IIITB Data Science
Batch C44

Telecom Churn Case Study

By -
Apoorva Nedunoori.
Heena Jain.
Divya Shah.

Problem Statement

- 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.
- There are two models of payment in telecom industry - Postpaid and Prepaid. In postpaid models, when customers want to switch to another operator, they inform existing operators to terminate the service. However, prepaid customers just stop using the services. Thus, churn prediction is usually more critical for prepaid customers.
- Customers who have not utilised any revenue-generating facilities such as mobile internet, outgoing calls, SMS etc. over a given period of time, can be categorised as '**Revenue Based Churn**'.
- Customers who have not done any usage, either incoming or outgoing - in terms of calls, internet etc. over a period of time, can be categorised as '**Usage Based Churn**'.
- In this case study, we will be using Usage based churn definition to identify probable churns.

Business Objective & Data

- The dataset contains customer-level information for a span of four consecutive months - June, July, August and September. The months are encoded as 6, 7, 8 and 9, respectively.
- 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, we have done understood typical customer behaviour during churn by data analysis.
- As we will be working in 4-month window, the first two months are 'good phase', the third month is 'action phase' and last month is 'churn' phase.
- The agenda is to identify key indicators for high probability churn customers.

Overall Approach

1. Data Cleaning
2. Data Analysis - Understanding customer behaviour
3. Defining target variable
4. Feature Engineering
5. Handling Class Imbalance
6. Model Building
7. Model Evaluation
8. Conclusion & recommendation

Problem Solving Methodology

Data Cleaning

- Reading and understanding Data
- Cleaning the data
- Missing value treatment
- Filtering high value customer for 'Good' phase.
- Null value imputation

Data Preparation

- Defining Churn variable
- Identifying class imbalance
- Feature Engineering
- Treating multicollinearity
- Train-test data split

Model Building

- Treating class imbalance with SMOTE technique
- Logistic Regression Model building
- Feature selection using RFE, p-value and VIF.

Model Evaluation

- Plotting ROC curve
- Finding optimal cut-off by TPR-FPR curve.
- Predictions on test set using final LR model

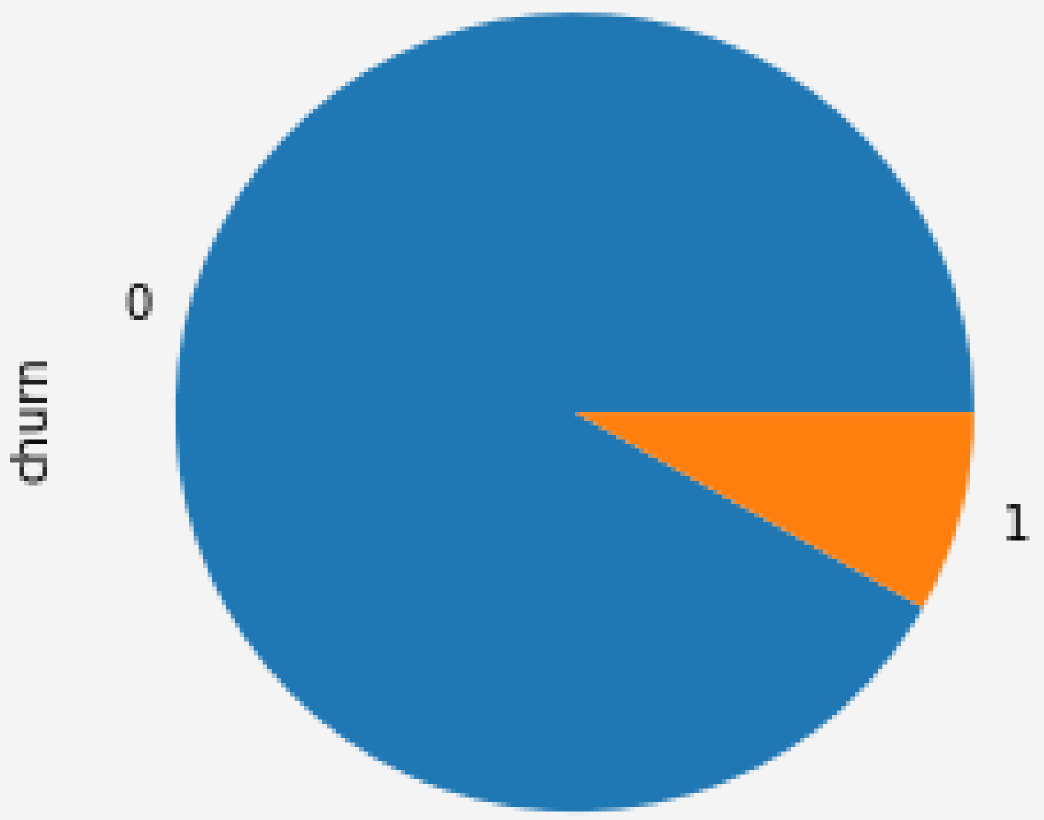
Data Cleaning

- Identified columns important to understand customer behaviour. The null values three columns namely 'date_of_last_rech_data', 'total_rech_data' and 'max_rech_data' helped in finding customers who have not done recharge for telecom service.
- Dropped the columns which were not adding any value to analysis and had more than 50% missing data.
- Dropped date time columns as they were already explained by other columns

High Value Customers

- The 70th quantile value to determine high value customer is 478.0. i.e. the customers who recharge monthly at average price of 478 and above are high value customers.
- It is found that total 30,001 customers are high value customers based on average recharge value for previous two months.

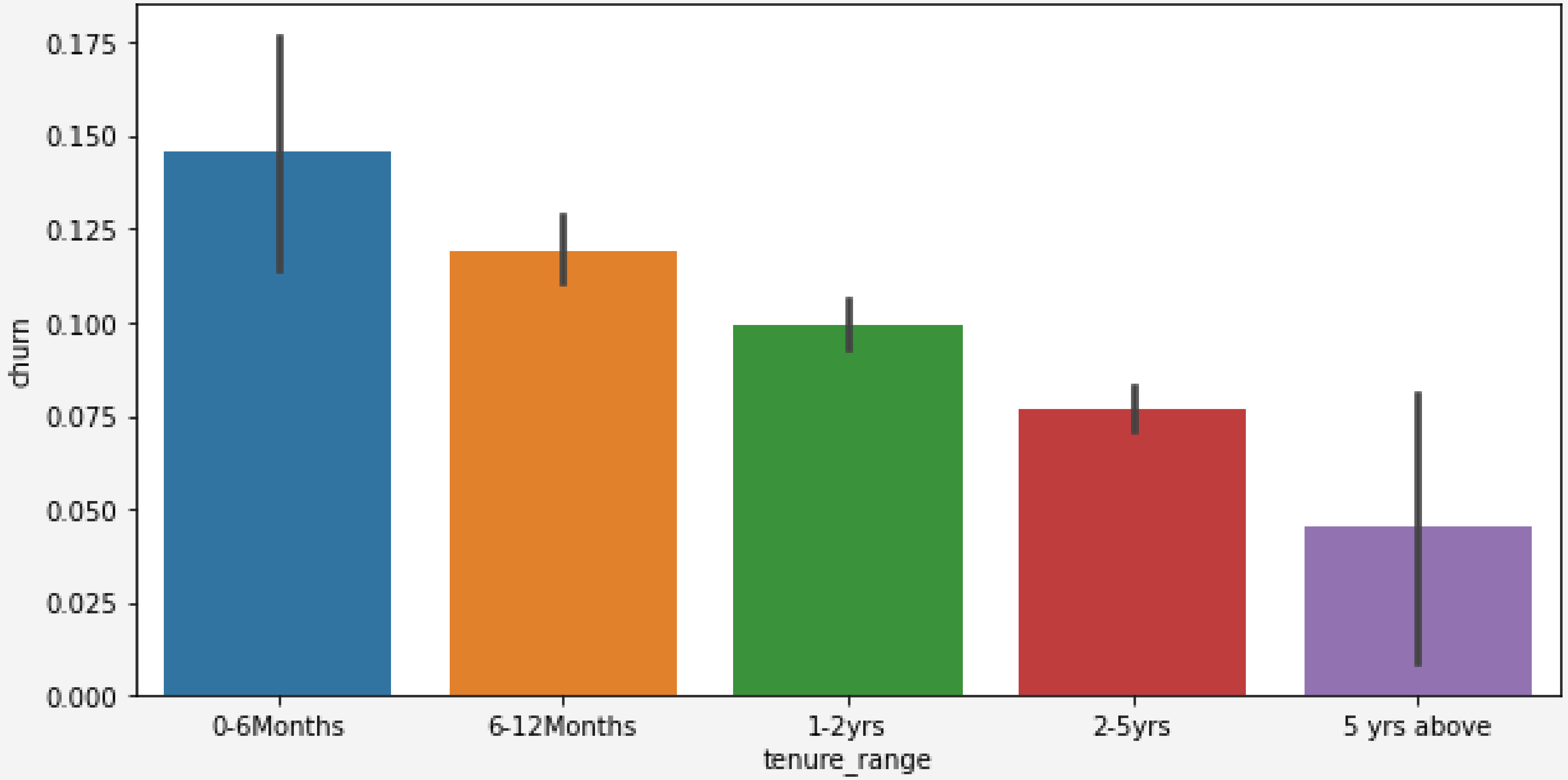
Churn Variable



- Derived 'Churn' variable using four columns namely total_ic_mou_9, total_og_mou_9, vol_2g_mb_9, vol_3g_mb_9.
- It is found that:

1. Non-churn Customers: 91.86%
2. Churn Customers : 8.13%

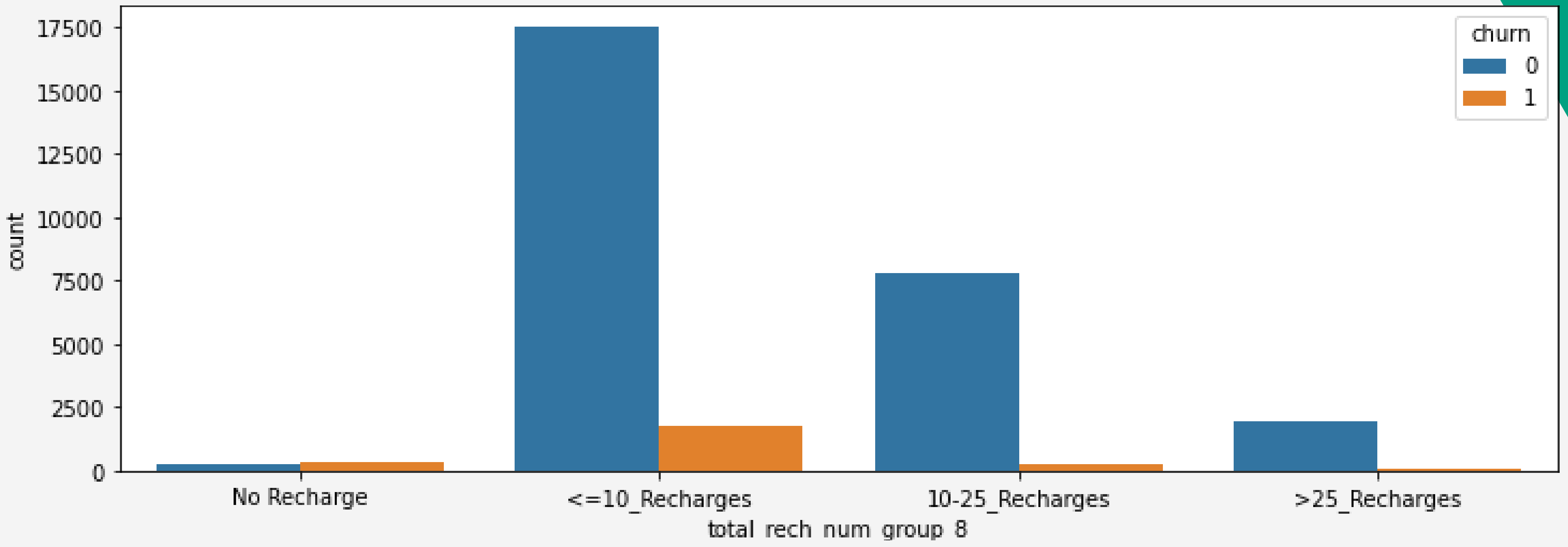
Churn Vs Tenure

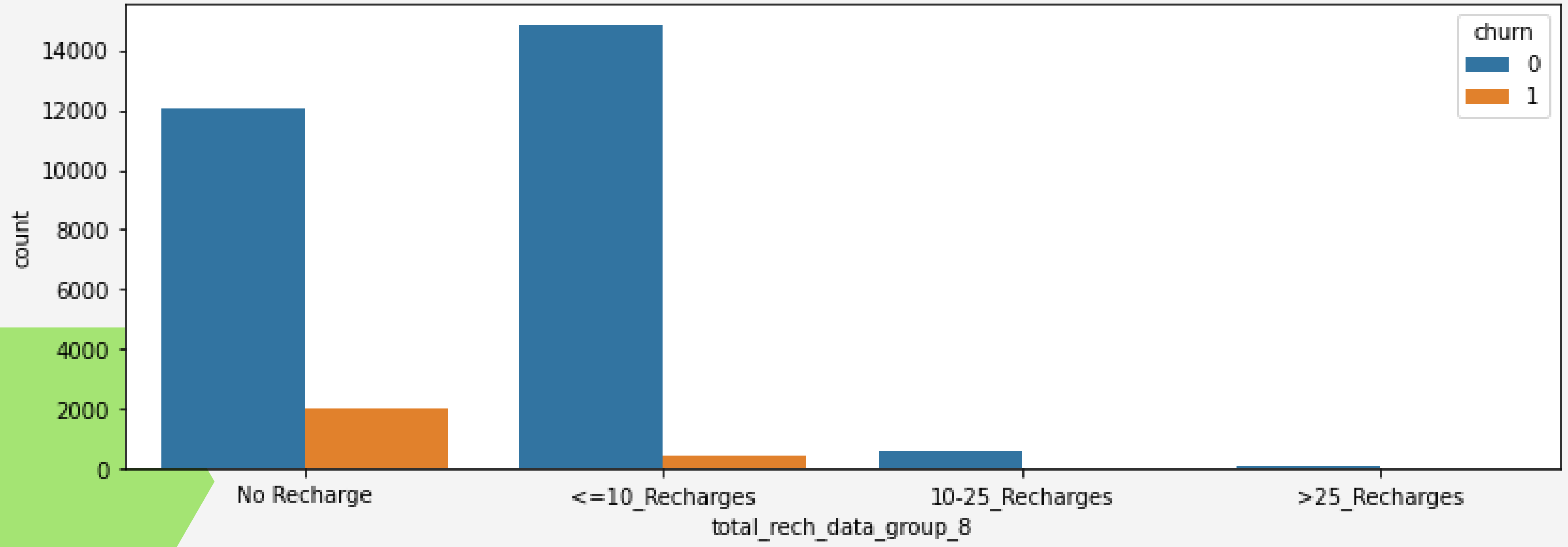


- Derived 'Tenure' variable to understand for how many months customers have been using telecom service.
- It is observed that, there is a maximum probability of high value customer to chun, when he/she is using services from 0 to 6 months.
- Post 6 months of usage probability of churning decreases gradually.

- After checking co-rrelations of all attributes with churn it is found that; Average outgoing calls and calls on roaming for 6th and 7th month are positively correlated with churn.
- However, average revenue per user and number of recharge for 8 th month are negatively correlated with churn.
- After analysing these variables it is found that:

Churn decreases with increase in recharge for numbers.

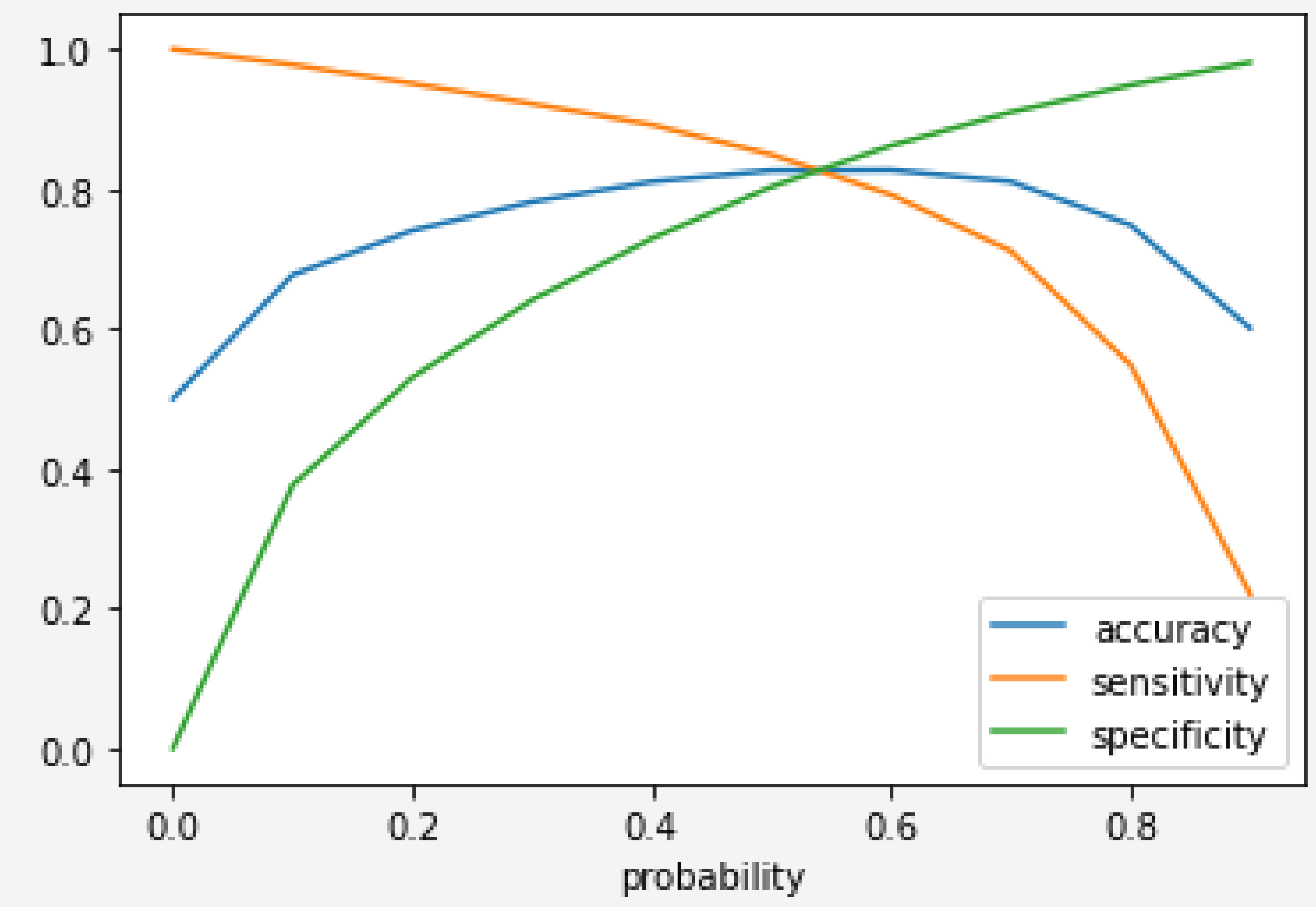


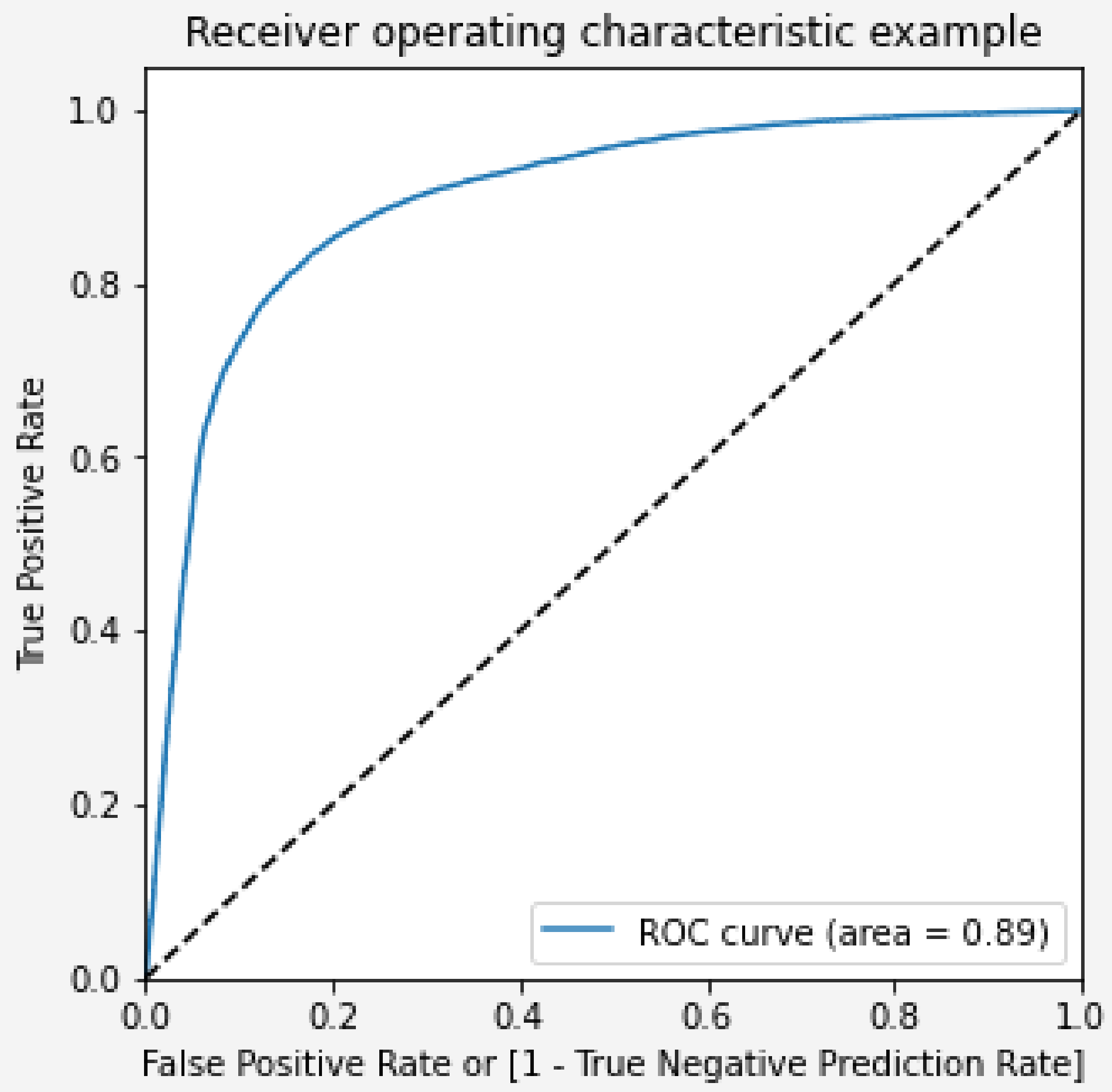


Churn decreases with increase in total recharge rate of data.

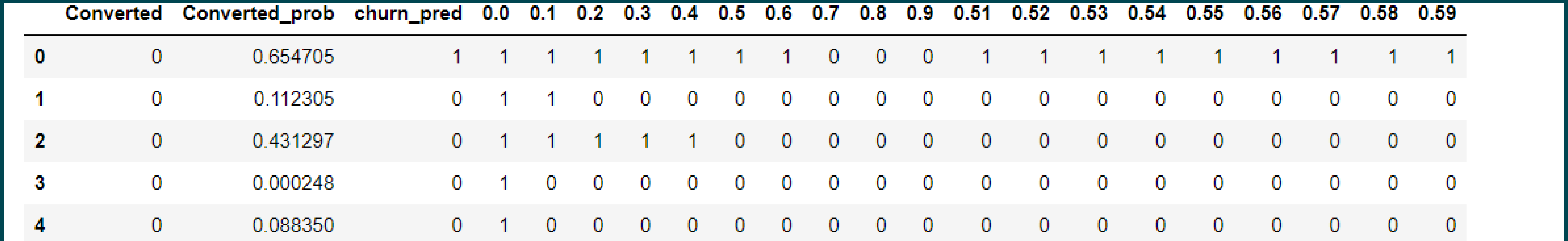
Model Building

- Splitting data into train and test with 70:30 ratio with 127 columns.
- Scaling of numerical columns using Min-Max Scaler.
- Treating class imbalance using SMOTE technique.
- Building Logistic Regression model.
- Choosing top 15 features using RFE and VIF
- Plotting ROC curve and TPR vs FPR curve to choose optimal cut-off point.
- Performed evaluation on test dataset.

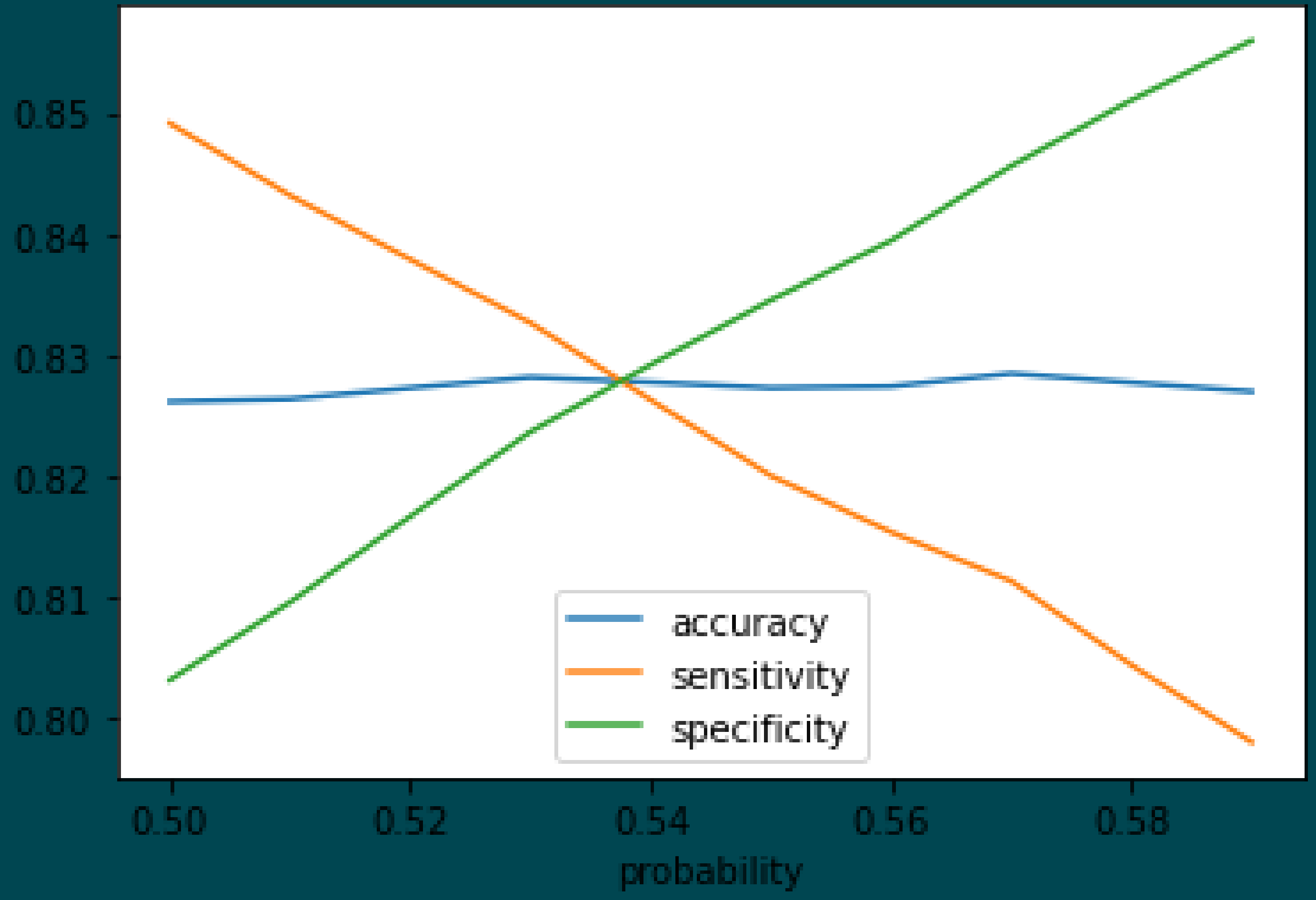




Model Evaluation



	Converted	Converted_prob	churn_pred	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.51	0.52	0.53	0.54	0.55	0.56	0.57	0.58	0.59
0	0	0.654705	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1
1	0	0.112305	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0.431297	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0.000248	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0.088350	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0



- After assuming cut-off point as 0.5 and plotting optimal-cutoff curve we found that the cut off lies between 0.5-0.6.
- Hence, after checking probabilities for all points from 0.51 to 059 we found, 0.54 as the optiomal cut-off point.

Evaluation Metrics of Train Data

- Accuracy = 82.77%
- Sensitivity = 82.62%
- Specificity = 82.91%
- False Positive Rate = 17.07%
- Precision = 82.87%
- True Negative Prediction Rate = 82.68%

Evaluation Metrics of Test Data

- Accuracy = 83.63%
- Sensitivity = 80.26%
- Specificity = 83.94%
- False Positive Rate = 16.05%
- Precision = 31.24%
- True Negative Prediction Rate = 97.90%

- AUC of ROC of test data is 0.87.
- The accuracy of the predicted model is approximately 84.0 %
- The sensitivity of the predicted model is: 80.0 %
- As the model created is based on a sentivity model, i.e. the True positive rate is given more importance as the company needs to focus on churn customers.

Conclusions

- Final logistic regression model is built with 15 features for high value customers.
- A cut-off probability is used to predict the customer will churn or not which is 0.54.
- The customers who have predicted probability of 0.54 and more are most likely to churn.
- The final model has sensitivity of 80% which means 80% of the predicted churns are true churns.
- Local Incoming for Month 8, Average Revenue Per Customer for Month 8 and Max Recharge Amount for Month 8 are the most important predictor variables to predict churn.

Recommendations

- Customers with high positive values in 'loc_ic_t2m_mou-_8' are going to churn. Drop in incoming call usage can signify customer churn. Company needs to revise its plans and offer customers with suitable plans if incoming calls are charged. Incoming calls may reduce for poor network in customers' current location.
- Customers with high positive values in 'last_day_rch_amt_8' are going to churn. Sales team may focus on the customer whose 'last_day_rch_amt_8 >= 35. Drop in last recharge amount in action phase is indicator of churn.
- Customers with high positive values in 'avg_arpu_6_7' are going to churn. Sales team may focus on the customer whose avg_arpu_6_7 > 160. Drop in average revenue per user in action phase is important indicator of customer churn. Revenue is generated by recharges made by the customer, suitable offers can be given to the customer by checking other information such as drop in outgoing calls, drop in data recharge etc.
- Customers with Tenure (age of network) < 6 months are more likely to churn than others. Customers using the telecom operator for more than 1000 days or 3 years are proven to be loyal customers (aon>1000). Churned customers have tenure of 26 months, that means customers using the telecom operator for 2 years or less (800 days or less) might churn.
- Positive value of 'total_rech_num_group_8'> 10 is a strong indicator of churn. About 12 % of the churned customers' total number of recharge dropped by 10 in action phase. Sales team may provide customers with dropping number of recharge with suitable recharge options.
- Positive values of 'max_rech_amt_group_8' and 'avg_rech_amt_data' indicate drop in maximum recharge amount and recharge data in action phase, which in turn indicates high probability of Customer churn.
- High positive values of 'vol_3g_mb' means for churned customers means huge drop in 3g data usage in action phase for customer who churned. Telecom company may need to revise its 3g data plans in order to retain customers using 3g data.