TELECOM CHURN GROUP STUDY

By

Deepak Palsavdiya

Bhavini Bhavesh Heniya

INTRODUCTION

Based on the provided information, the objective is to predict customer churn in the telecommunications industry, specifically focusing on the Prepaid model in the Indian and Southeast Asian markets. Churn prediction is crucial as it helps in retaining high-value customers and reducing revenue leakage. The churn is defined based on usage data during the last phase of customer engagement, and the analysis needs to be done using data from the first three months to predict churn in the ninth month. Here are the steps you can take to conduct this analysis:

1) Data Collection and Preparation
2) Feature Selection
3) Data Analysis
4) Model Selection and Training
5) Churn Prediction

OBJECTIVE

Our goal is to predict customer churn and identify the key factors influencing it.

We will employ multiple machine learning algorithms to build prediction models, assess their accuracy and performance, and determine the best-suited model for our business case.

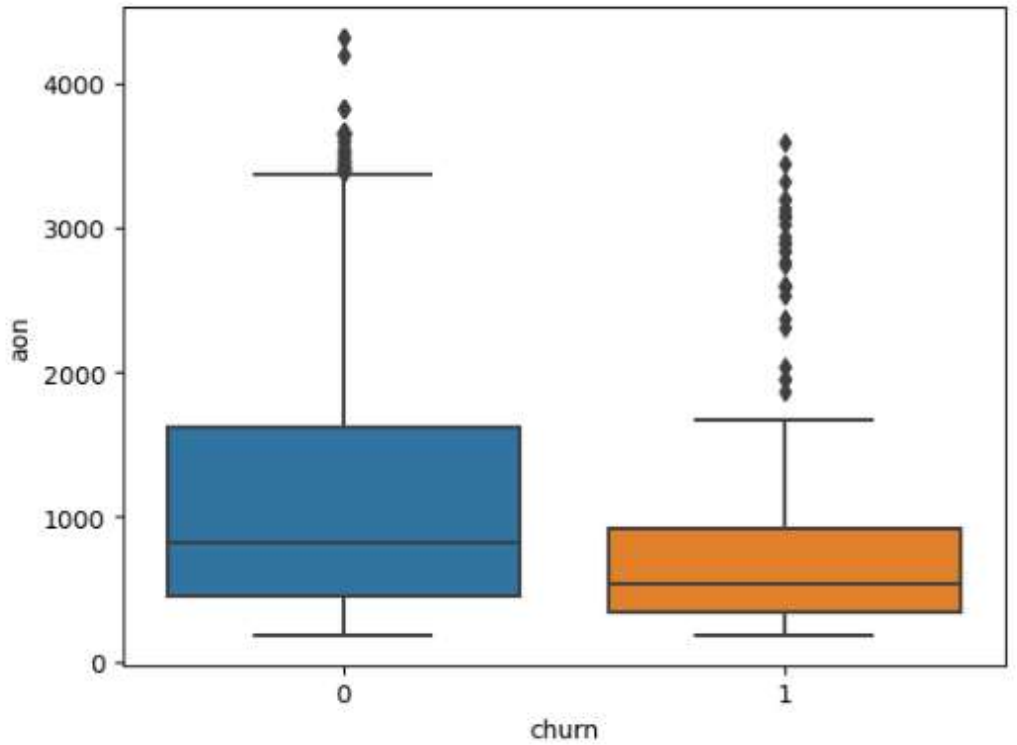
Finally, we will provide an executive summary outlining our findings and recommendations.

STRATEGY

1) Data Understanding: Thoroughly analyzing and grasping the dataset's intricacies.
2) Data Cleansing: Cleaning and refining the dataset to ensure accuracy and consistency.
3) High-Value Customer Identification: Filtering and identifying customers with high value for focused analysis.
4) Target Variable Definition: Clearly defining the variable that indicates customer churn.
5) Data Preparation: Preparing the data for analysis and modeling.
6) Data Modeling: Generating dummy variables for categorical features. Splitting the data into training and test sets. Addressing class imbalance in the data.
7) Logistic Regression: Utilizing Recursive Feature Elimination (RFE) technique for variable selection. Building the logistic regression model. Evaluating the model using metrics such as accuracy, specificity, and sensitivity. Making predictions on the test data. Tuning hyperparameters for optimal performance.
8) Model Selection: Evaluating various models and selecting the most suitable one for the business case.

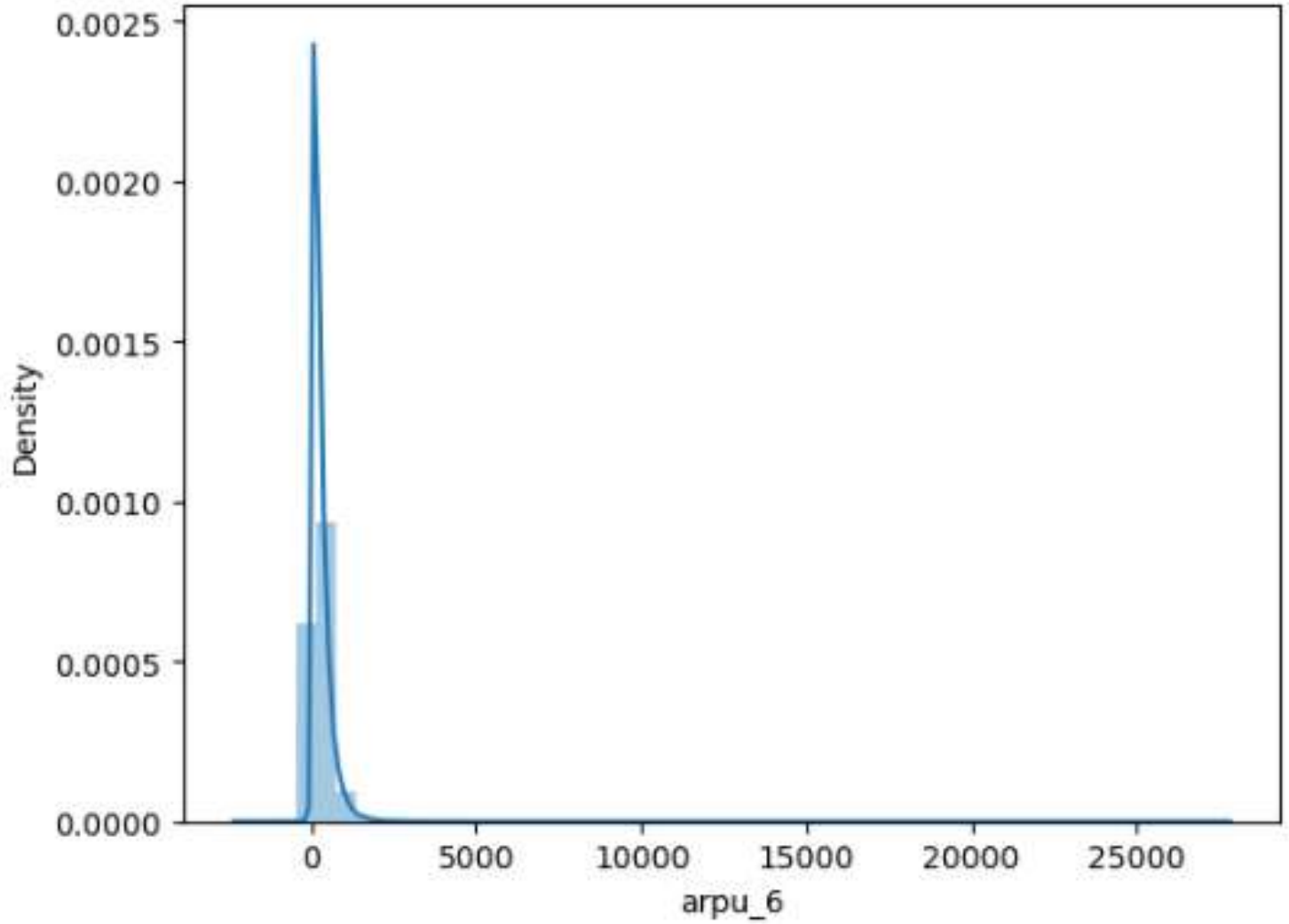
DATA UNDERSTANDING

- The dataset comprises 99,999 rows and 226 columns.
- Handling missing values is necessary for the given input dataset.
- The variable 'Churn' signifies whether a customer has churned or not. Our objective is to develop predictive models for churn prediction.
- Outliers are identified within the dataset.
- There is an imbalance in the 'Churn' variable, indicating a disparity in class distribution.
- We will address missing values by dropping columns or rows with significant missing data.



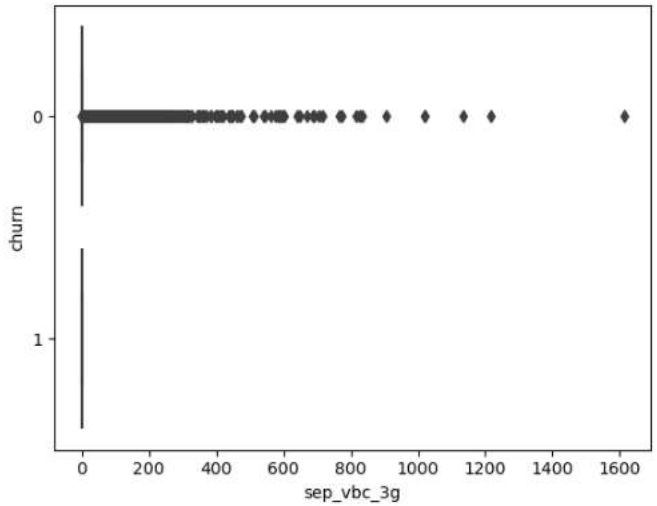
EXPLORATORY DATA ANALYSIS

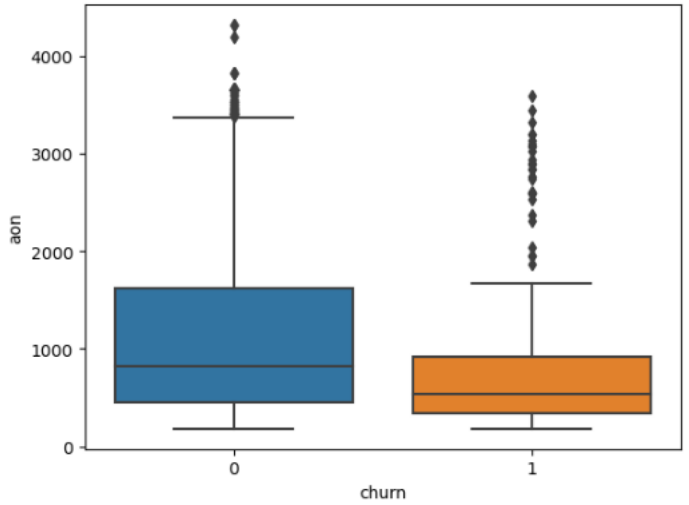
UNIVARIATE

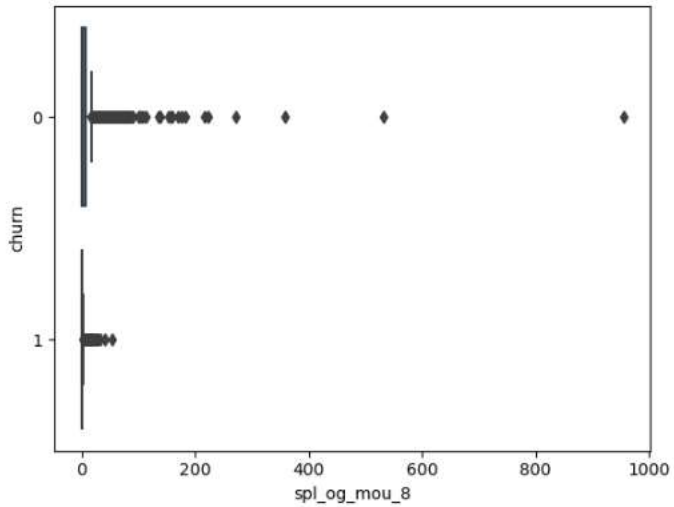


EXPLORATORY DATA ANALYSIS

BIVARIATE







EXPLORATORY DATA ANALYSIS - SUMMARY

#EDA - Summary

Calls Revenue(3 columns):

#Invalid Values : Having minimum values as negatives, indicating some customers are making loss to the company. These columns are either invalid or not adding value to our prediction, can be dropped from the dataset.

#Standardise: Revenue columns can be rounded to 2 decimal places.

#Minutes of usage(60+ columns):

#Usage minutes is generally 0 except for few outliers, for below variables:

#Roaming Incoming ISD Incoming Special Incoming Others STD incoming T2F STD outgoing T2F Outgoing Others ISD Outgoing Local Outgoing T2C (Customer care calls)

#Most of the columns have outliers.

EXPLORATORY DATA ANALYSIS - SUMMARY

Aggregating Columns based on Incoming and Outgoing, or Aggregating based on Each Type of Incoming Calls and Outgoing Calls and looking at the metrics will give a better understanding of the data.

#Recharge (12 Numeric + 3 Date columns)

#Data Type Conversion:

#Data in numeric columns are integers, so can be converted to int type.

#Date columns need to be converted to date type

#Data 2G And 3G(22 Columns)

#Most of the columns have median as O and have outliers

#vbc_3g columns need column renaming as it needs month to be encoded to its number.

#Standardise: Columns can be rounded off to 2 decimal places.

EXPLORATORY DATA ANALYSIS - SUMMARY

#Age on Network (1 Column)

#Feature can be derived from AON column.

#Churn (Dependent Variable)

#There exists a Class Imbalance in the dataset, where actual churn customers are only 6% of the dataset.

#Reviewing the Dropped Columns:

#More columns will be lost because of dropping missing value columns, while it can be handled to be imputed by considered 0 as missing values follow a pattern where Calls only users have blanks for Data related columns and the vice versa.

#Feature Engineering - Thoughts

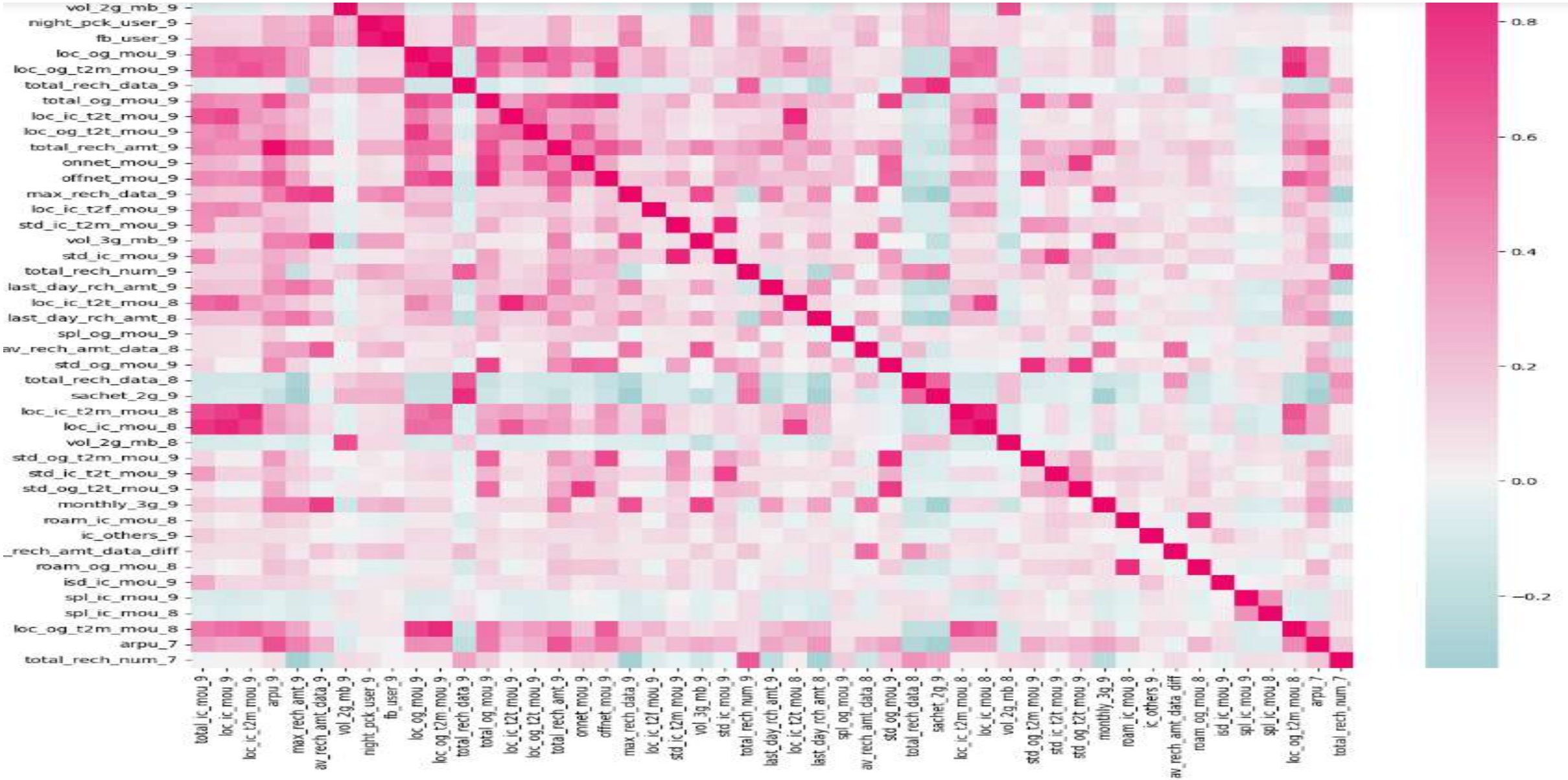
#Derive no. of years the customer is using network from AON

#Derive fields to indicate the type of user the customer is: Uses Both Calls and Data, Only Calls, Only Data, Only Incoming calls, Only Outgoing calls, etc.

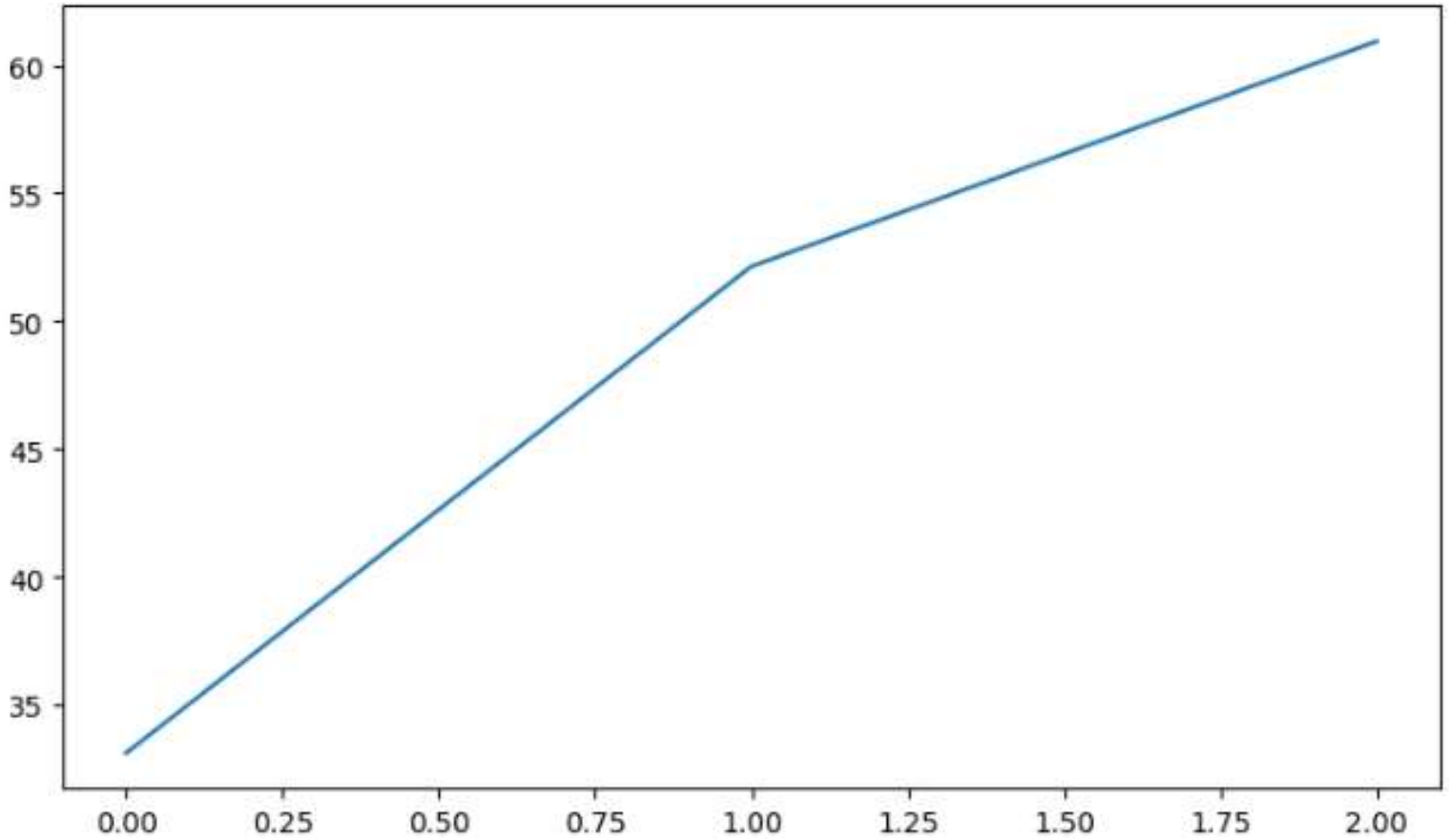
#Bin the customers into different segments based on Service usage, Recharge amount, Usage/Recharge pattern.

#Calls to Customer Care is a Key indicator that customer is not happy with the services, derive columns like time over call

HEAT MAP


vol_2g_mb_9
night_pck_user_9
fb_user_9
loc_og_mou_9
loc_og_t2m_mou_9
total_rech_data_9
total_og_mou_9
loc_ic_t2t_mou_9
loc_og_t2t_mou_9
total_rech_amt_9
onnet_mou_9
offnet_mou_9
max_rech_data_9
loc_ic_t2f_mou_9
std_ic_t2m_mou_9
vol_3g_mb_9
std_ic_mou_9
total_rech_num_9
last_day_rch_amt_9
loc_ic_t2t_mou_8
last_day_rch_amt_8
spl_og_mou_9
av_rech_amt_data_8
std_og_mou_9
total_rech_data_8
sachet_2g_9
loc_ic_t2m_mou_8
loc_ic_mou_8
vol_2g_mb_8
std_og_t2m_mou_9
std_ic_t2t_mou_9
std_og_t2t_mou_9
monthly_3g_9
roam_ic_mou_8
ic_others_9
_rech_amt_data_diff
roam_og_mou_8
isd_ic_mou_9
spl_ic_mou_9
spl_ic_mou_8
loc_og_t2m_mou_8
arpu_7
total_rech_num_7
total_ic_mou_9
loc_ic_mou_9
loc_ic_t2m_mou_9
arpu_9
max_rech_amt_9
av_rech_amt_data_9
vol_2g_mb_9
night_pck_user_9
fb_user_9
loc_og_mou_9
loc_og_t2m_mou_9
total_rech_data_9
total_og_mou_9
loc_ic_t2t_mou_9
loc_og_t2t_mou_9
total_rech_amt_9
onnet_mou_9
offnet_mou_9
max_rech_data_9
loc_ic_t2f_mou_9
std_ic_t2m_mou_9
vol_3g_mb_9
std_ic_mou_9
total_rech_num_9
last_day_rch_amt_9
loc_ic_t2t_mou_8
last_day_rch_amt_8
spl_og_mou_9
av_rech_amt_data_8
std_og_mou_9
total_rech_data_8
sachet_2g_9
loc_ic_t2m_mou_8
loc_ic_mou_8
vol_2g_mb_8
std_og_t2m_mou_9
std_ic_t2t_mou_9
std_og_t2t_mou_9
monthly_3g_9
roam_ic_mou_8
ic_others_9
av_rech_amt_data_diff
roam_og_mou_8
isd_ic_mou_9
spl_ic_mou_9
spl_ic_mou_8
loc_og_t2m_mou_8
arpu_7
total_rech_num_7
0.8
0.6
0.4
0.2
0.0
−0.2

MODEL ANALYSIS - PCA



HYPER PARAMETER TUNING

Logistic Regression

Best Score: 0.9985
Sensitivity: 0.9
Specificity: 0.97
AUC: 0.985

Using Random Forest

Best Score: 0.995
Sensitivity: 0.9
Specificity: 1.0
AUC: 1.0

Looks like random forest is overfitting and the sensitivity is very low. So going with logistic and PCA

CONCLUSION

The primary churn indicators identified from the analysis include total incoming call minutes of usage in the action phase (total_ic_mou_8), the difference in total recharge amounts (total_rech_amt_diff), total outgoing call minutes of usage in the action phase (total_og_mou_8), average revenue per user (arpu), roaming incoming call minutes of usage in the action phase (roam_ic_mou_8), roaming outgoing call minutes of usage in the action phase (roam_og_mou_8), STD incoming call minutes of usage in the action phase (std_ic_mou_8), STD outgoing call minutes of usage in the action phase (std_og_mou_8), and average recharge amount during the action phase (av_rech_amt_data_8).

To reduce churn, the following steps can be taken:

1. Customized Discounts: Offer special discounts tailored to individual customers based on their usage patterns.

2. Enhanced Internet Services: Provide additional internet services upon recharge to enhance customer satisfaction and value.

3. Customer Engagement: Engage in personalized conversations with customers to understand their needs and preferences better. Fulfilling customer desires can enhance loyalty.

4. Lower Data Tariffs: Reduce tariffs on data usage to make it more affordable for customers, addressing a significant aspect of their telecom needs.

5. Improved Network Coverage: Enhance 2G coverage, especially in areas where 3G is unavailable. Additionally, invest in expanding the 3G network to regions lacking current 3G coverage. Improved network coverage ensures better service quality and customer satisfaction, reducing the likelihood of churn.

THANK YOU

Deepak Palsavdiya

Bhavini Bhavesh Heniya