Telecom Customer Churn Case Study

BY- PANTHANGI VANI

INTRODUCTION

Churn prediction is one of the most popular Big Data use cases in business. It consists of detecting customer who are likely to cancel a subscription to a service.

Churn is a problem for telecom companies because it is more expensive to acquire a new customer than to keep your existing one from leaving.

PROBLEM STATEMENT

Business problem overview

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. Given the fact that it costs 5-10 times more to acquire a new customer than to retain an existing one, **customer retention** has now become even more important than customer acquisition.

For many incumbent operators, retaining high profitable customers is the number one business goal.

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

In this project, we will analyse customer-level data of a leading telecom firm, build predictive models to identify customers at high risk of churn and identify the main indicators of churn.

OBJECTIVE

To predict customer churn

Highlighting the main variables/ factors influencing Customer Churn.

Use various ML algorithms to build prediction models, evaluate the accuracy & performance of these models.

Finding out the best model for our business case & providing executive summary.

DESCRIPTION

Source dataset is in csv format.

Dataset contains 7043 rows and 14 columns.

There is no missing values for the provided input dataset.

Churn is the variable which notifies whether a particular customer is churned or not.

METHODOLOGIES

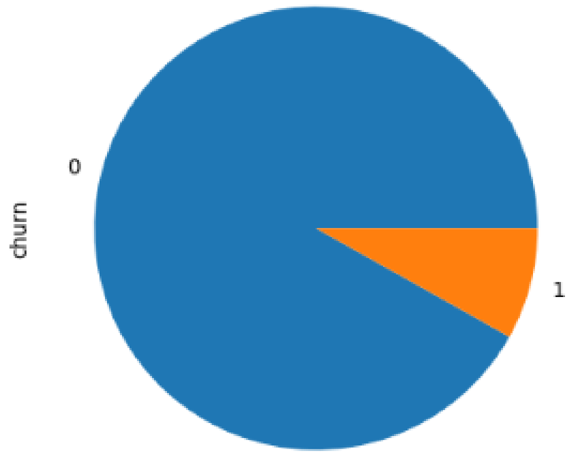
EDA(Exploratory Data Analysis): The dataset consists of 12 variables in all. A few are continuous, rest are categorical. The control variable was customer.

Model building which includes defining the purpose if model, determine the model boundary , build the model, create an interface & export the model.

Evaluating machine learning algorithm is an essential part of project.

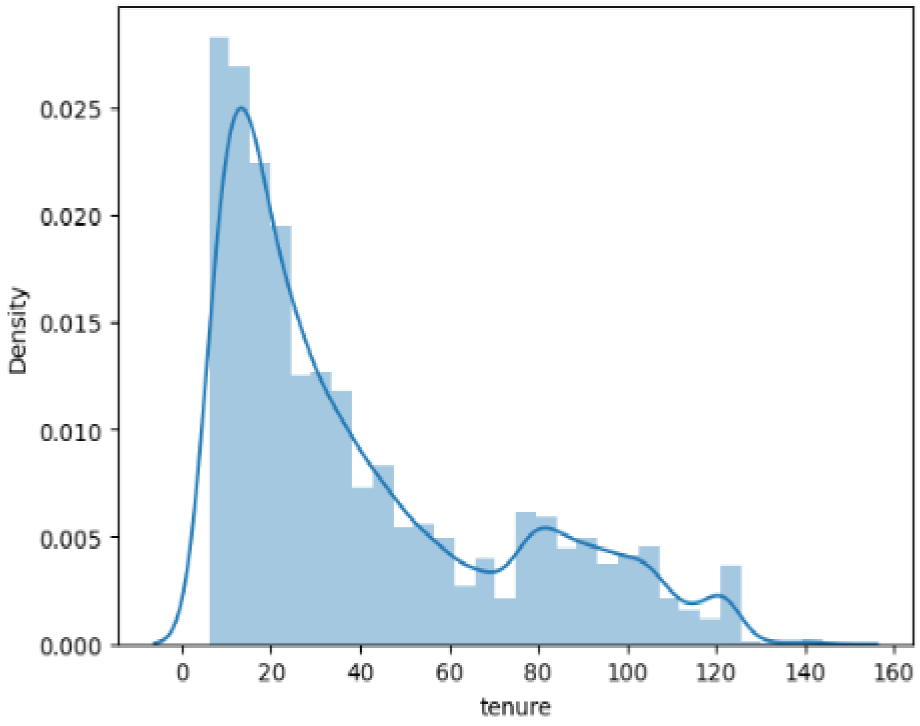
Churn/non churn percentage

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

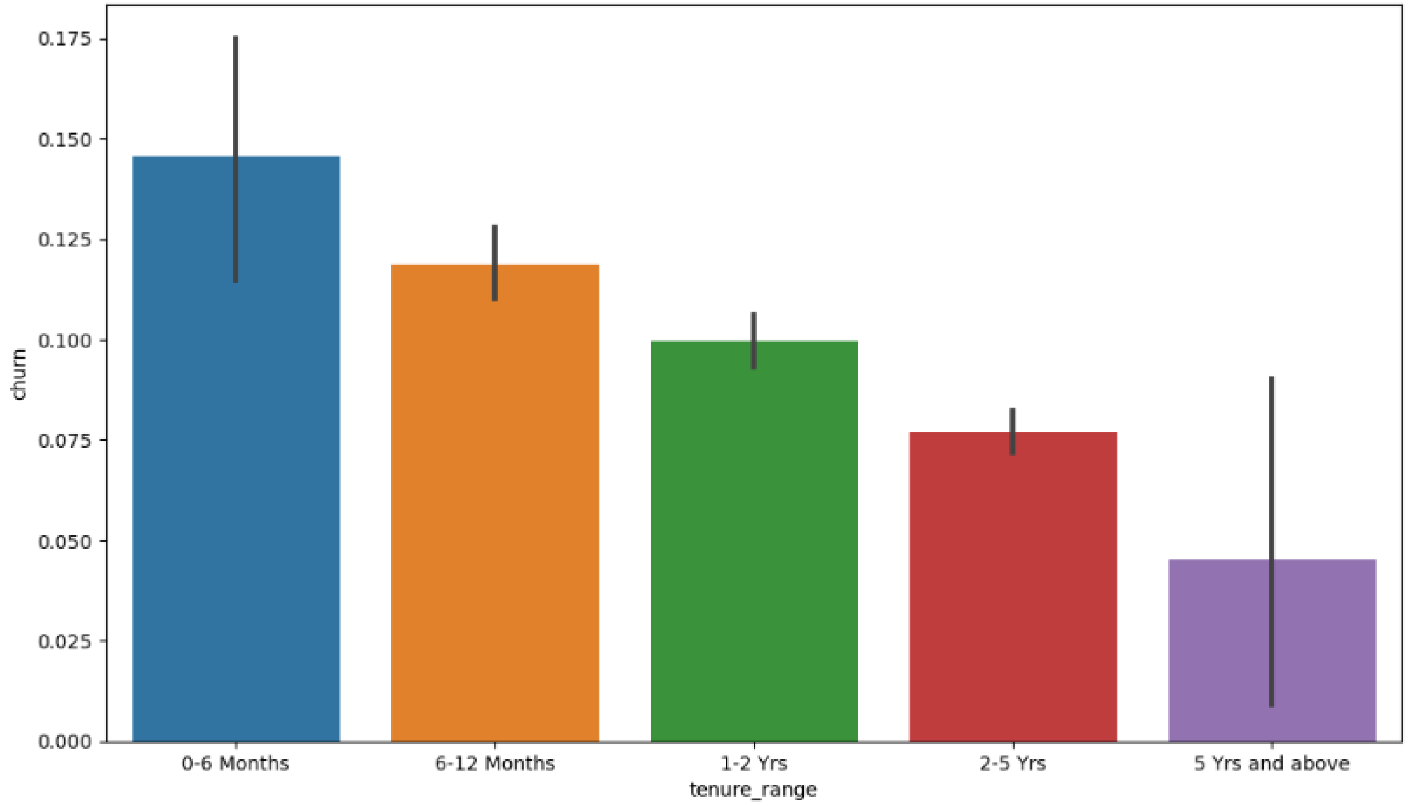


As we can see that 91% of the customers do not churn, there is a possibility of class imbalance

Distribution of the tenure variable

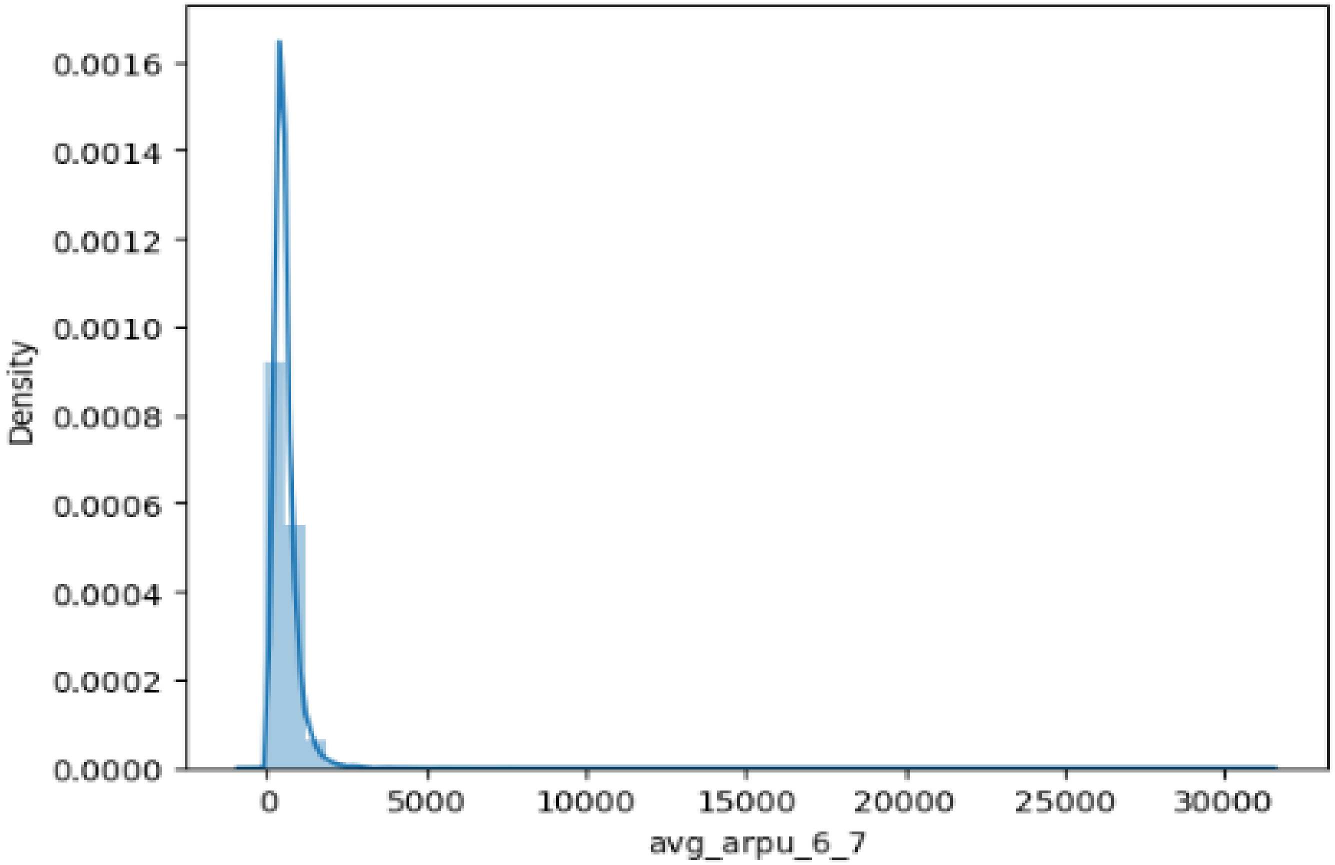


Bar plot for tenure range

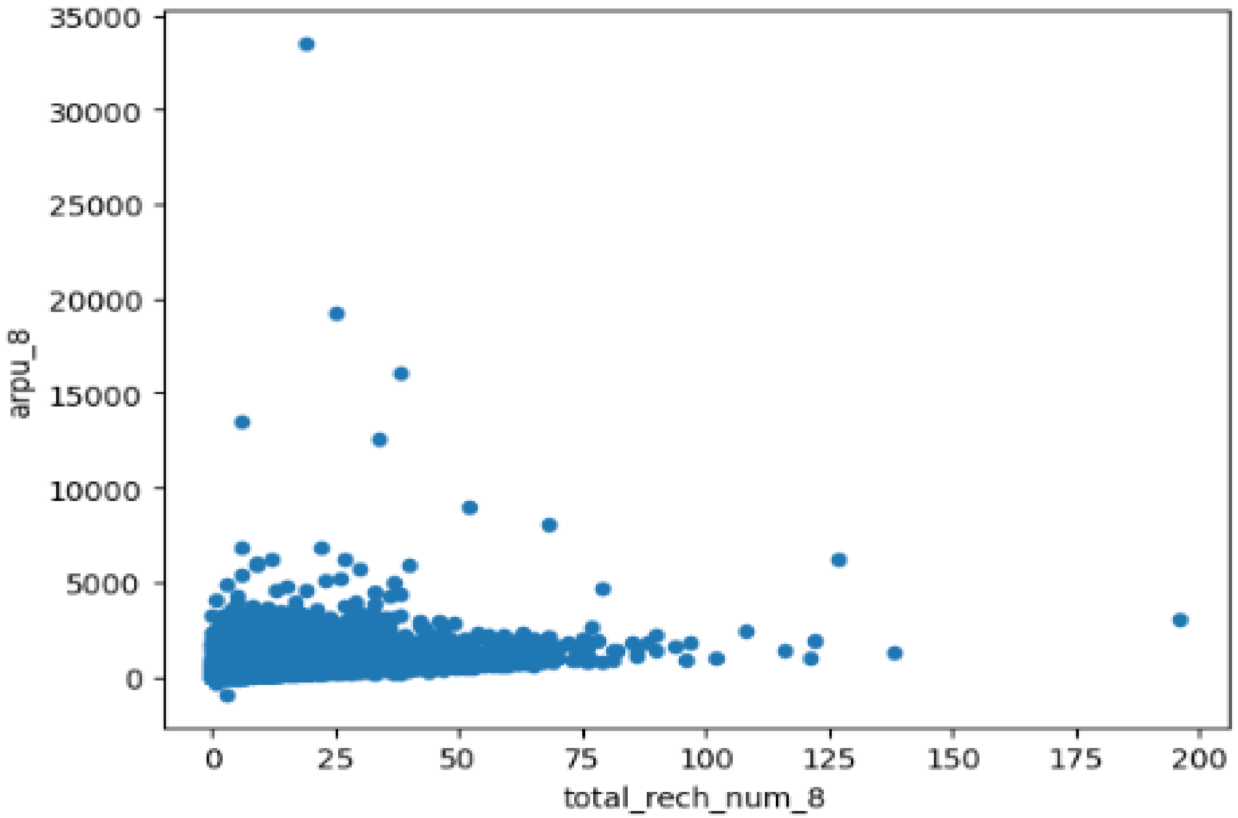


It can be seen that the maximum churn rate happens within 0-6 month, but it gradually decreases as the customer retains in the network.

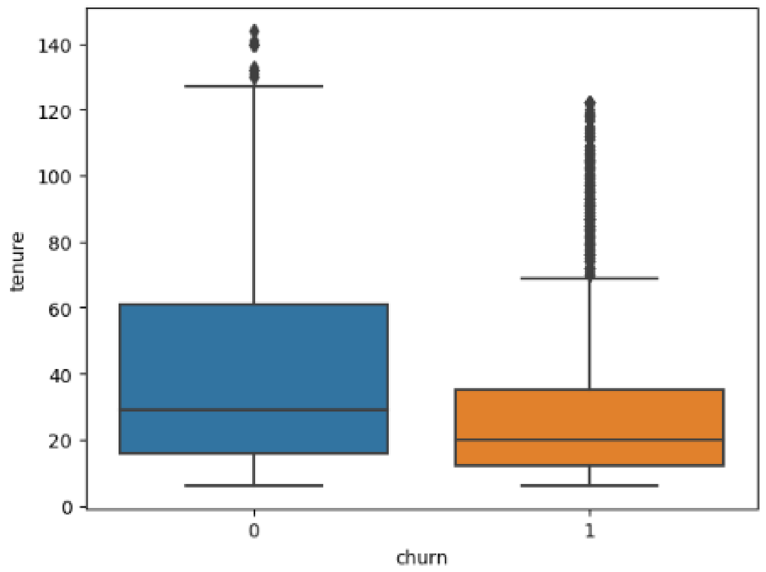
Distplot(telecom_data['avg_arpu_6_7’]



Scatter plot between total recharge and avg revenue for the 8th month

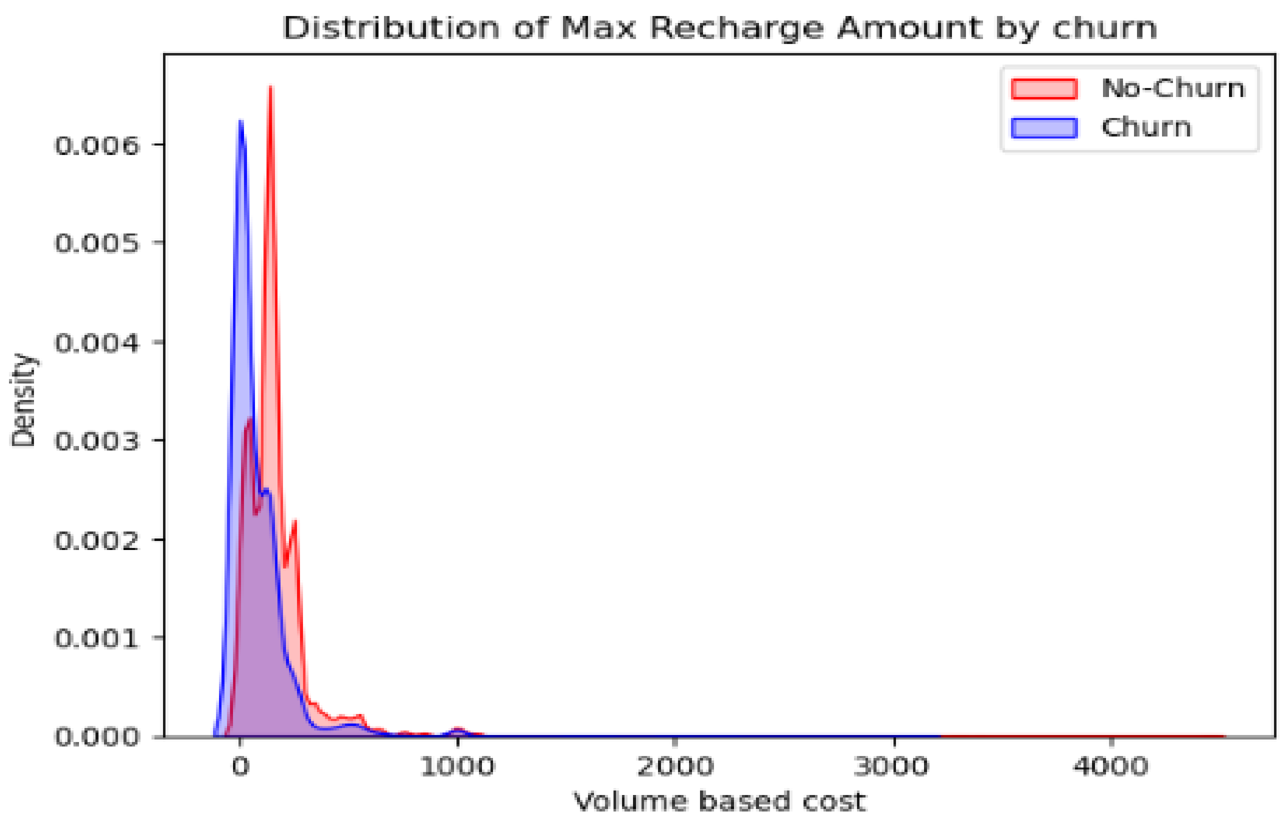


Boxplot between Churn & Tenure

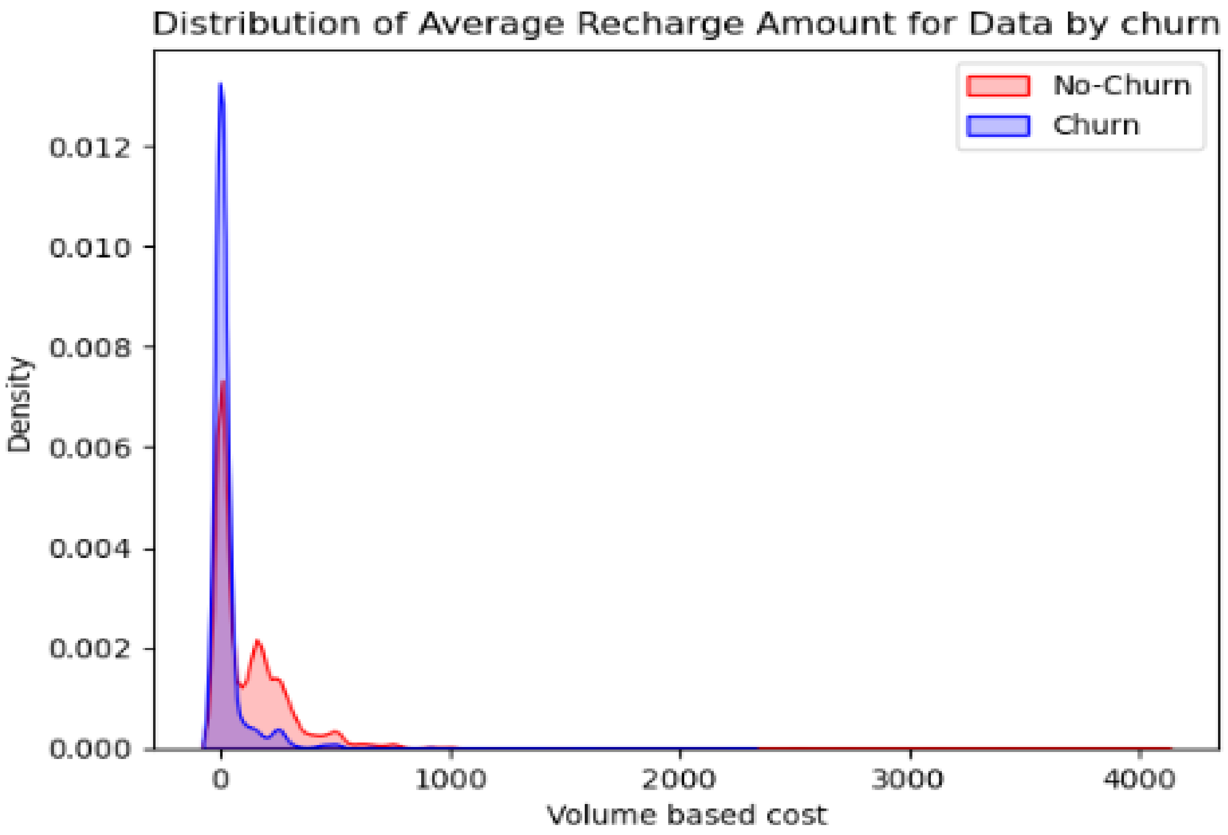


From the above plot , its clear tenured customers do no churn and they keep availing telecom services

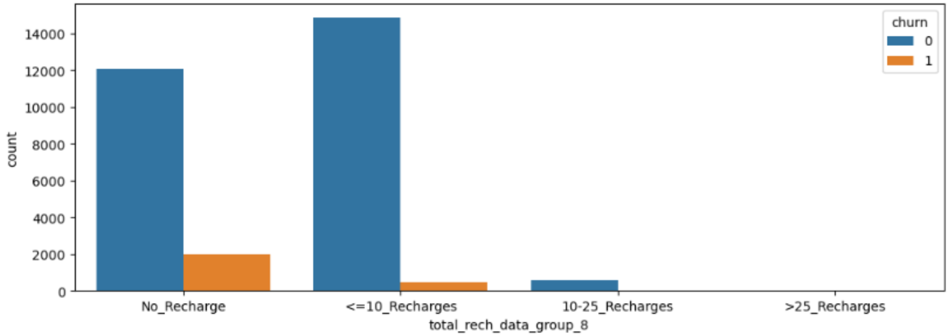
Plot between churn vs max rechare amount



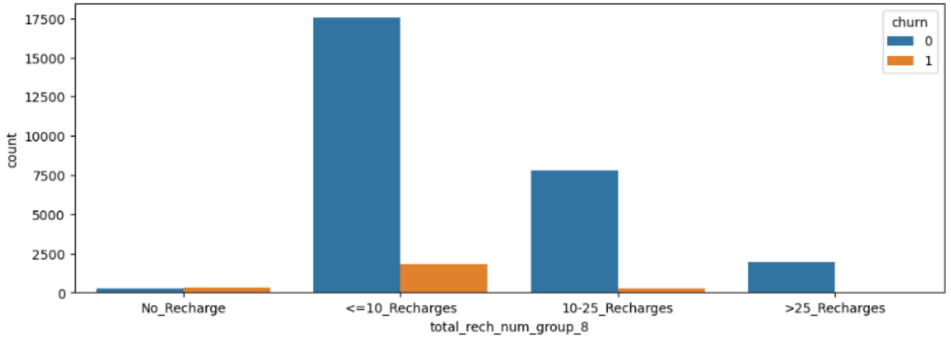
Churn vs max recharge amount



Results

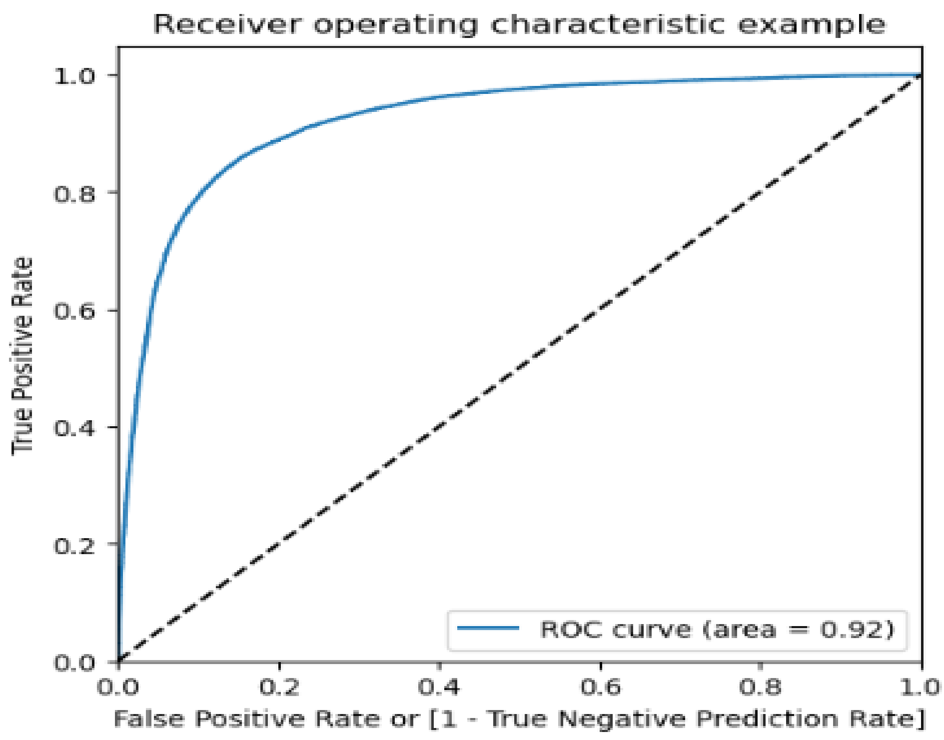



```
                    Distribution of total_rech_num_8 variable
 <=10_Recharges     19349
10-25_Recharges      8073
>25_Recharges        1996
No_Recharge           583
Name: total_rech_num_group_8, dtype: int64
```

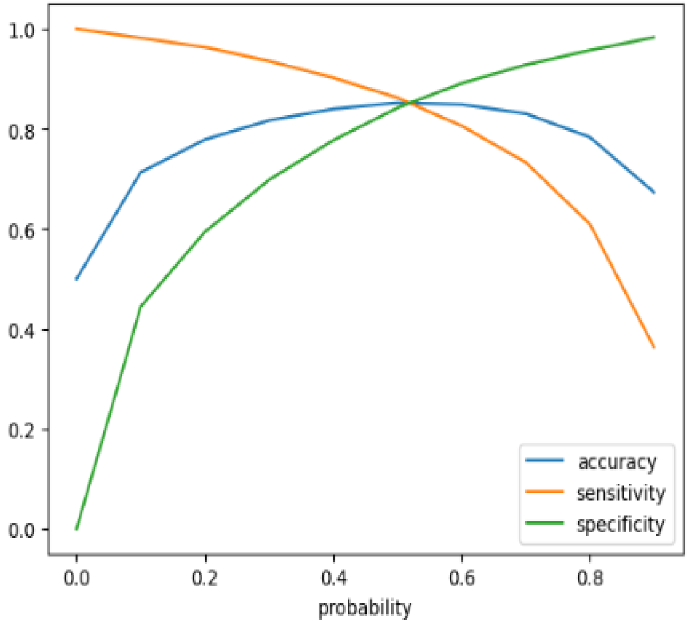


As the number of recharge rate increases, the churn rate decreases clearly.

The ROC curve for the obtained metrics



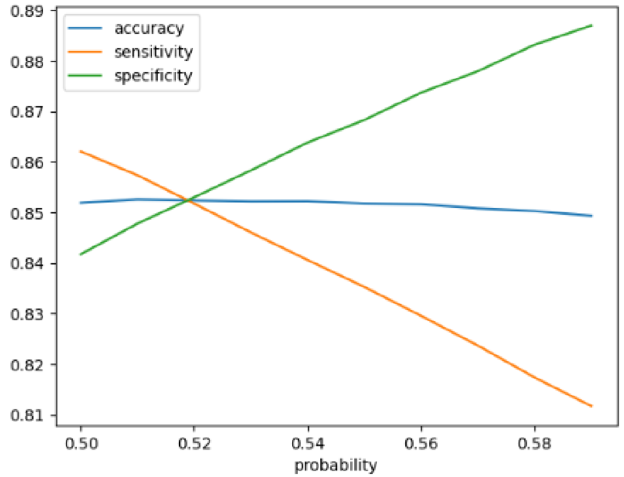
Accuracy, Sensitivity and Specificity for various probabilities calculated



Initially we selected the optimum point of classification as 0.5.

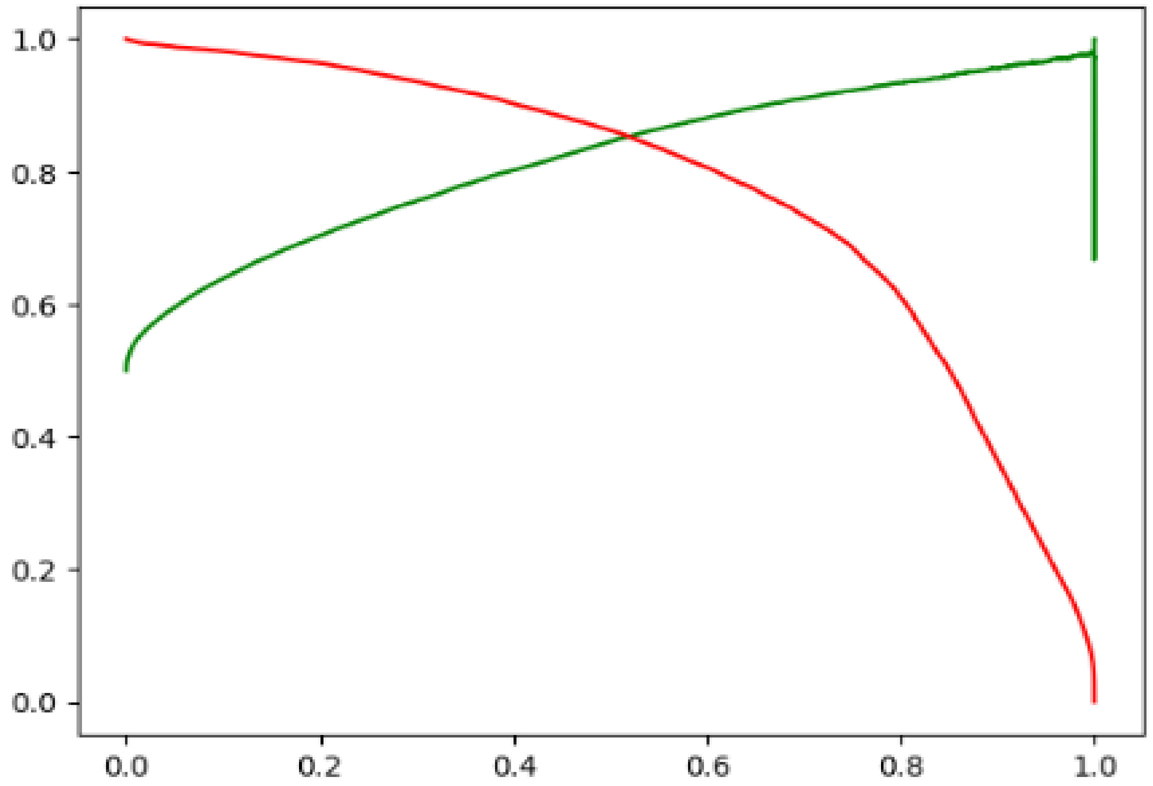
From the above graph, we can see the optimum cutoff is slightly higher than 0.5 but lies lower than 0.6. So lets tweek a little more within this range.

Accuracy, Sensitivity and Specificity for various probabilities calculated

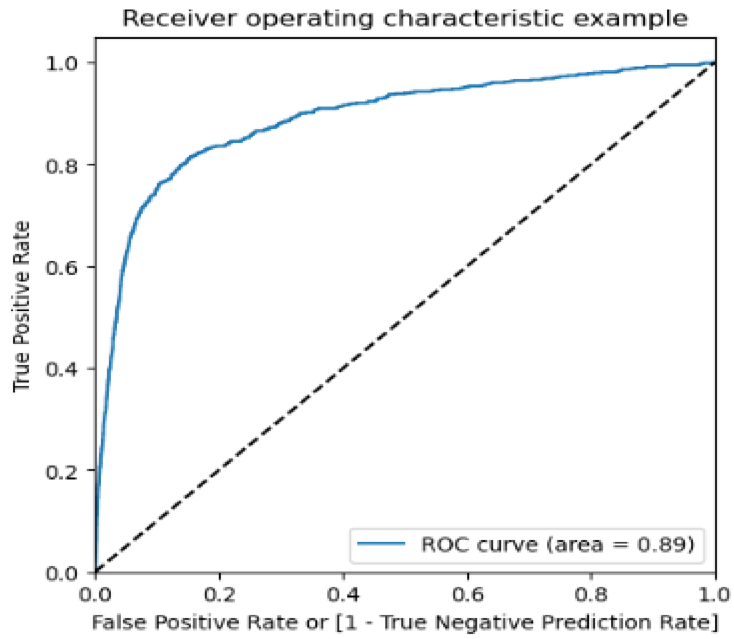


From the above graph we can conclude, the optimal cutoff point in the probability to define the predicted churn variabe converges at 0.52

Precision and recall tradeoff



ROC curve for the test dataset



The AUC score for train dataset is 0.92 and the test dataset is 0.89.

This model can be considered as a good model.