Telecom Churn Case Study

Arnav Narayan

Problem Statement

- In the telecom industry, customers can 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.

Methodology

Step-1:

Data Importing, Inspecting, Cleaning & Manipulation

a) Filtering high-value customers

b) Tagging Churn Customers

c) Handling NA or Missing Values.

d) Dropping of Unnecessary Columns (i.e., which are not taken for in Analysis)

e) Dropping of Columns having large number of missing values.

f) Imputation of Values where required.

Step-2:

Exploratory Data Analysis

- Univariate Analysis.
- Bivariate data analysis.

Step-3:

Model Building Preparation & Validation

a) Test-Train Split

b) Scaling

Step-4:

Model Evaluation

a) Creating a data frame with the actual conversion flag and predicted probabilities Creating a new column 'Predicted’

b) Finding the Optimal Cutoff

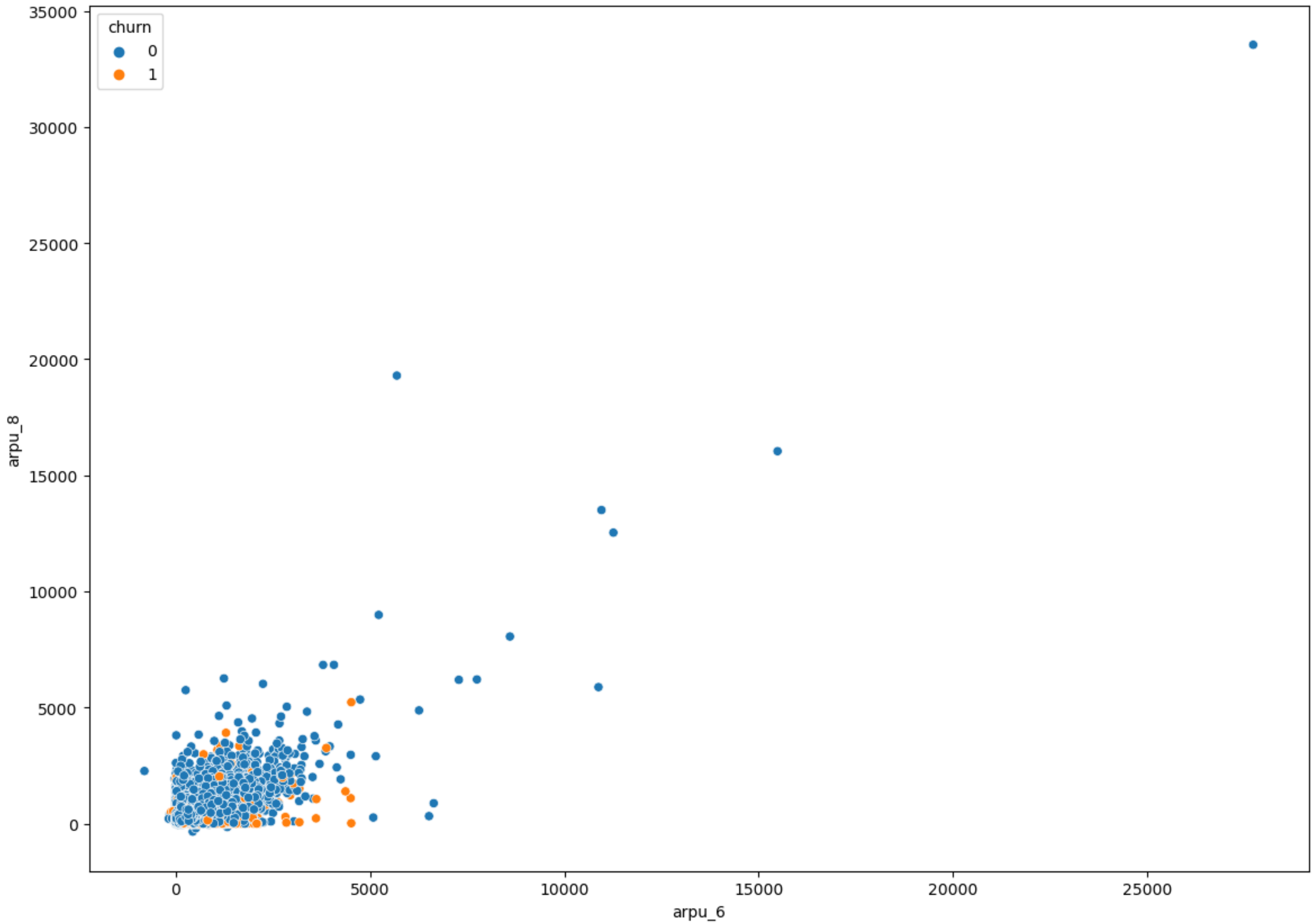
Step-5:

Making Predictions based on the Test Set and deriving conclusion.

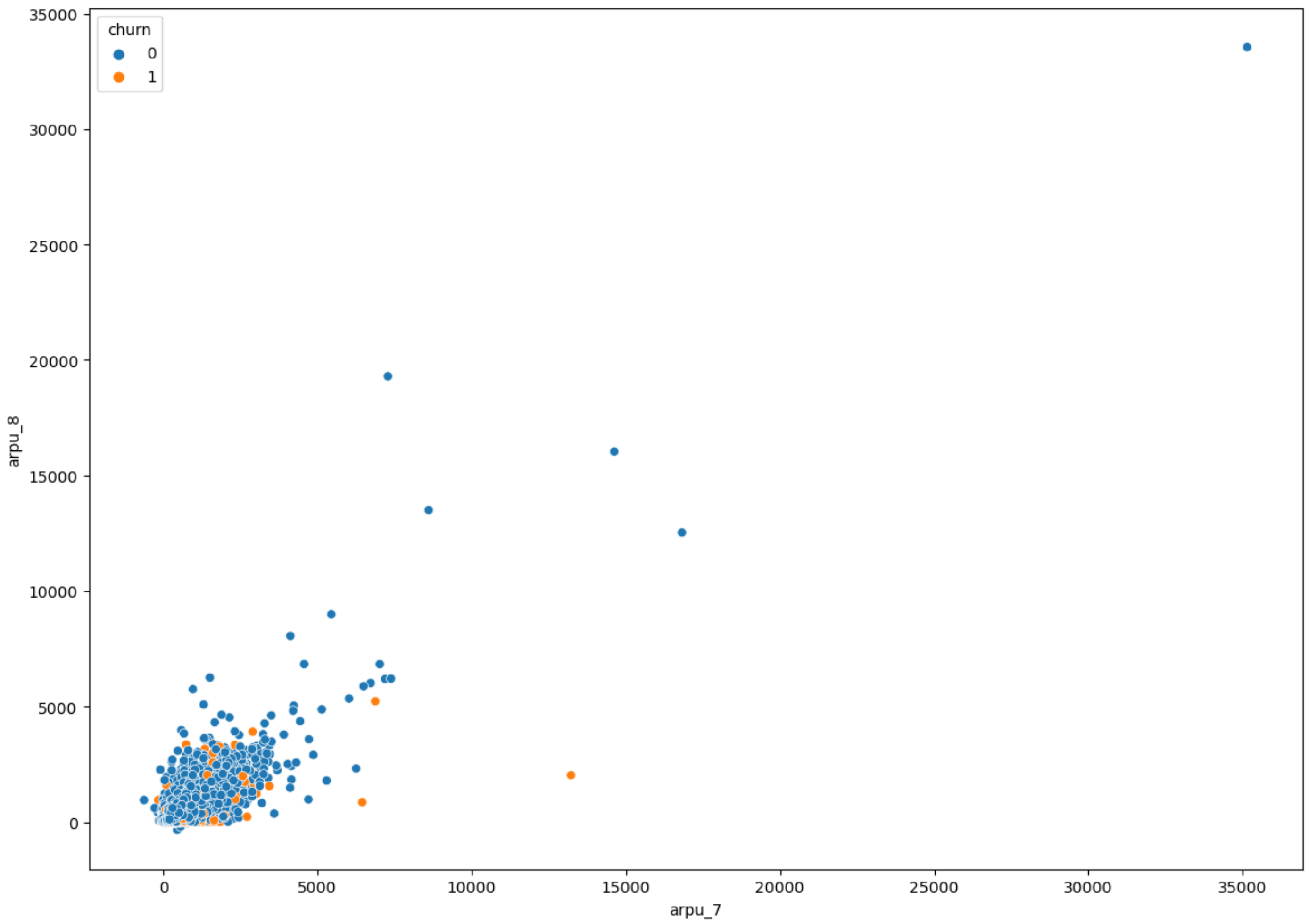
Data Cleaning and Preparation

- We dropped all the columns which had more than 30% values missing. Since 30% is a sufficient amount we dropped them.
- We Checked the other remaining columns and drop columns which are not required for our analysis.
- The columns with null or missing values, we imputed them with 0.0
- We Checked the other remaining columns and dropped the columns which are not required for our analysis

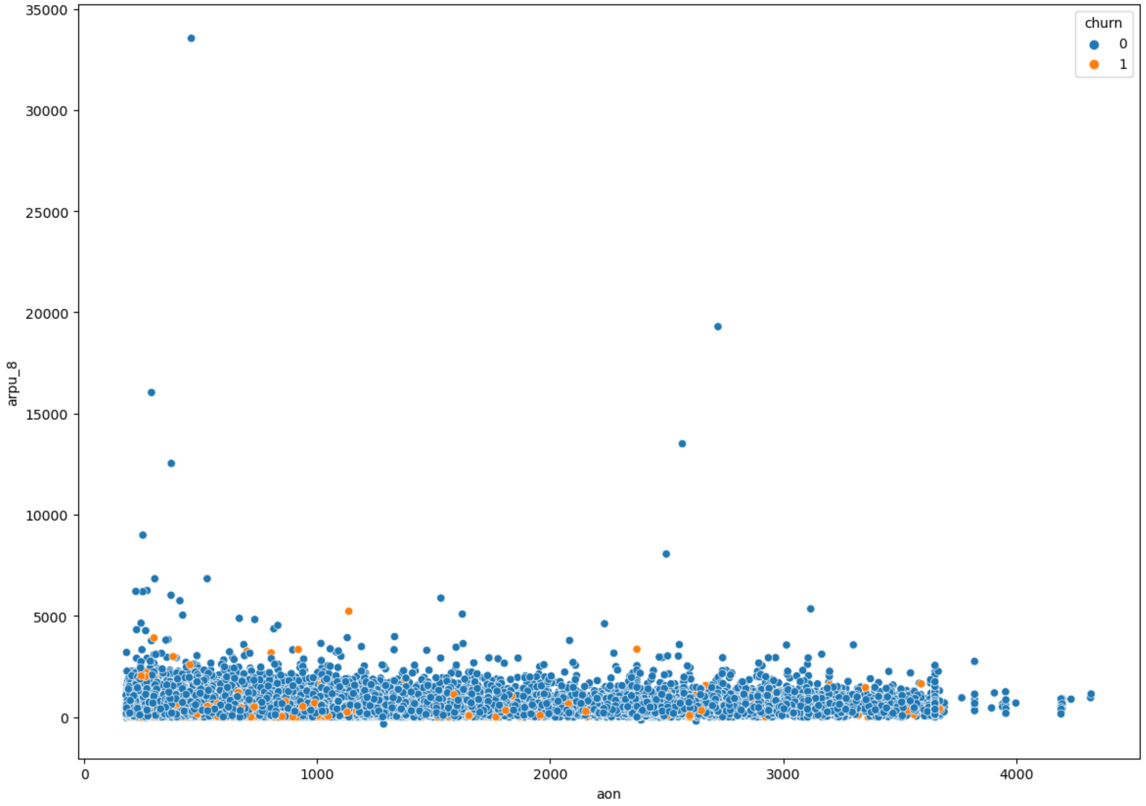
EDA



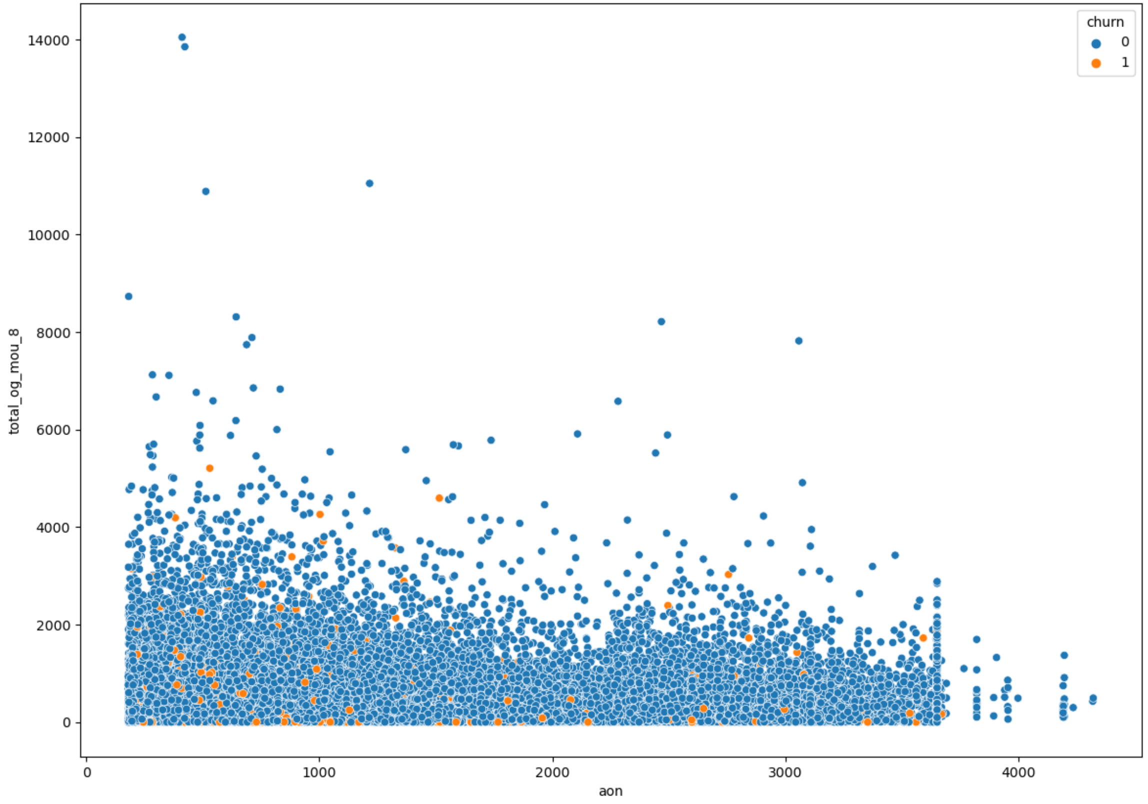
churn
0
1
35000
30000
25000
20000
15000
10000
5000
0
arpu_8
0
5000
10000
15000
20000
25000
arpu_6



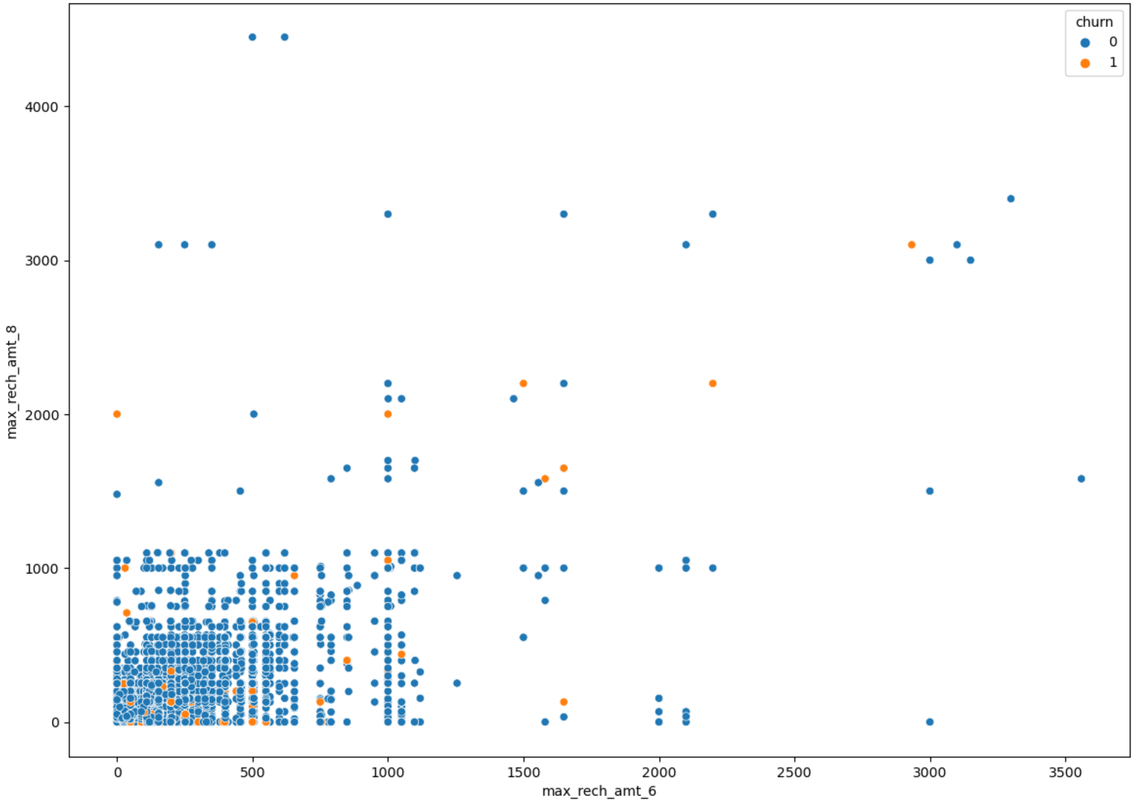
We can observe that customers with lesser "arpu" in 6,7 are more likely to churn



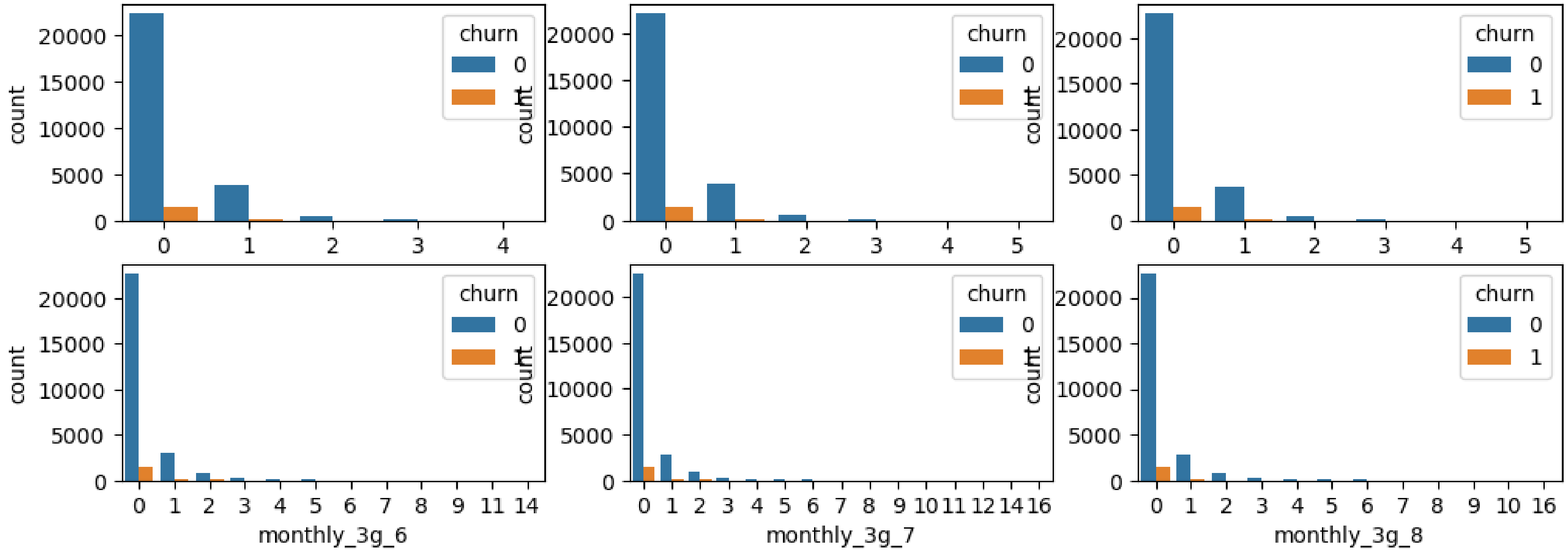
We can observe that customers with lesser "arpu_8" and "aon" are more likely to churn.



We can observe that customers with lesser "total_og_mou_8" and "aon" are more likely to churn.



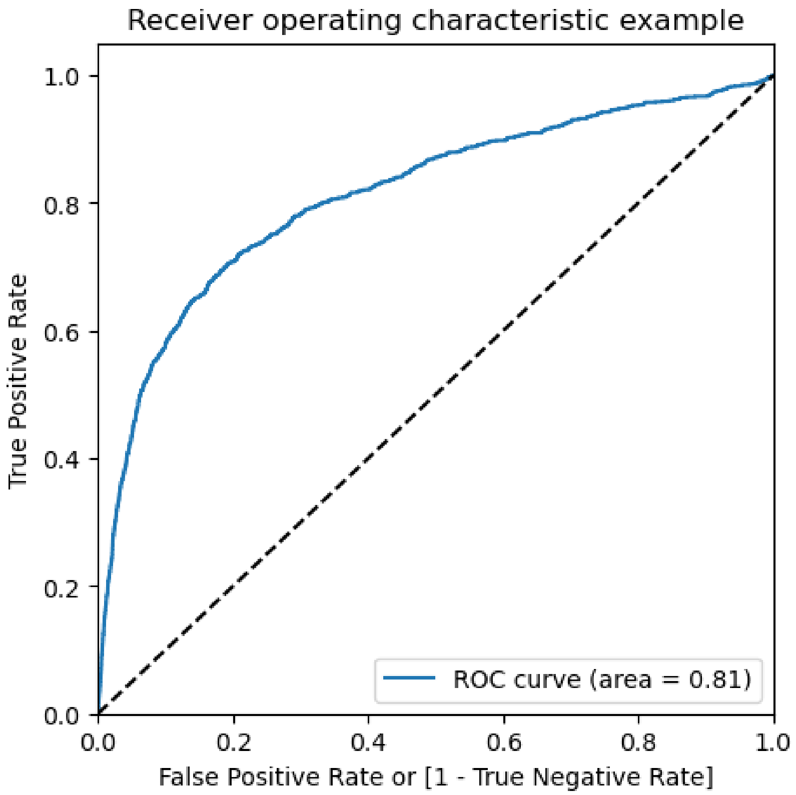
we can clearly observe that lesser "max_rech_amt_6" and "max_rech_amt_8" indicates higher churn chance.s

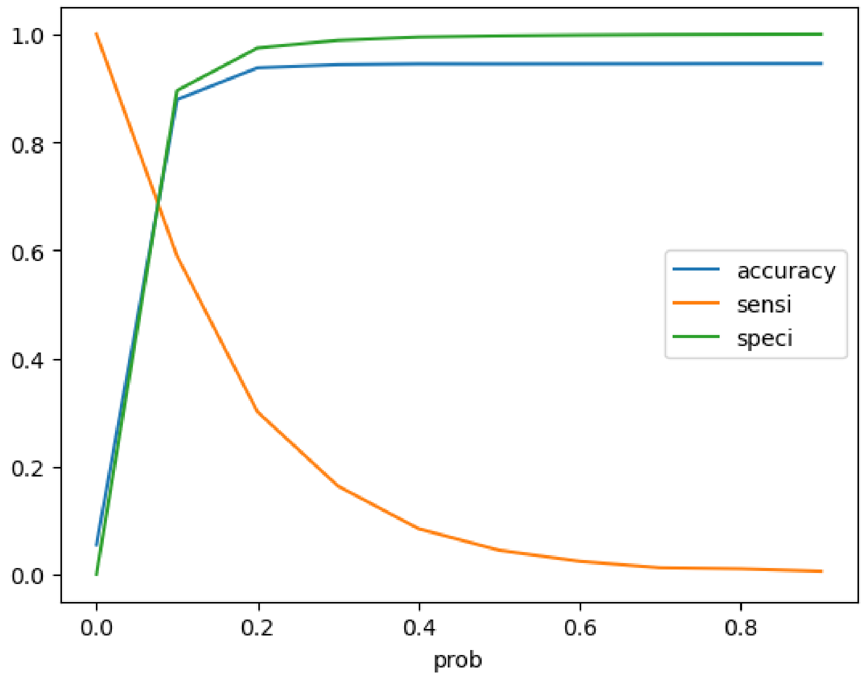


from the plot above we can observe that Higher use of 2g or 3g indicates lesser chances of churn.

Model Building

- Splitting the Data into Training and Testing Sets.
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Generalized Logistic Regression Results.
- Feature Selection Using RFE.
- Building Model.
- Assessing the model.
- Predictions made based on test data set.





1) Since we know that the perfect ROC Curve should be a value close to 1. We are getting a value of 0.81 indicating a good predictive model.

2) From the curve above, we see that 0.15 is the optimum point to take it as a cutoff probability.

Conclusion

- So using Logistic regression we are getting an accuracy of 94% on train data and 92% on test data.
- Roaming outgoing in month 7 and 8 are strong indicators, hence telecom company should reduce the costs for roaming. They need to provide good offers to the customers who are using services from a roaming zone.
- MOU is one of the major factors.
- Max recharge in the good phase are indicators of churn.
- We can clearly see most of the critical features are form the action phase, which is inline with the business understanding that action phase needs more attention.