Telecom Churn Case Study

Sheetal Saxena

Viren Pawar

Mohan Prasad

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.

- 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.**

Objectives

- 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.
- After identifying important predictors, display them visually – we will use plots, summary tables etc. - whatever you think best conveys the importance of features.
- **recommend strategies to manage customer churn** based on your observations.

99999

Data Understanding

- The provided *Telecom* dataset has around 9000 data points. This dataset consists of various attributes such as Churn, Age on Net,average revenue per user , service packs, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.

- The target variable, in this case study, is the column ‘Churn’ which tells whether customer will leave this network.

Case Study Approach

I. Data Preparation, Cleaning and EDA
II. Test-train Split and Scaling
III. Model Building
IV. Model Evaluation
V. Prediction on data sets
VI. Conclusion
VII. Recommendations

Data Preparation, Cleaning & Exploratory Data Analysis

Step 1: Importing Data

Step 2: Inspecting the Dataframe

Step 3: Data Cleaning

Step 4: EDA

Step 5: Data Preparation

Importing and Inspecting the data

- There are 99999 rows and 226 columns in the dataframe.
- There are columns which give the details about the incoming call, On net usage , offnet usage, average recharge amount etc .

Data Cleaning

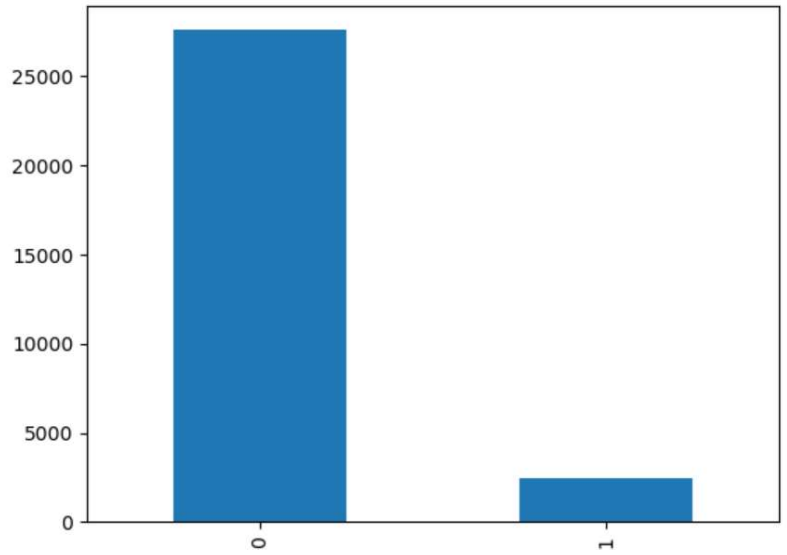
- Missing Values Analysis (Percentage Wise)

arpu_3g_6	74.846748
arpu_3g_7	74.428744
arpu_3g_8	73.660737
arpu_3g_9	74.077741
arpu_2g_6	74.846748
arpu_2g_7	74.428744
arpu_2g_8	73.660737
arpu_2g_9	74.077741
night_pck_user_6	74.846748
night_pck_user_7	74.428744
night_pck_user_8	73.660737
night_pck_user_9	74.077741
fb_user_6	74.846748
fb_user_7	74.428744
fb_user_8	73.660737
fb_user_9	74.077741

Data Cleaning

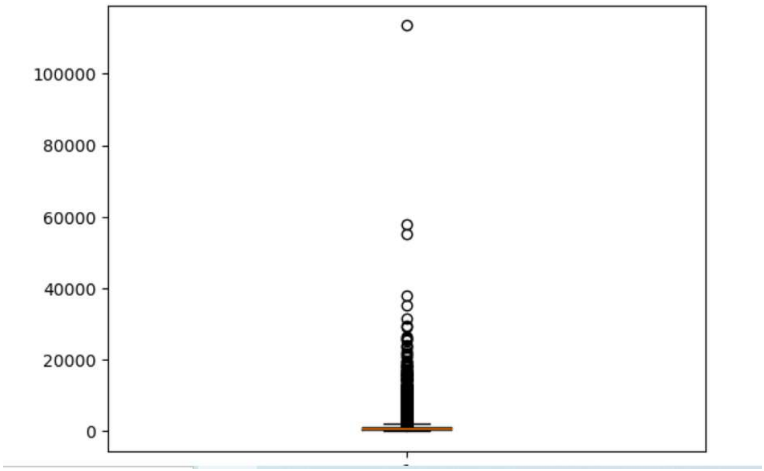
- After analyzing the null value percentage of all columns present in telecom Datasethave 75% of null values
- We have dropped the columns having Missing Values.
- We have imputted onnet, offnet, roam_og, loc_og, std_og, isd_og, spl_og, og_others as 0 as total_outgoing minutes of usage is 0 for customer
- Also, imputed the incoming calls columns like roam_ic, loc_ic, std_ic, spl_ic, isd_ic, ic_others as 0 as total_ic_mou is 0 for customer
- We have filtered out **high-value customers**
- Calculated the average recharges for 6 and 7^{th} month.

Univariate Analysis



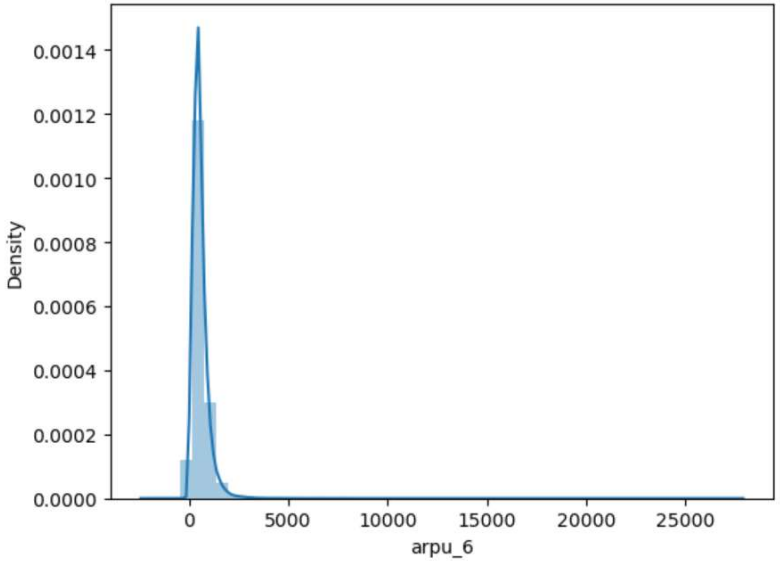
- As seen in the graph maximum is for 0 and less is for 1

Univariate Analysis



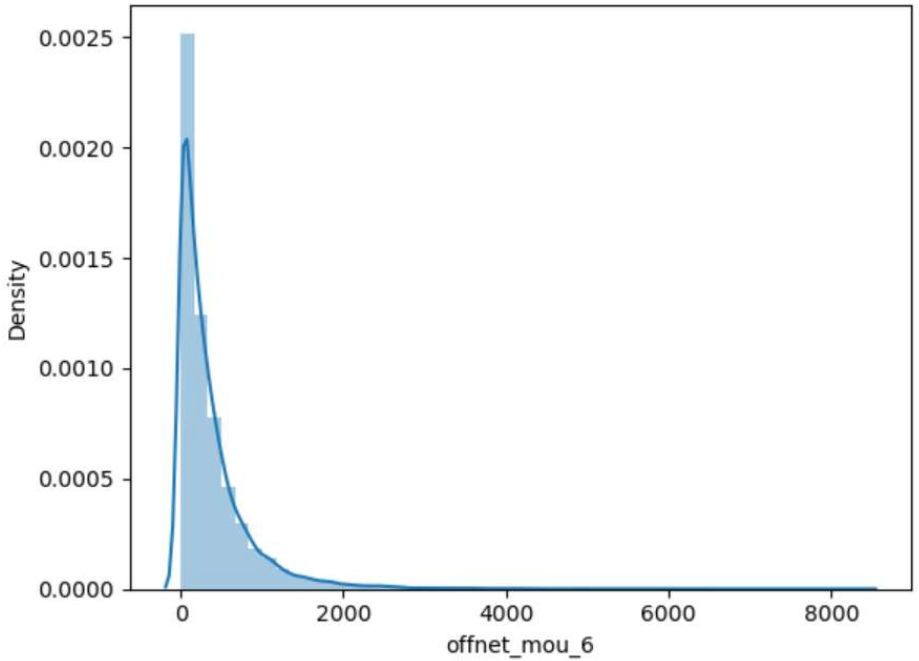
- This graph was plotted for the total rech amount for 6th month
- As seen in the graph there are many outliers

Average revenue per user



- As seen in the graph the maximum average revenue per user is 27731.

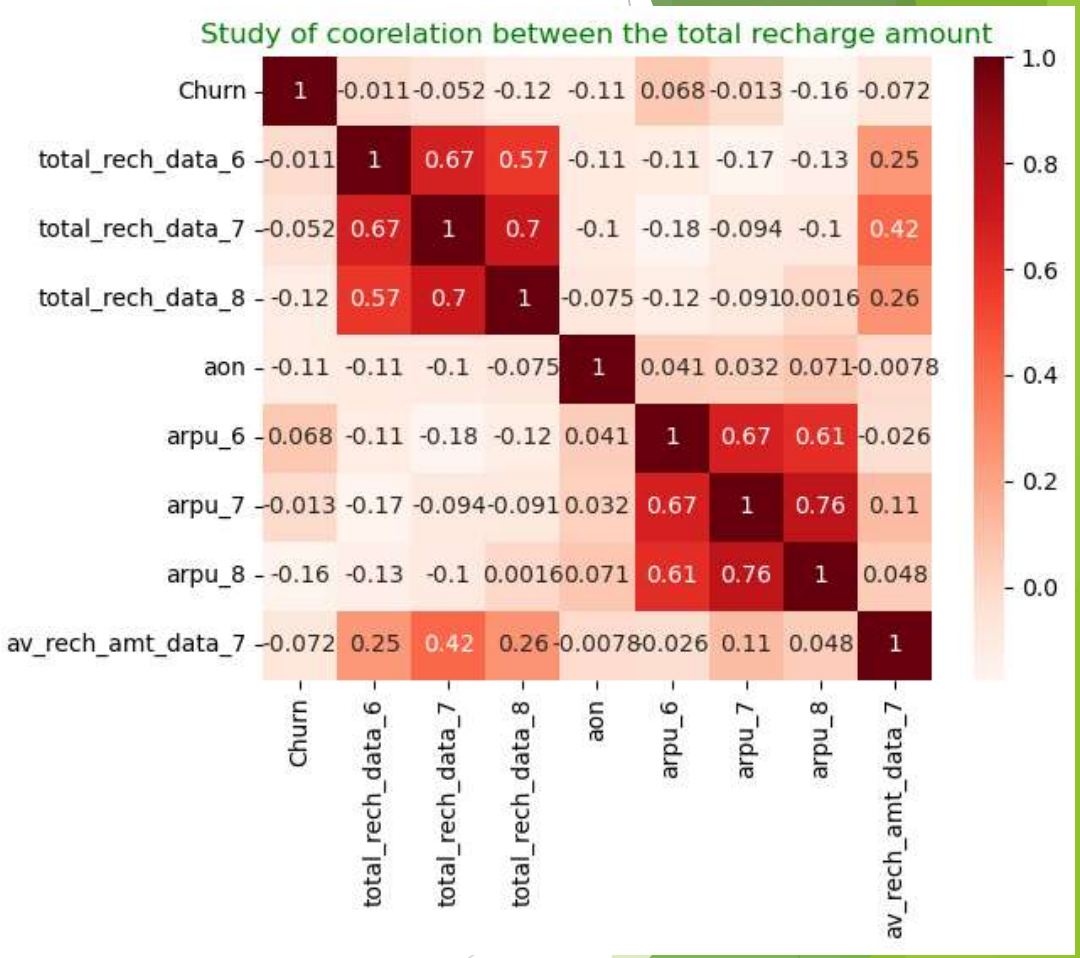
Offnet Minutes of usage



- As seen in the graph the offnet minutes of usage for 6th month is 8362

Multivariate Analysis

- There is a positive correlation with Average unit per user and average amount recharge
- -Churn has a positive correlation with average revenue per user for 6th month.



Test-train Split and Scaling

Test-train Split and Scaling

- Step 1: Test-Train Split

  The dataset is split into Training and Testing Data in the ratio of 70:30 .

- Step 2: Feature Scaling
  - The Feature Scaling is done by using *StandardScalar* function.
  - For training data, *fit_transform* function is used.
  - For testing data, *transform* function is used.
  - This will be different for different models.

Model Building

We have prepared various models based on following algorithms

- Principle component Analysis and Regression
- Logistic Regression with RFE and VIF
- Decision Tree
- ADA Boosting with Decision Tree
- Random Forest

Model Building

As per the algorithms , we have taken the features and built the model

Model Evaluation parameters for all models

Step 1: Accuracy

Step 2: Sensitivity and Specificity

Step 3: Precision and Recall

Determining the precision and recall for various models test datasets

MethodologyTest -	Precision	Test - Recall
--PCA with regression	37	71.33
--Logistic Regression	40.7	71.33
--Decision Tree	73	46
--ADA Boosting with DT	69.1	52.3
--Random Forests	74	50.0

Conclusion

- We see that almost on all models the values are coming very similar to each other and more often than not there is a trade off between precision and recall
- Since both the metrics are important, we feel that going ahead with Random forests or Decision Trees

Recommendations

As per our analysis, following factors would affect the Churn :

- Total Incoming Minutes of usage in the August
- Total Incoming Minutes of usage in the July
- 2G data pack
- Roaming
- Sachet 2g

Thankyou !!