TELECOM CHURN PREDICTION

CAPSTONE PROJECT

Anjali KN

INTRODUCTION

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.

To reduce customer churn, telecom companies need to **predict which highly profitable customers are at risk of churn.**

Thus, churn prediction is usually more critical (and non-trivial) for prepaid customers, and the term ‘churn’ should be defined carefully.

BUSINESS OBJECTIVE

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 behaviour during churn will be helpful.

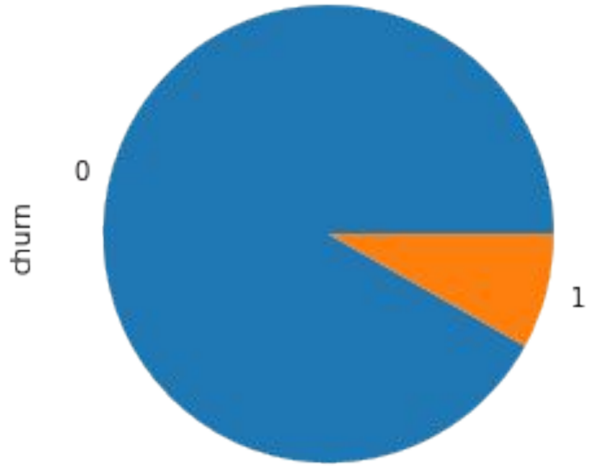
In churn prediction, we assume that there are **three phases** of the customer lifecycle :

1. The ‘good’ phase
2. The ‘action’ phase
3. The ‘churn’ phase

Steps

- ➔ Preprocess data
- ➔ Conduct appropriate exploratory analysis to extract useful insights
- ➔ Train a model, tune model hyperparameters, etc.
- ➔ Evaluate the models using appropriate evaluation metrics.
- ➔ choose a model based on an evaluation metric with proper justification.

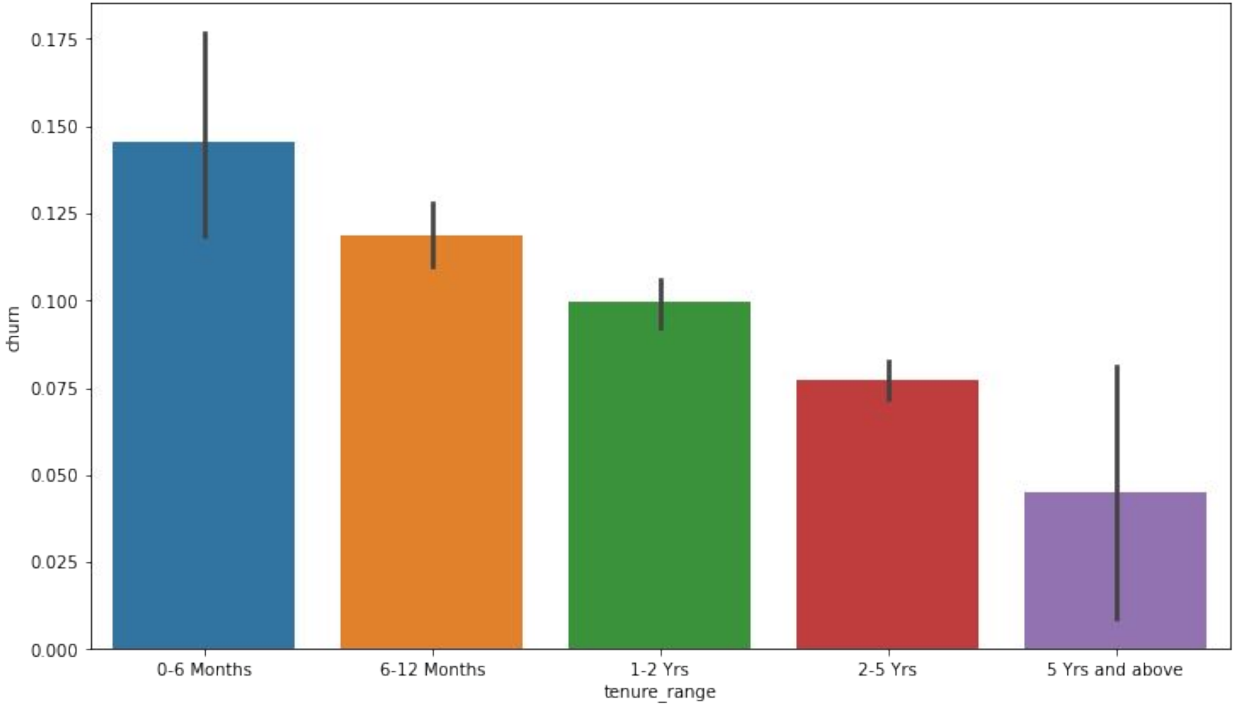
Lets find out churn/non churn percentage



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

Since this variable `churn` is the target variable, all the columns relating to this variable(i.e. all columns with suffix `_9`) can be dropped forn the dataset.

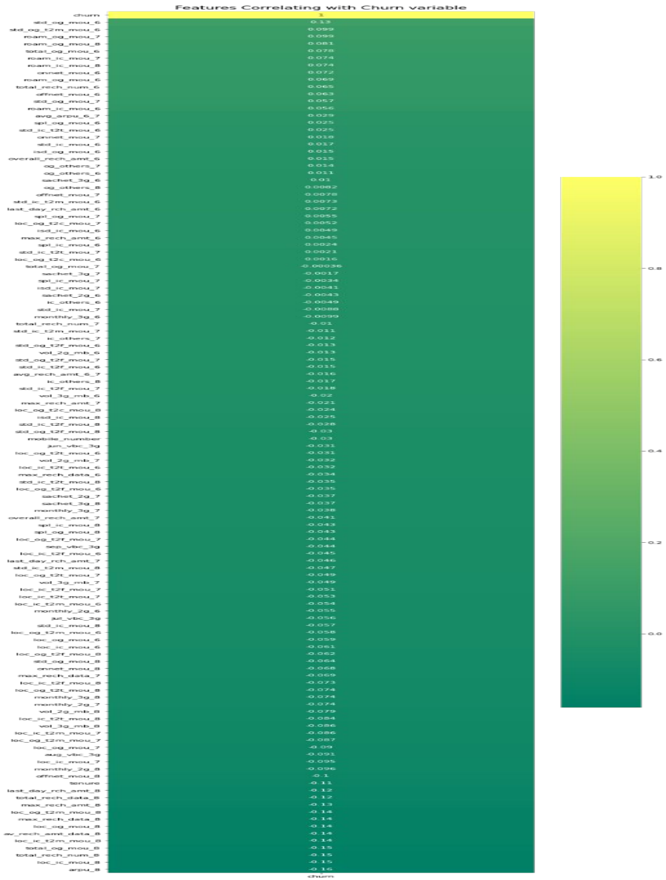
Plotting a bar plot for tenure range



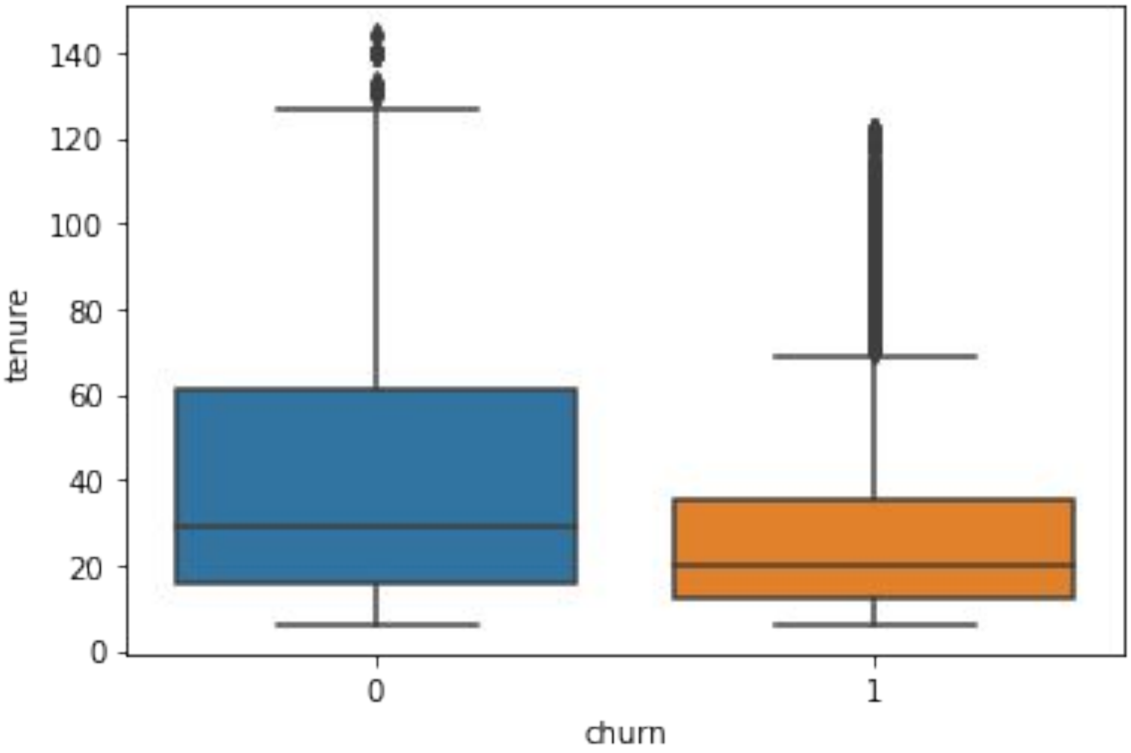
It can be seen that the maximum churn rate happens within 0-6 month, but it gradually decreases as the customer retains in the network.

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.

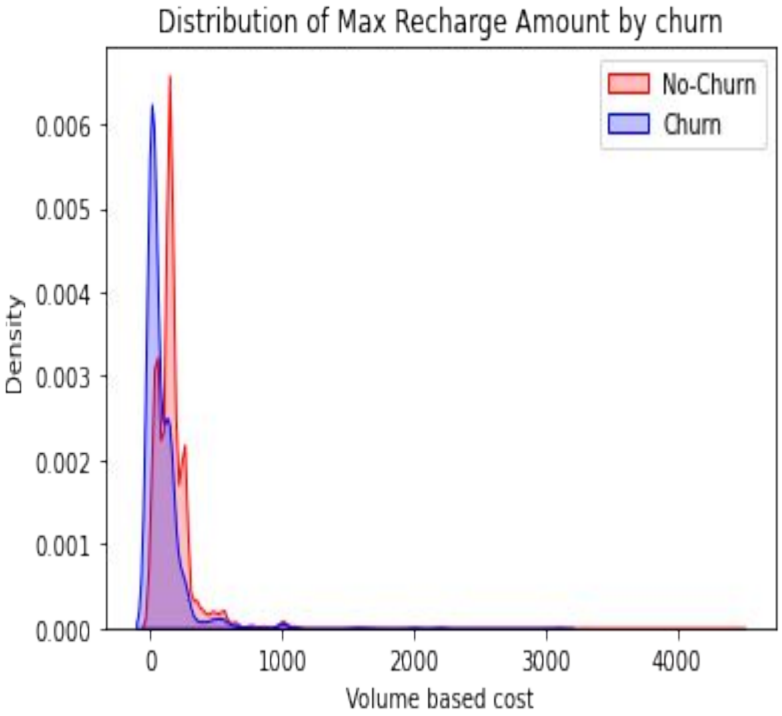
Checking Correlation between target variable(SalePrice) with the other variable in the dataset

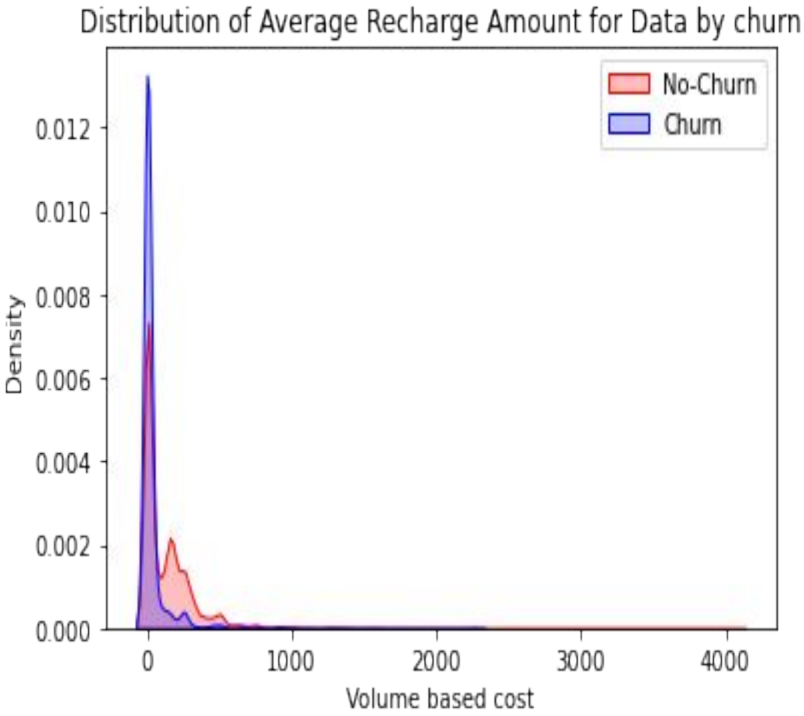


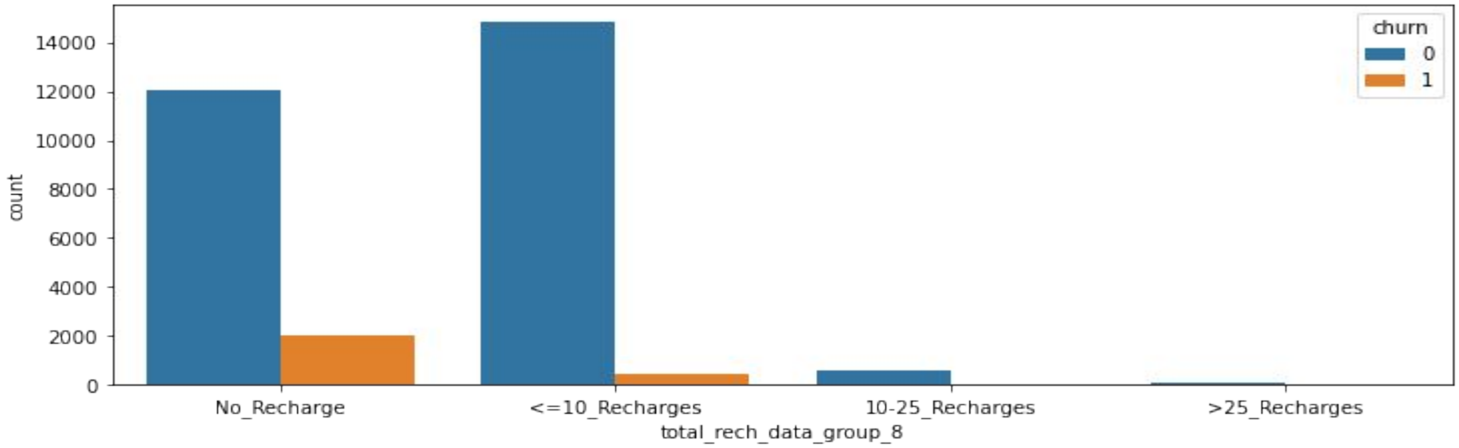
- Avg Outgoing Calls & calls on romaning for 6 & 7th months are positively correlated with churn.
- Avg Revenue, No. Of Recharge for 8th month has negative correlation with churn.

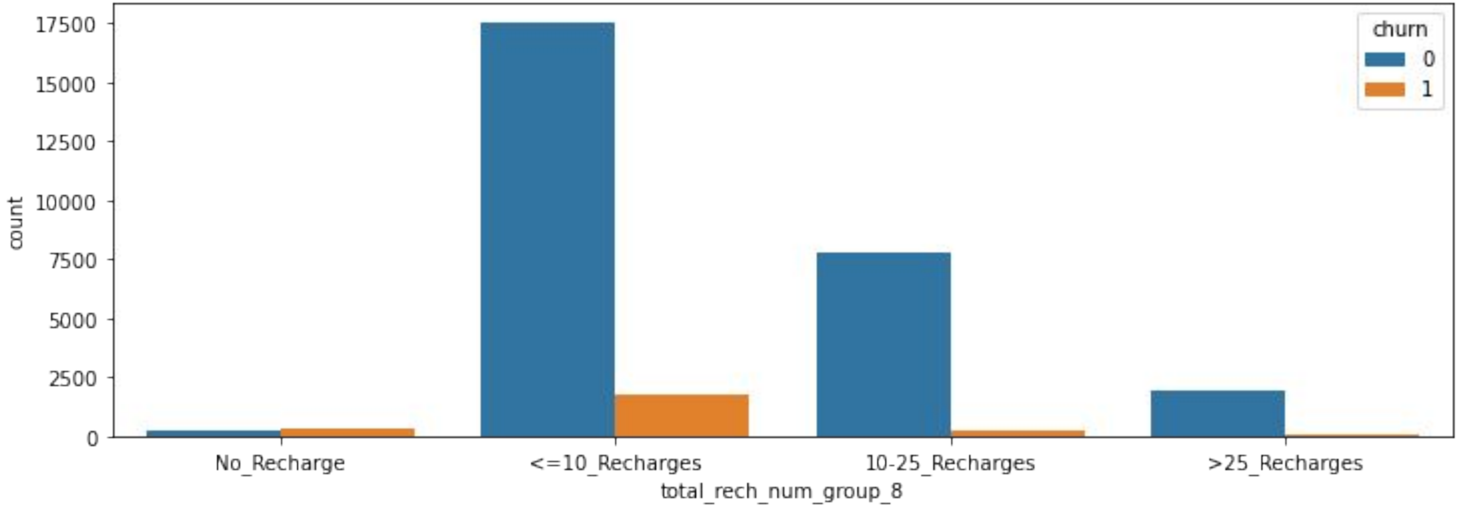


From the above plot , its clear tenured customers do no churn and they keep availing telecom services


Distribution of Max Recharge Amount by churn
No-Churn
Churn
Density
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0
1000
2000
3000
4000
Volume based cost


Distribution of Average Recharge Amount for Data by churn
No-Churn
Churn
Density
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0
1000
2000
3000
4000
Volume based cost


churn
0
1
count
14000
12000
10000
8000
6000
4000
2000
0
No_Recharge
<=10_Recharges
10-25_Recharges
>25_Recharges
total_rech_data_group_8


churn
0
1
count
17500
15000
12500
10000
7500
5000
2500
0
No_Recharge
<=10_Recharges
10-25_Recharges
>25_Recharges
total_rech_num_group_8

MODEL BUILDING

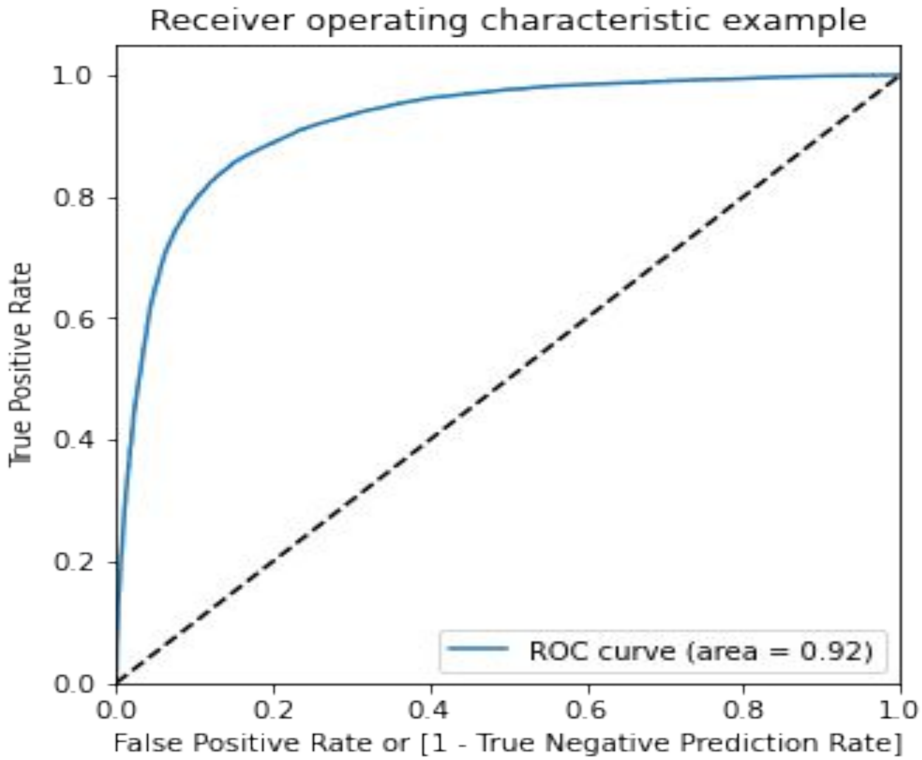
- create y dataset for model building.
- Split the dataset into train and test datasets
- Apply scaling on the dataset

Data Imbalance Handling

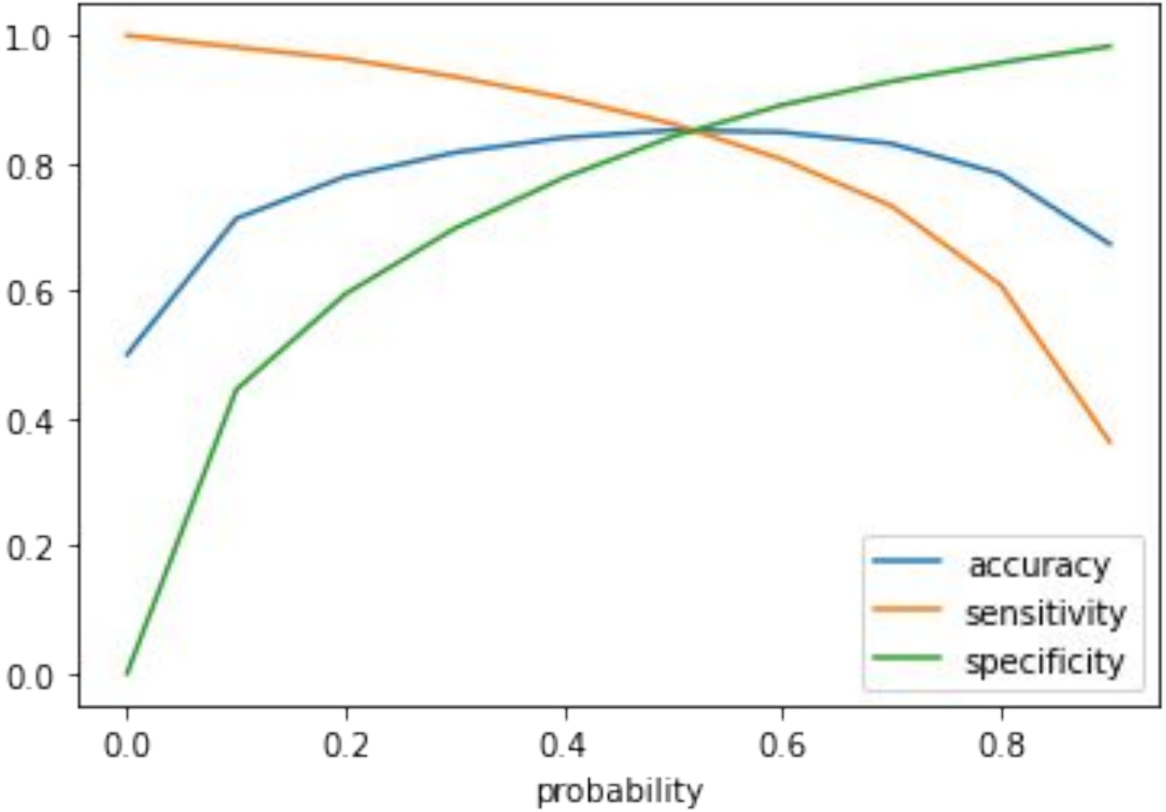
Using SMOTE method, we can balance the data w.r.t. churn variable and proceed further

1. **Logistic Regression using Feature Selection (RFE method)**
2. ***Assessing the model with StatsModels***
3. ***Creating a dataframe with the actual churn flag and the predicted probabilities***
4.

Plotting the ROC Curve

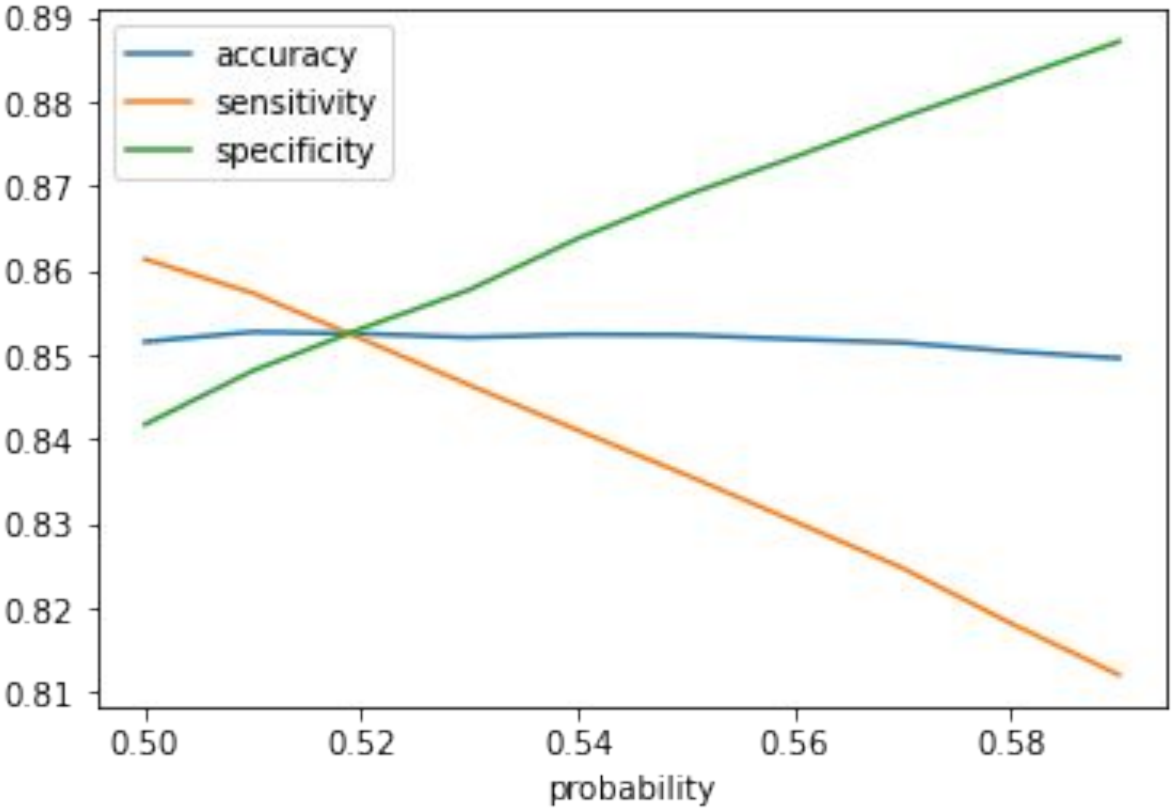


Plotting accuracy sensitivity and specificity for various probabilities



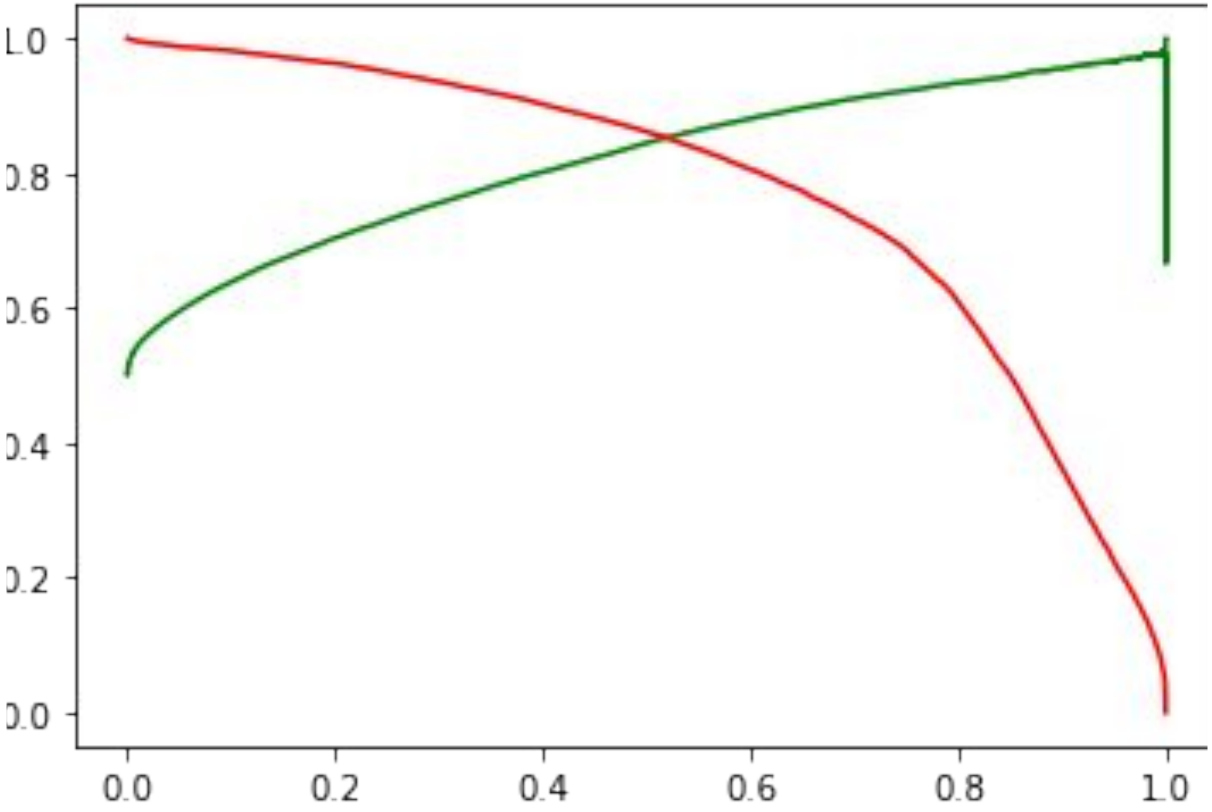
From the above graph, we can see the optimum cutoff is slightly higher than 0.5 but lies lower than 0.6. So lets tweek a little more within this range.

Plotting accuracy sensitivity and specificity for various probabilities calculated above.



From the above graph we can conclude, the optimal cutoff point in the probability to define the predicted churn variabe converges at `0.52`

Precision and recall tradeoff



From the above graph we can conclude, the optimal cutoff point in the probability to define the predicted churn variabe converges at `0.52`

Making predictions on the test set

1. Transforming and feature selection for test data
2. Predicting the target variable
3. Metrics Evaluation

Explaining the results¶

The accuracy of the predicted model is: 84.0 %
The sensitivity of the predicted model is: 81.0 %

As the model created is based on a sensitivity model, i.e. the True positive rate is given more importance as the actual and prediction of churn by a customer

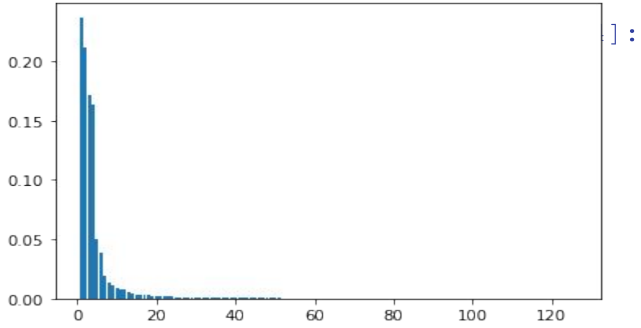
Performing Logistic Regression

Confusion Matirx for y_test & y_pred

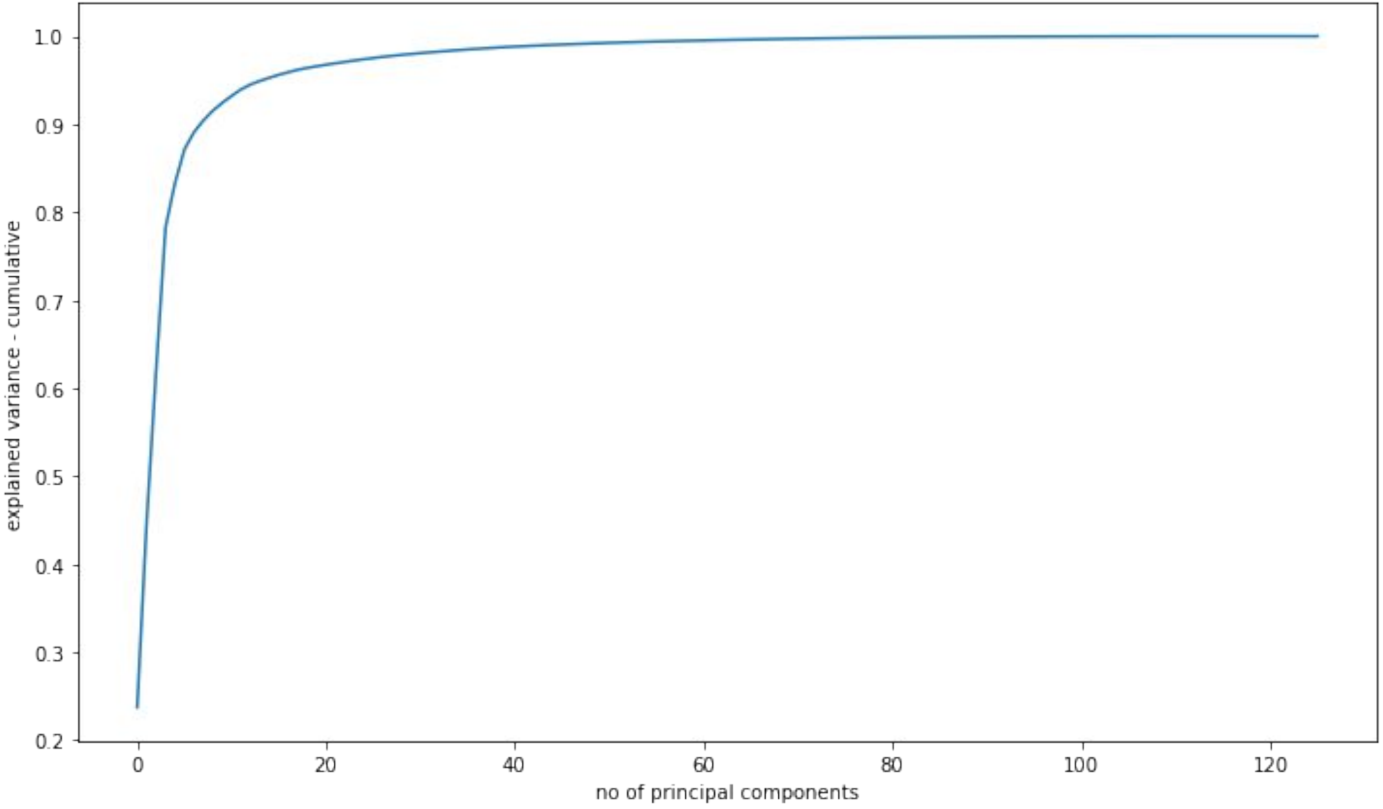
[[6761 1511]

[126 603]]

Accuracy of the logistic regression model with PCA: 0.818131318742362



Making a scree plot



90% of the data can be explained with 8 PCA components*

Confusion Matirx for y_test & y_pred
[[6247 2025]
[184 545]]

Accuracy of the logistic regression model with PCA: 0.7545828241306521

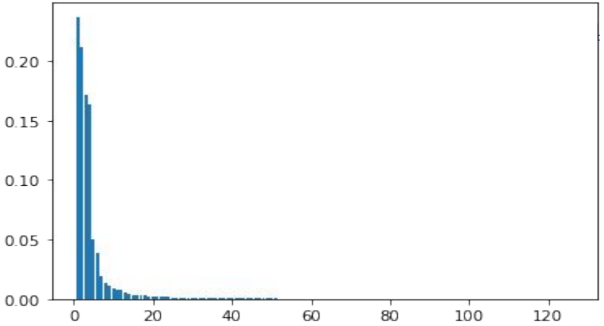
Performing Logistic Regression

Confusion Matirx for y_test & y_pred

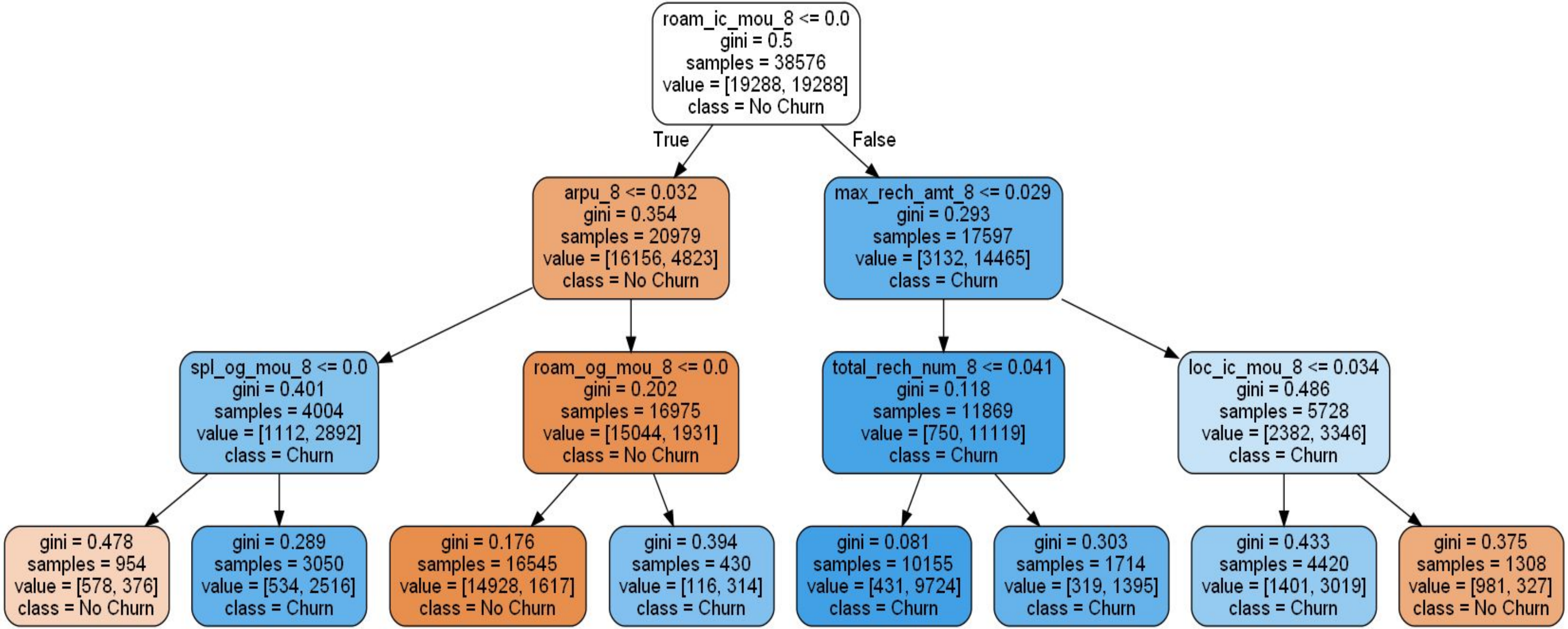
[[6761 1511]

[126 603]]

Accuracy of the logistic regression model with PCA: 0.818131318742362



Decision Tree

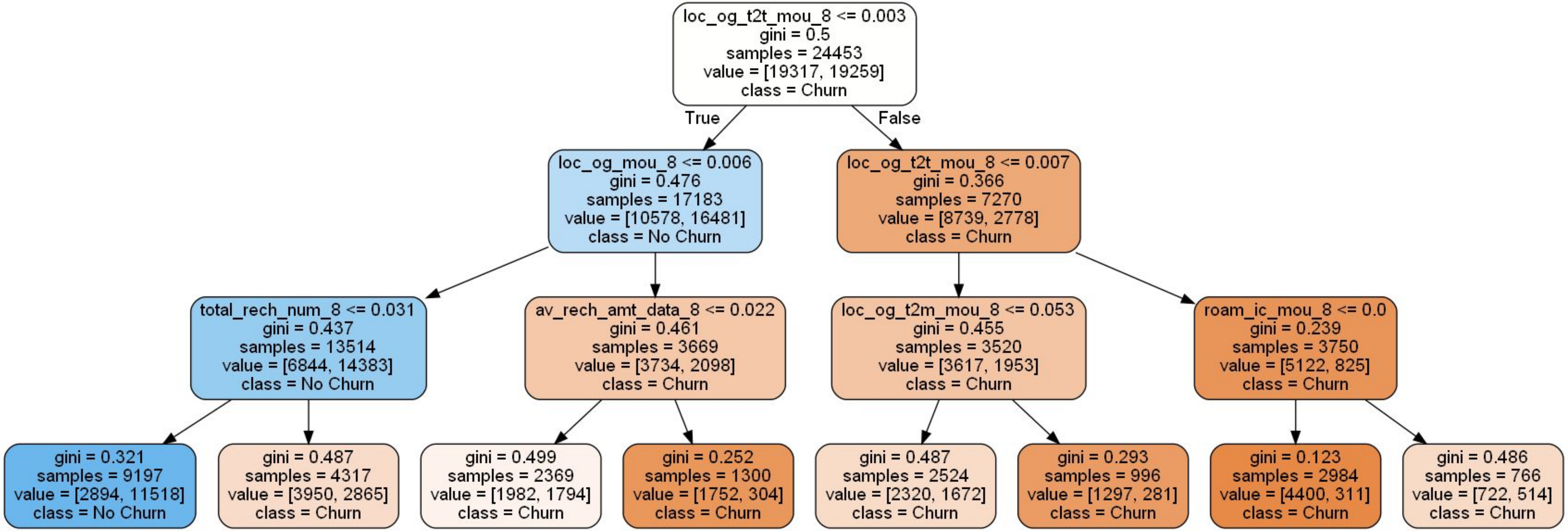


Accuracy of various models

Model	Accuracy
Logistic Regression	84.000000
Logistic Regression with PCA	0.754583
Decision Tree	0.843462

Logistic Regression Model. Decision Tree Classifier. Random Forest Classifier. The above models were initially created with default parameters which did not give accurate results and the score metrics were not good. Then we hypertuned each model and recreated them with the best estimators. The hyper tuned model showed an increase in the classification scores though marginally.

Random Forest



Summery of scores

Train Accuracy : 0.6381169639153879
Train Confusion Matrix:
[[18478   810]
 [13150 6138]]

Test Accuracy : 0.9002333074102877
Test Confusion Matrix:
[[7899 373]
 [525 204]]

3. Conclusion

Top 7 Features affecting churn

- roam_og_mou_8
- roam_ic_mou_8
- arpu_8
- max_rech_amt_8
- total_og_mou_8
- last_day_rch_amt_8
- av_rech_amt_data_8

Our Random Forest model is a decent model. We are able to predict with accuracy of 90.05 %

.

Thank you