TELECOM CHURN CASE STUDY

Identification of typical behaviors of high value customers during Churn in order for the firm to make business decisions based on them

Team Members -Rakshith P Shetty, Rucha Pusdekar & Rajeev Dubey

INDEX

- Problem Statement & Expected Outcome
- Understanding Data
- Steps taken for the analysis
- Data Cleaning
- EDA
- Data Preparation
- Feature Selection
- Model Evaluation
- Recommendation

Problem Statement & Expected Outcome

Background:

- Customers in the telecom business can select among many service providers and actively switch from one operator to another. The telecommunications business has an annual churn rate of 15-25% in this highly competitive market
- Given that it costs 5-10 times more to acquire a new customer than it does to retain a current one, customer retention has now overtaken customer acquisition as the most critical factor.

Problem Statement:

- In this project, you will evaluate customer-level data from a large telecom company, develop predictive models to identify customers at high risk of churn, and identify the main churn indicators.

Expected Outcome:

- The number one company goal is to keep high-profitable clients. To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.

Understanding Data

- The dataset includes customer-level data for four consecutive months: June, July, August, and September. The months are represented by the numbers 6, 7, 8, and 9.
- The business goal is to predict churn in the final (i.e. ninth) month using data (features) from the first three months. Understanding normal client behaviors during turnover will be beneficial in performing this duty efficiently.
- Assumed 3 stages of Churn
  - The ‘good’ phase: In this phase, the customer is happy with the service and behaves as usual.
  - The ‘action’ phase: The customer experience starts to sore in this phase, for e.g. he/she gets a compelling offer from a competitor, faces unjust charges, becomes unhappy with service quality etc. In this phase, the customer usually shows different behavior than in the ‘good’ months. Also, it is crucial to identify high-churn-risk customers in this phase, since some corrective actions can be taken at this point
  - The ‘churn’ phase: In this phase, the customer is said to have churned.

Steps taken for analysis

Step 1
Data Cleaning:
Load, Understand and clean the dataset

Step 2
Data Preparation:
Filter for high value customers, Tag Churners

Step 3
EDA:
Understand the data though visualization

Step 4
Data split & Resampling :
Train test Split, SMOTE method is used to create class balance

Step 5
Preprocessing:
Feature Scaling by MinMaxScaler

Step 6
Model Selection:
Create multiple models and train the models with the resampled data

Step 7
Model Evaluation:
Confusion matrix, ROC Curve, Accuracy, Specificity & Sensitivity

Step 8
Prediction on Test set :
Compare eval metrics of Train and test dataset and consider the best model

Step 1: Data Cleaning

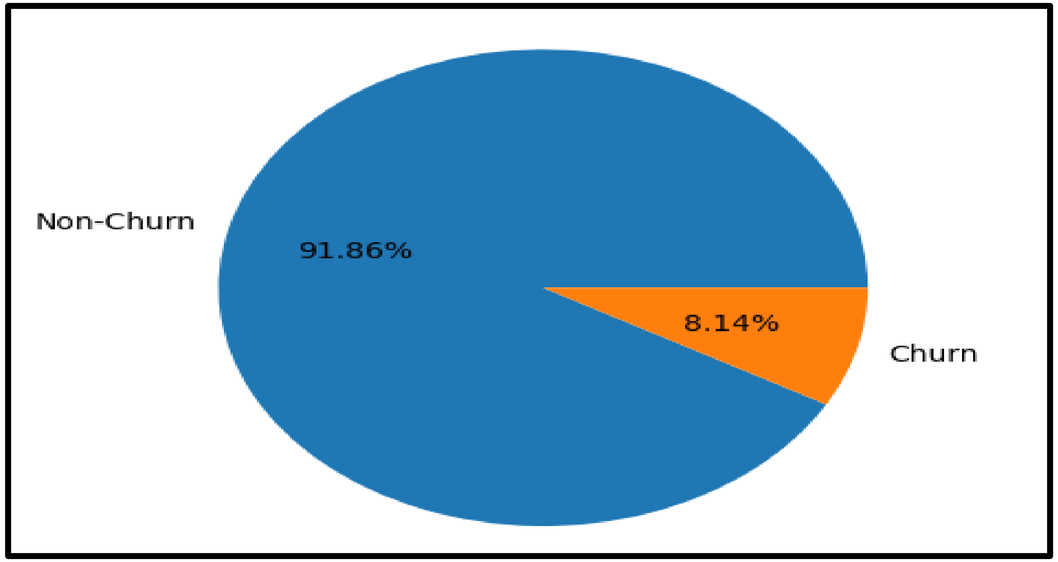
- Columns which had a single unique value/Date. These columns were removed since they offer no value to our data.
- Missing values in recharge & minutes of usage related columns were replaced by Zero. As the missing data says that the particular customer as not used that particular service
- Columns which is irrelevant or data imbalance with respect to analysis have been dropped.

Step 2: Data Preparation

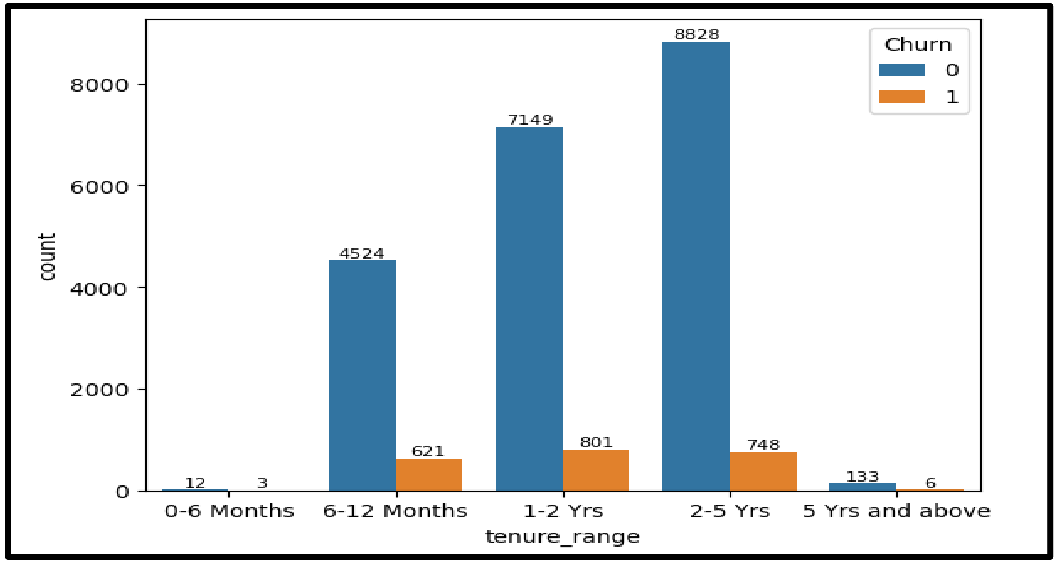
- As desired by the telecom company, an analysis be performed on High-Value customers, which are the top 30% of customers, based on the average recharge amount in the first two months (6&7), also known as the Good phase.
- Tag churners: Those who haven't made any calls (incoming or outgoing) and haven't used mobile internet even once during the churn period. To tag churners, utilize the parameters total_ic_mou_9, total_og_mou_9, vol_2g_mb_9, and vol_3g_mb_9.
- Drop all the columns that belongs to month 9.

Step 3: EDA

Churn Percentage



Distribution of Tenure

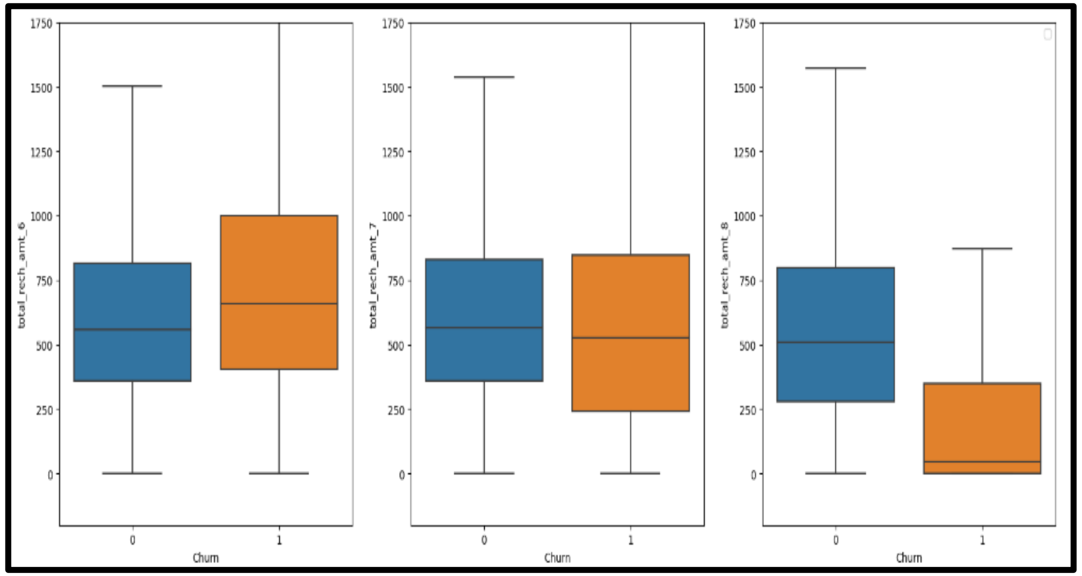


Inference:

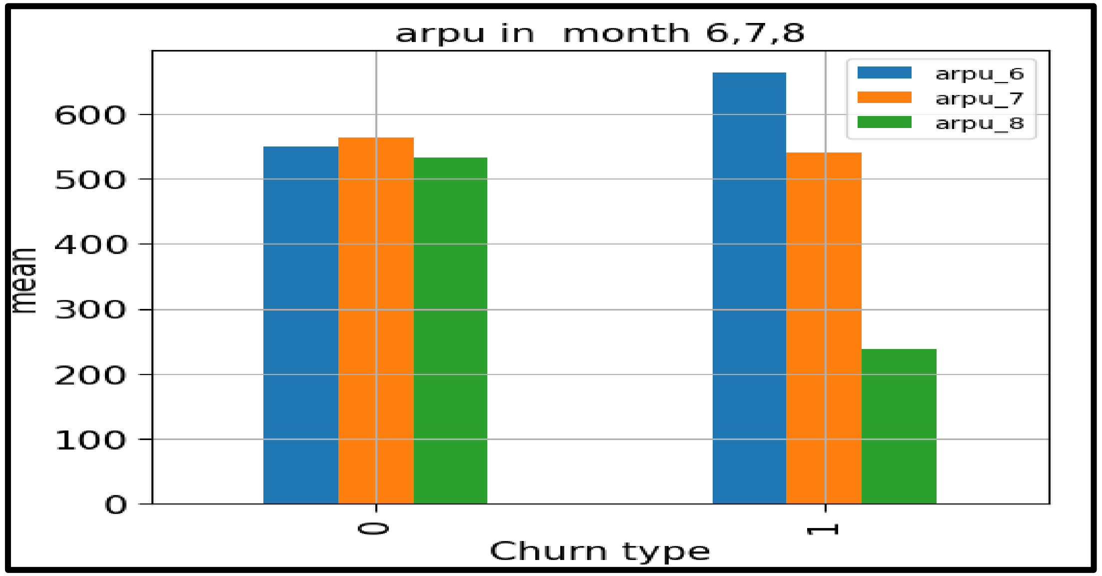
- Churn: Churn rate in high value customers is a little more than 8%.
- Tenure: If a tenure of a customer is above 5 years they are most likely to continue with the telecom service.

Step 3: EDA

Distribution of total recharge amount



Distribution of Average Rev per User

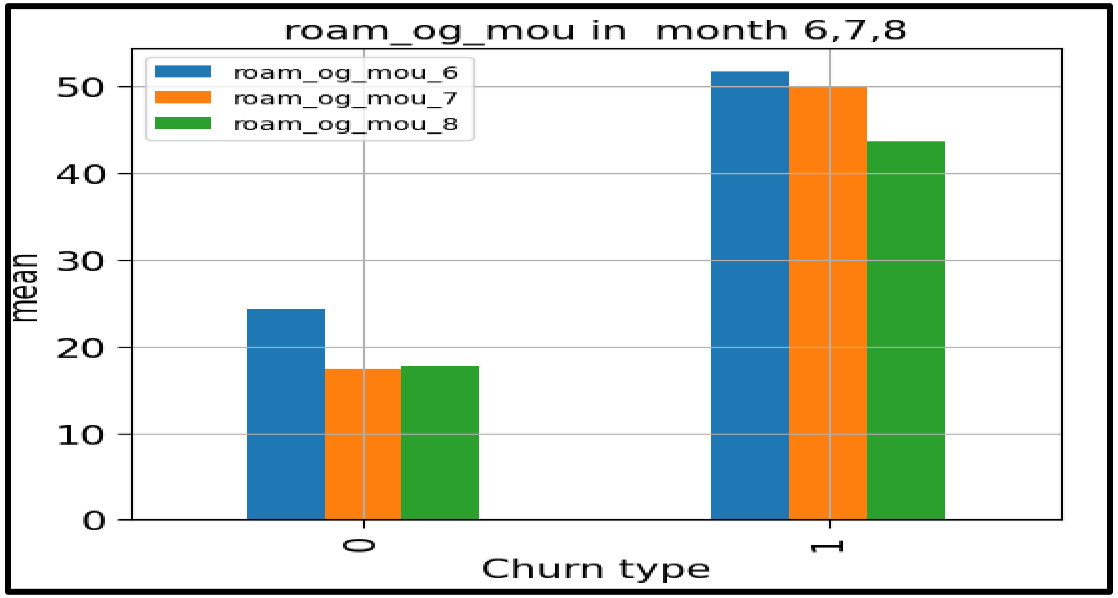


Inference:

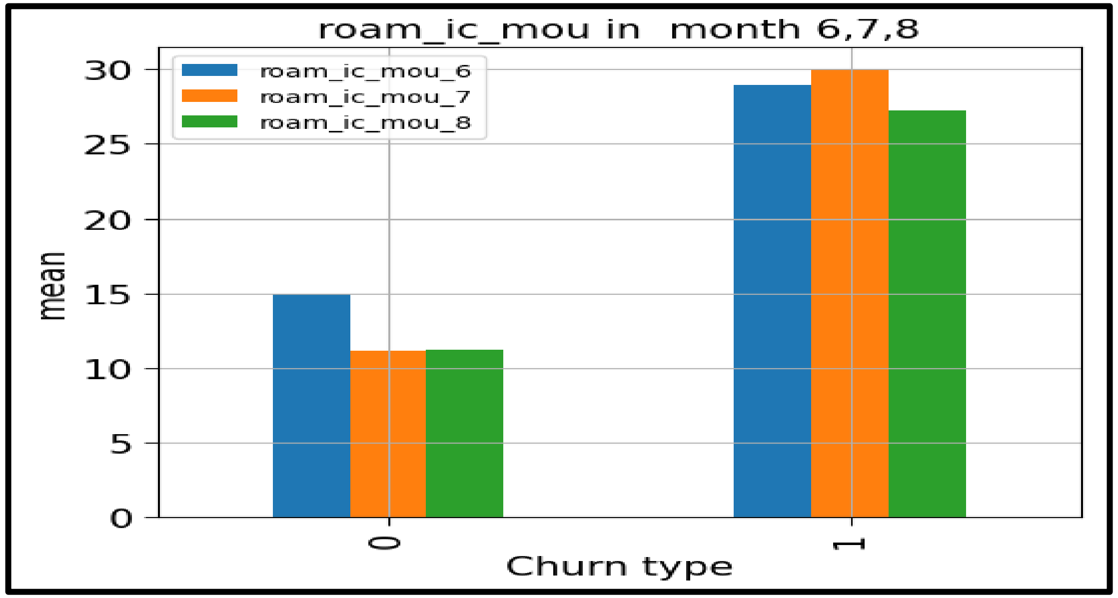
- Total recharge amount: It can observed that total recharge amount for churned customers decreases MoM .
- Average Rev per User : Decrease in Average rev per user is a strong indication of churn.

Step 3: EDA

Distribution of Roaming Min in Outgoing call



Distribution of Roaming Min in incoming call

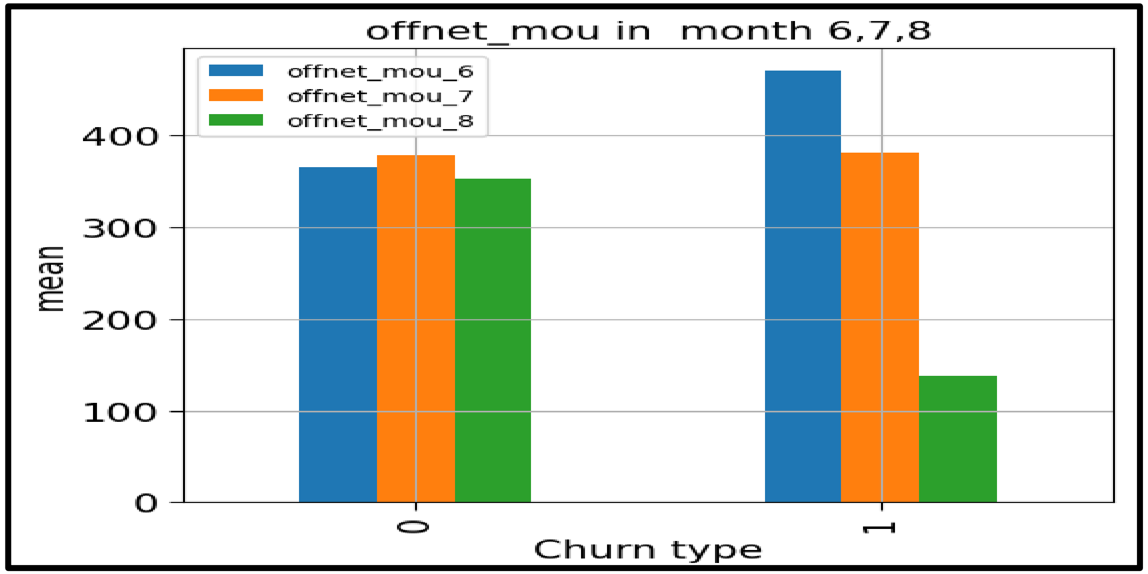


Inference:

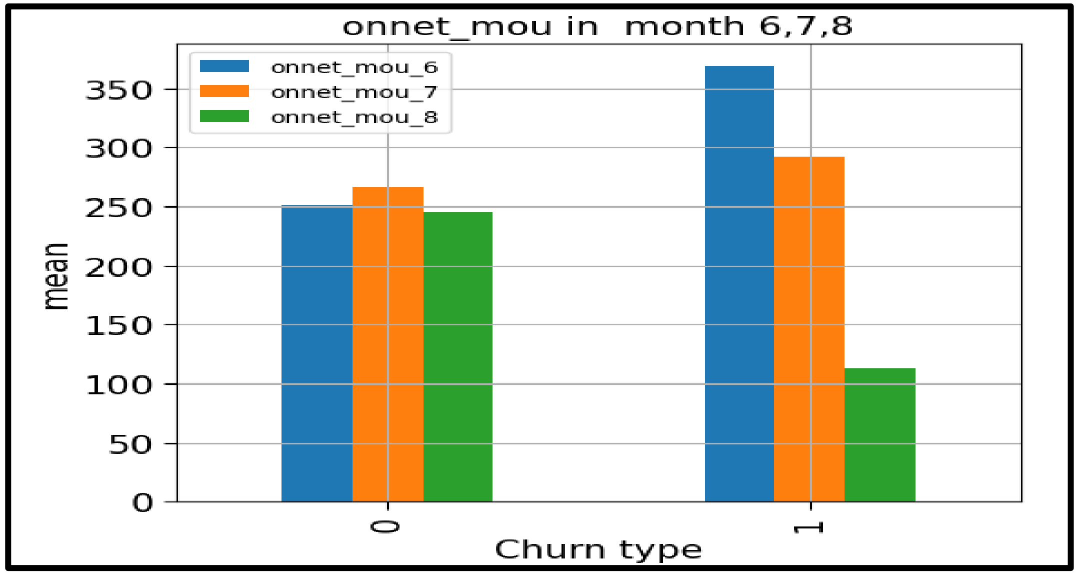
- Roaming Min in Outgoing call : Churn customers are more actively are having longer conversational when compared to average of total customer.
- Roaming Min in incoming call : Churn customers are more actively are having longer conversational when compared to average of total customer.

Step 3: EDA

Distribution of Minutes outside the operator T network



Distribution of Minutes Within the operator T network

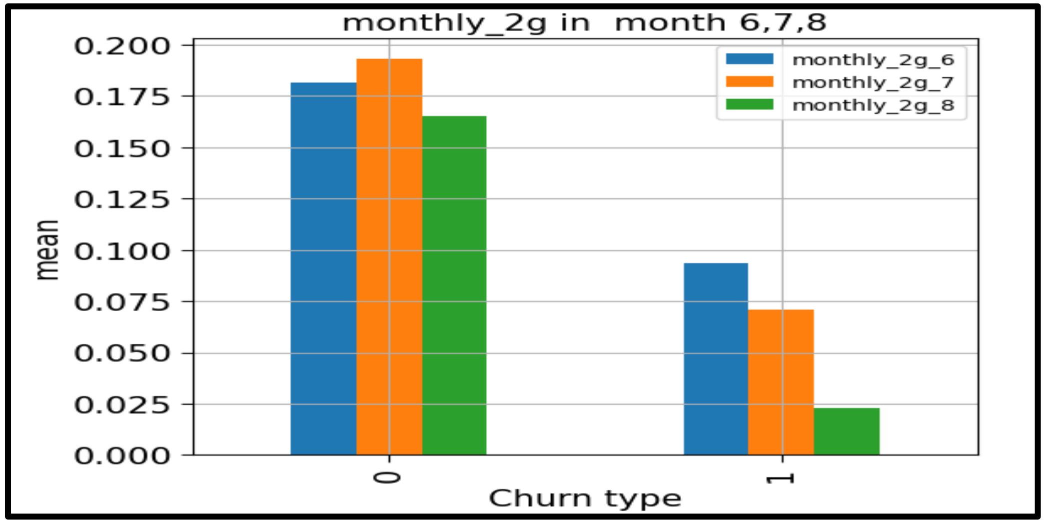


Inference:

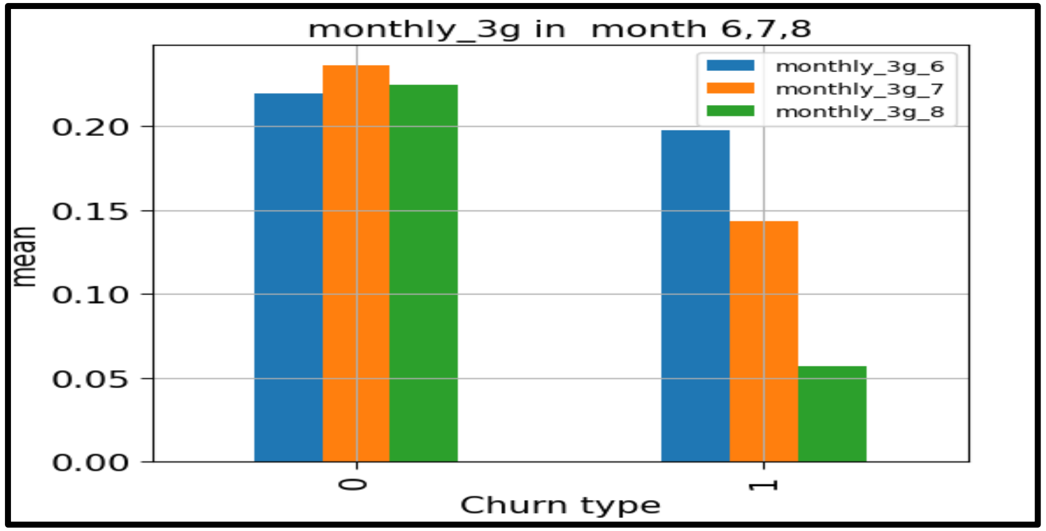
- Minutes outside the operator T network : Decrease in Minutes of usage outside the operator T network is strong indicator of Churn.
- Minutes Within the operator T network: Decrease in Minutes of usage Within the operator T network is strong indicator of Churn.

Step 3: EDA

Distribution of Monthly 2G



Distribution of Monthly 3G



Inference:

- Monthly 2G/3G :2G and 3G usage for churned customers drops in 8th month.

Step 4: Data Split and Resampling

- Splitting Train & Test Sets at 70:30 ratio along with stratify parameter so that the classes are equally divided between in train and test dataset.
- We observed that the churn percentage is just 8.14% of the whole dataset, indicating that there is a class imbalance in the dataset. Over sampling SMOTE is being used to balance out class imbalance in the train dataset.

Step 5: Preprocessing

- Feature scale all the numerical variable using Min Max Scaler.

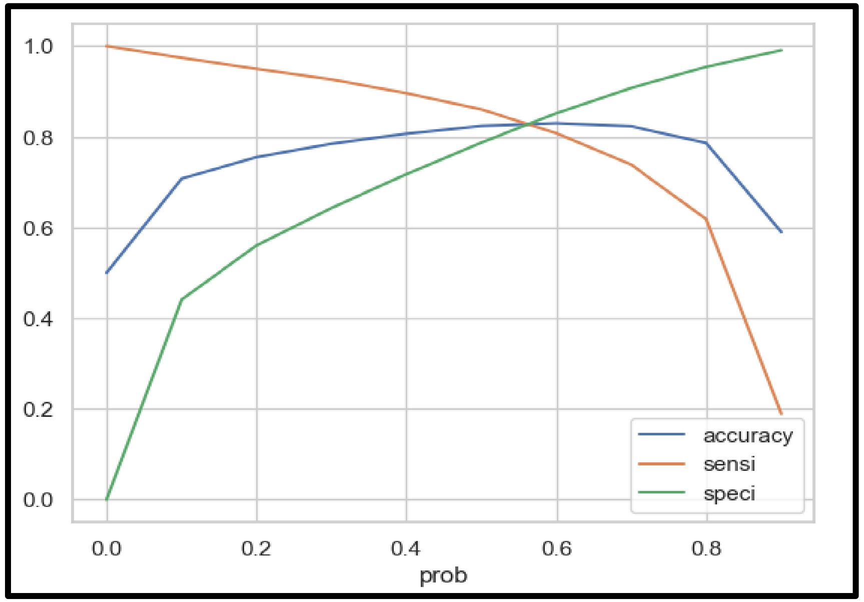
Step 6: Model Selection

- Different Models used for the given Dataset
  - Logistic Regression Using PCA.
  - Random Forest.
  - Decision Tree.
  - Logistic Regression With RFE.

Step 7: Model Evaluation- 0.5 Cutoff

Actual/Predicted	Not_Churn	Churn
Not_Churn	15187	4104
Churn	2696	16595

Metrics	Score
Accuracy	0.82
Sensitivity	0.86
Specificity	0.79
Recall	0.86

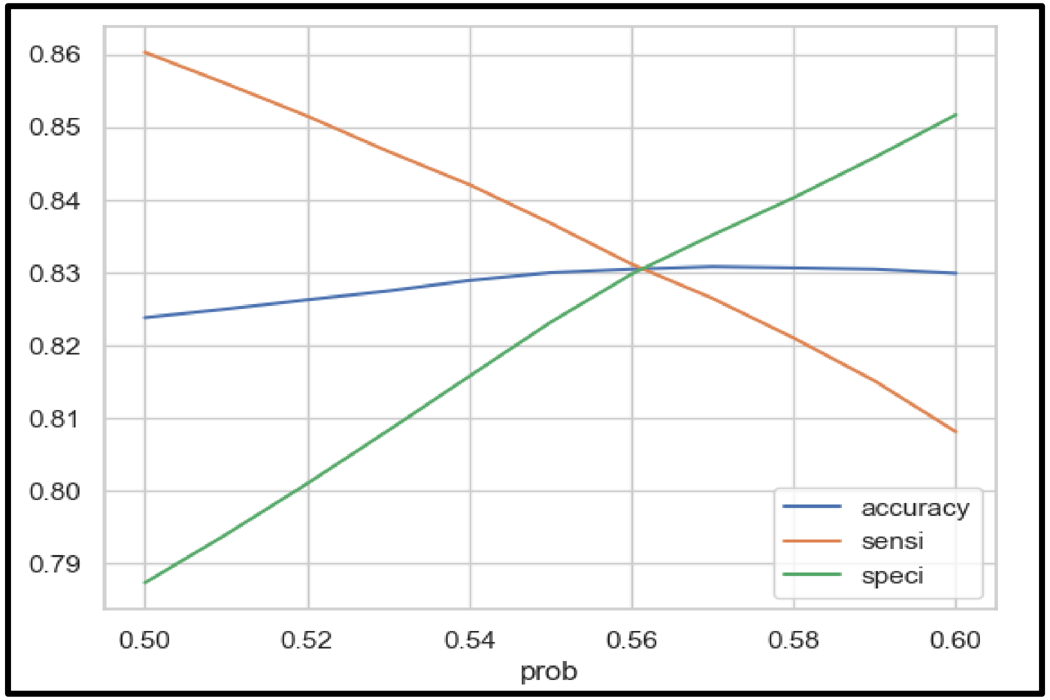


- Based on the above Plot we have considered lead score of 0.56 as the cutoff . Anyone with the lead score of above 0.56 will considered as Converted.

Step 7 :Model Evaluation – 0.56 Cutoff

Actual/Predicted	Not_Churn	Churn
Not_Chrun	16006	3285
Churn	3258	16033

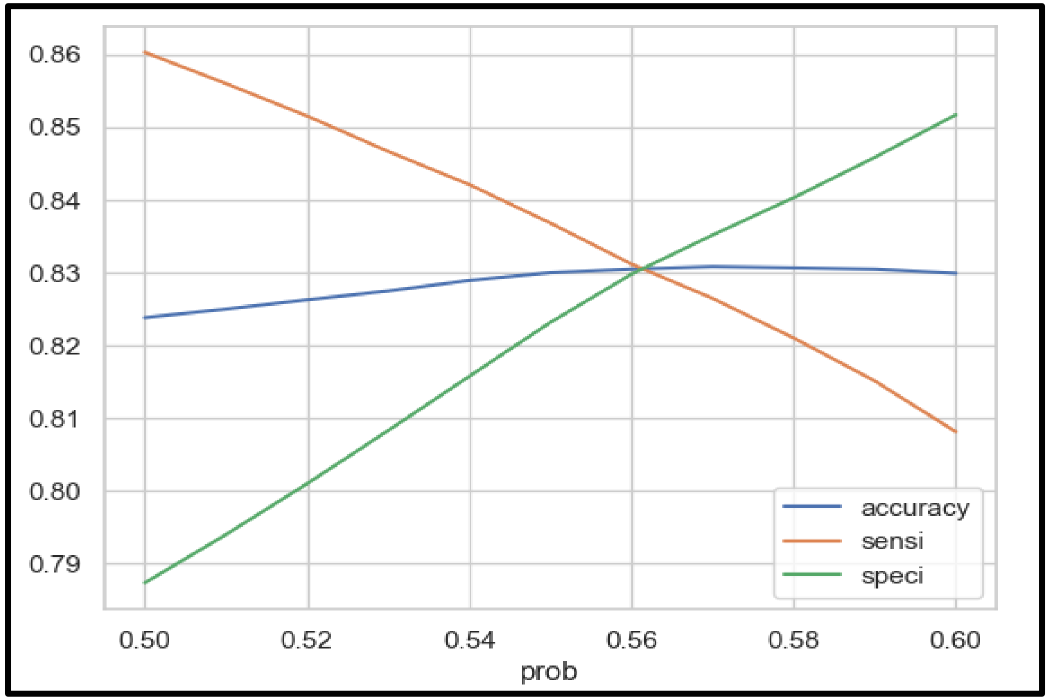
Metrics	Score
Accuracy	0.83
Sensitivity	0.83
Specificity	0.83
Recall	0.83



Step 8: Prediction on Test Dataset

Actual/Predicted	Not_Churn	Churn
Not_Churn	6848	1421
Churn	159	573

Metrics	Score
Accuracy	0.82
Sensitivity	0.78
Specificity	0.83
Recall	0.78



Recommendation

- Based on tenure, Customers who are with less than 4 years are more likely to churn. So, company should concentrate more on that segment by rolling out new schemes to that group.
- Company must provide better 2G/3G area coverage where 2G/3G services are not good, it is a strong indicator of churn.
- It is observed that the recharge amount, volume-based cost drop for 8th month indicates Churn
- Incoming and Outgoing Calls on roaming for 8th month are strong indicators of churn
- Average revenue per user seems to be most important feature in determining churn
- prediction.