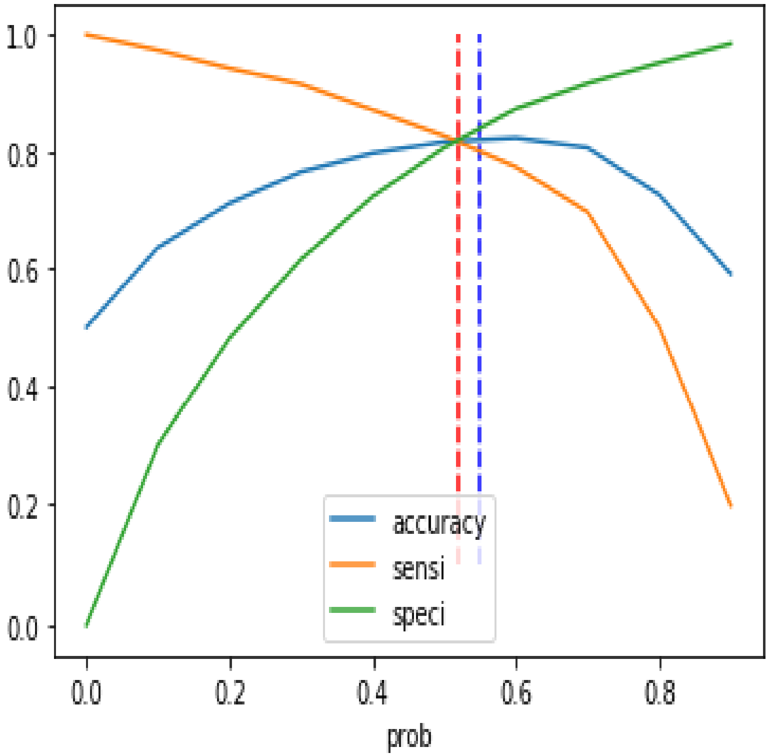
1. Confusion Matrix :

```
[ [17711   4240  ]
  [  3799    18152] ]
```

2. Sensitivity/Recall/True Positive Rate for the confusion matrix = 82.69%
Accuracy Score = 81.69% Specificity for the confusion matrix = 80.68% Precision for the confusion matrix = 81.06% False Positive Rate for the confusion matrix = 19.32%

Plotting the ROC Curve and cut off:





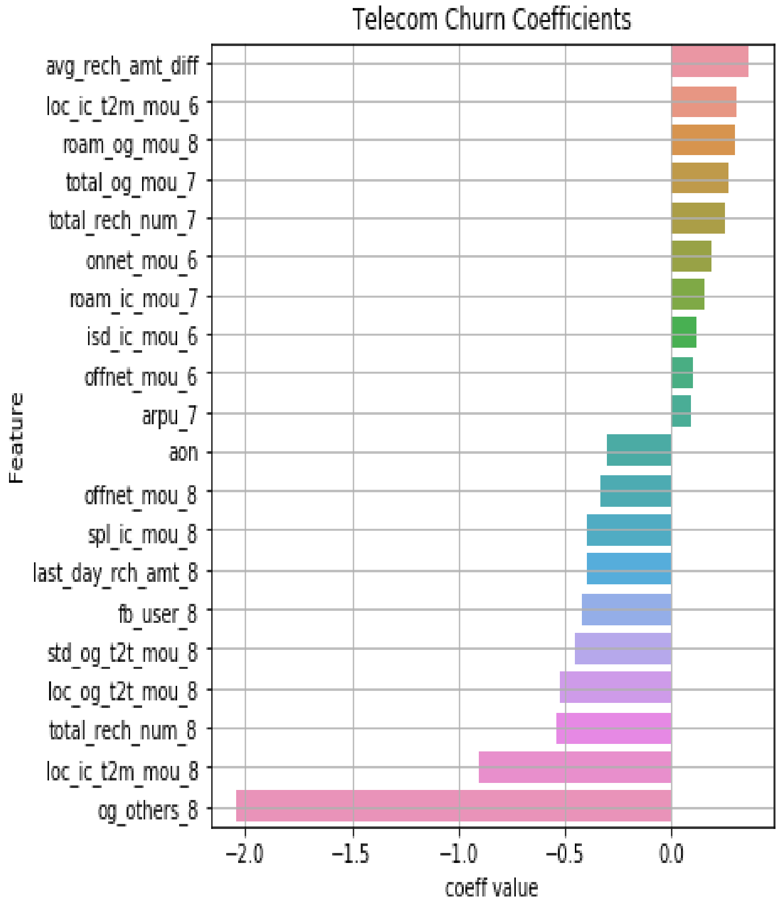
- Despite 0.52 being the optimum point to take as cutoff probability, threshold = 0.45 gives us the best (most acceptable) combination for accuracy and sensitivity.

Calculation of All metrics (3rd Model)

- Confusion matrix:
- [[4176 1263]
- [75 482]]

- Sensitivity/Recall/True Positive Rate for the confusion matrix = 86.54%
- Accuracy Score = 77.69%
- Specificity for the confusion matrix = 76.78%
- Precision for the confusion matrix = 27.62%
- False Positive Rate for the confusion matrix = 23.22%

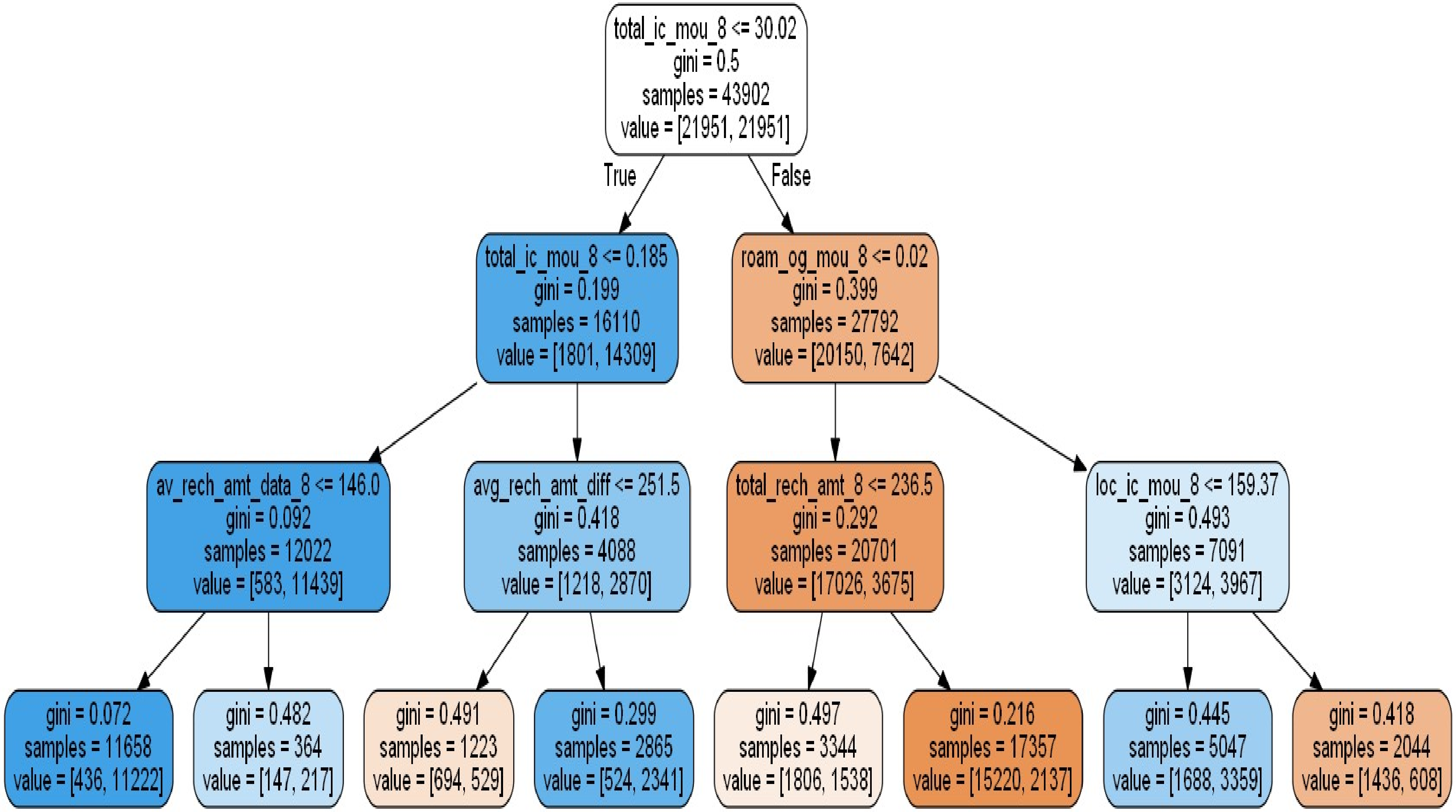
Telecom Churn Coefficients:



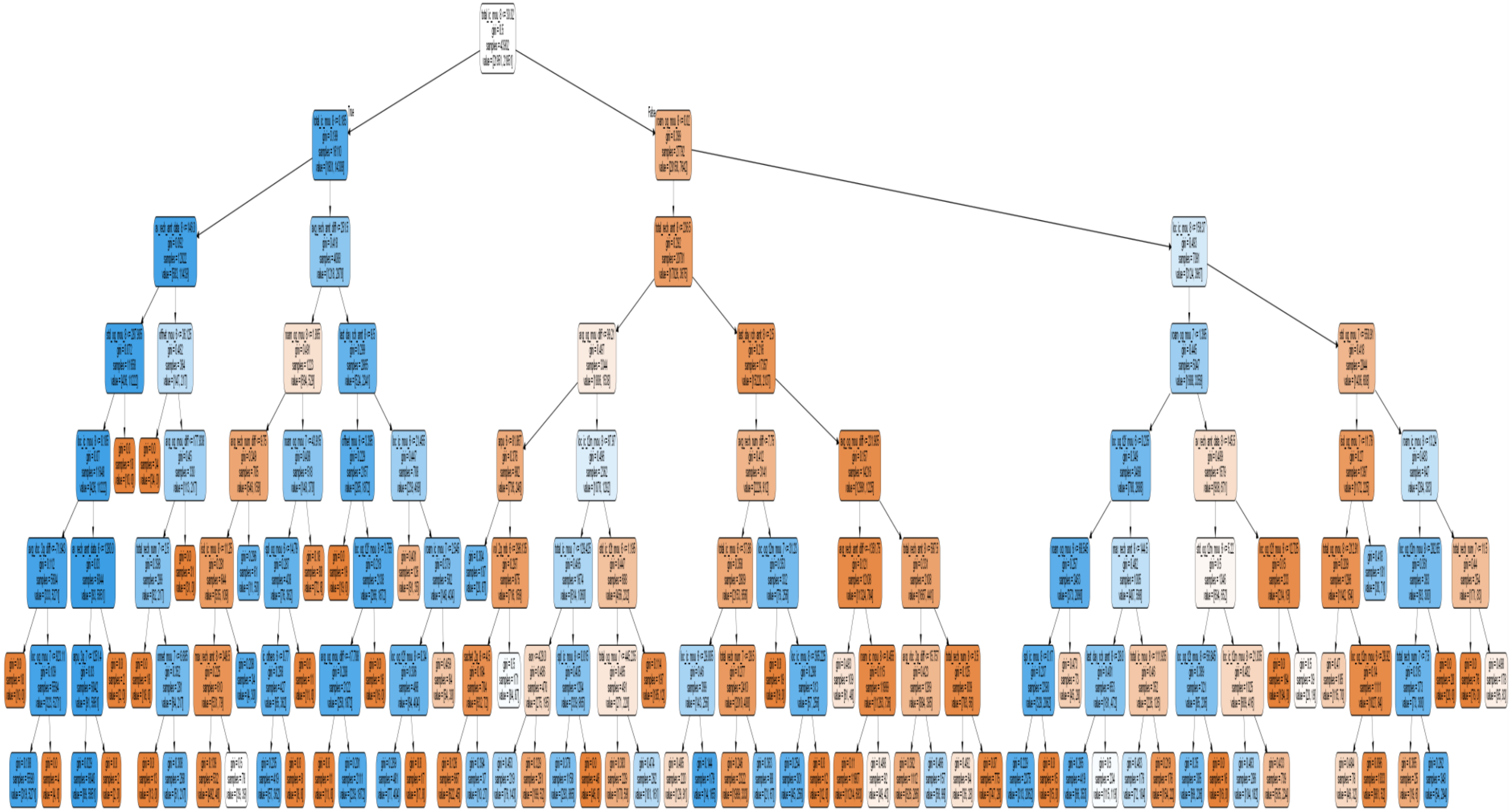
From the above plot, the most impactful (top 10) factors contributing to churn behaviour are

- og_others_8
- loc_ic_t2m_mou_8
- total_rech_num_8
- loc_og_t2t_mou_8
- std_og_t2t_mou_8
- fb_user_8
- last_day_rch_amt_8
- spl_ic_mou_8
- avg_rech_amt_diff
- offnet_mou_8

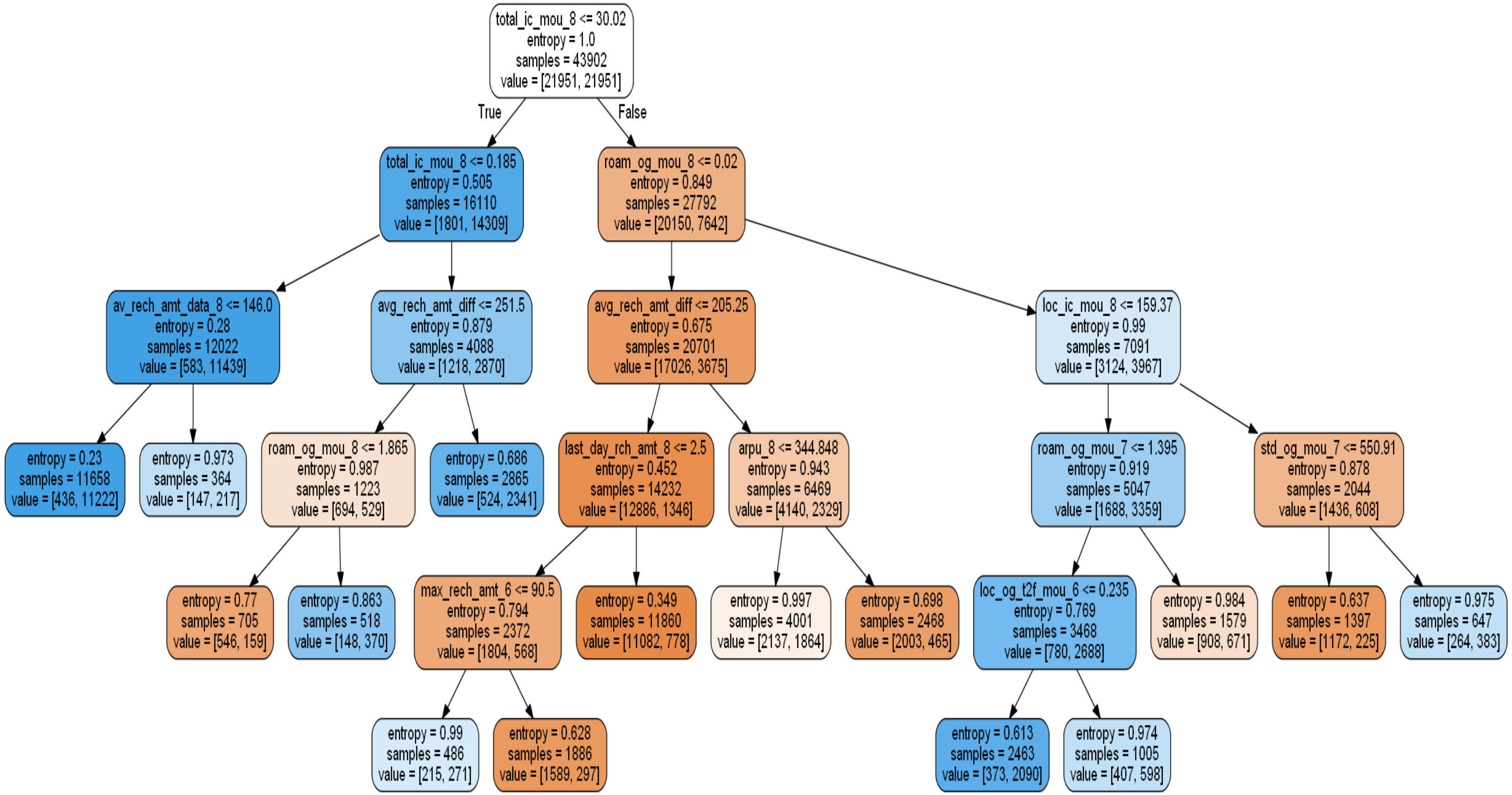
Decision Tree : Model 1



Decision Tree Model 2



Decision Tree Model 3:



total_ic_mou_8 <= 30.02
entropy = 1.0
samples = 43902
value = [21951, 21951]

True

False

total_ic_mou_8 <= 0.185
entropy = 0.505
samples = 16110
value = [1801, 14309]

roam_og_mou_8 <= 0.02
entropy = 0.849
samples = 27792
value = [20150, 7642]

av_rech_amt_data_8 <= 146.0
entropy = 0.28
samples = 12022
value = [583, 11439]

avg_rech_amt_diff <= 251.5
entropy = 0.879
samples = 4088
value = [1218, 2870]

avg_rech_amt_diff <= 205.25
entropy = 0.675
samples = 20701
value = [17026, 3675]

loc_ic_mou_8 <= 159.37
entropy = 0.99
samples = 7091
value = [3124, 3967]

entropy = 0.23
samples = 11658
value = [436, 11222]

entropy = 0.973
samples = 364
value = [147, 217]

roam_og_mou_8 <= 1.865
entropy = 0.987
samples = 1223
value = [694, 529]

entropy = 0.686
samples = 2865
value = [524, 2341]

last_day_rch_amt_8 <= 2.5
entropy = 0.452
samples = 14232
value = [12886, 1346]

arpu_8 <= 344.848
entropy = 0.943
samples = 6469
value = [4140, 2329]

roam_og_mou_7 <= 1.395
entropy = 0.919
samples = 5047
value = [1688, 3359]

std_og_mou_7 <= 550.91
entropy = 0.878
samples = 2044
value = [1436, 608]

entropy = 0.77
samples = 705
value = [546, 159]

entropy = 0.863
samples = 518
value = [148, 370]

max_rech_amt_6 <= 90.5
entropy = 0.794
samples = 2372
value = [1804, 568]

entropy = 0.349
samples = 11860
value = [11082, 778]

entropy = 0.997
samples = 4001
value = [2137, 1864]

entropy = 0.698
samples = 2468
value = [2003, 465]

loc_og_t2f_mou_6 <= 0.235
entropy = 0.769
samples = 3468
value = [780, 2688]

entropy = 0.984
samples = 1579
value = [908, 671]

entropy = 0.637
samples = 1397
value = [1172, 225]

entropy = 0.975
samples = 647
value = [264, 383]

entropy = 0.99
samples = 486
value = [215, 271]

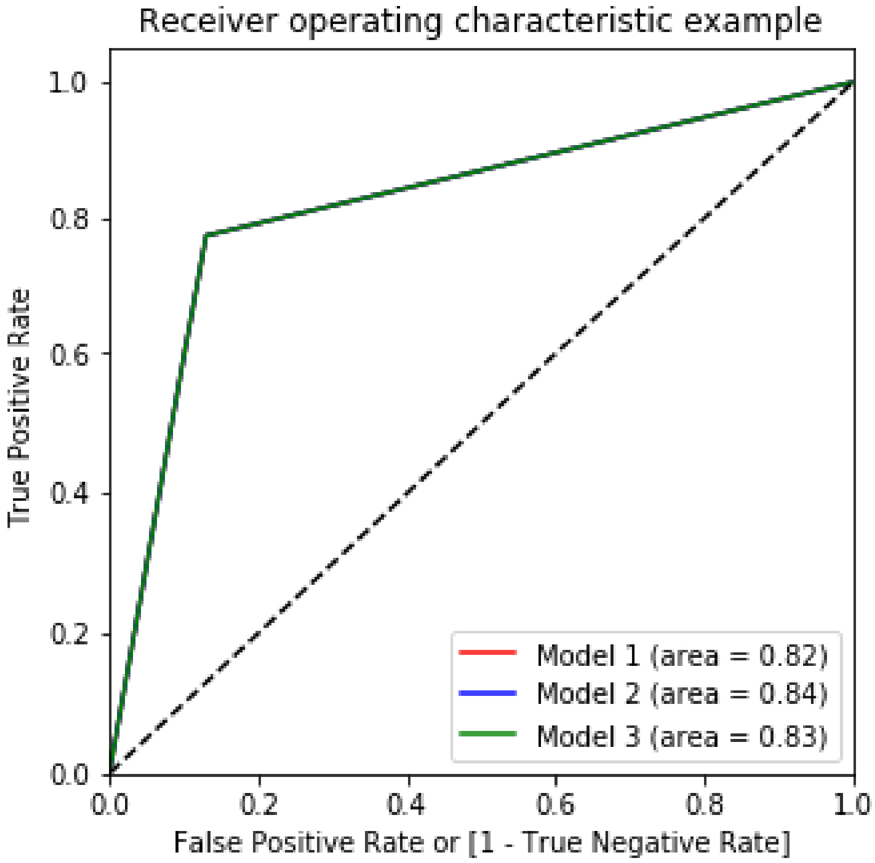
entropy = 0.628
samples = 1886
value = [1589, 297]

entropy = 0.613
samples = 2463
value = [373, 2090]

entropy = 0.974
samples = 1005
value = [407, 598]

ROC Curve:

Model 2 (largest area) gives the best results in accordance with our inference from the confusion matrix.



Final Inferences and suggestions:

- 2nd Decision tree Model is showing the most accuracy.
- If the total recharge made by the customer in August is more, then he is likely to retain with the network. Offers can be made to the customers to hold them back, unlimited data.
- Lesser the amount of recharge in the end of the month obviously there are more chances of the customer churning. Suggestion for discounts and coupons to reduce the amount of the recharge to be borne by the customer; Value added packs, ex. free night calling .
- As the number of outgoing calls (both local and std) have increased from the good phase to action phase, it is observed that the likelihood of the customer churning is high. Adding more minutes to the customer who are proned to the network and giving benefits for staying loyal.

- **Conclusion 2nd Decision Tree giving the most accuracy.**
- Confusion matrix:

[[4792 647]

- [111 446]] Sensitivity/Recall/True Positive Rate for the confusion matrix = 80.07% Accuracy Score = 87.36% Specificity for the confusion matrix = 88.1% Precision for the confusion matrix = 40.81% False Positive Rate for the confusion matrix = 11.9%
- Inference - Gives balanced values for all metrics.