TELECOM CHURN CASE STUDY

- Visnhu Sreekar
- Md.Shahbaz Shafiq Qureshi
- Sayali

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, you will analyze 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 key predictors, visually present them using a combination of plots, summary tables, or any suitable means to effectively convey the significance of these features.

Based on your observations, suggest tactics for mitigating customer churn.

DATA UNDERSTANDING

- The telecom dataset you've been provided with comprises approximately 9000 data points and encompasses various attributes like churn, Age on Net, average revenue per user, service packs, and more. These attributes may have varying degrees of relevance in determining whether a lead will ultimately convert or not.

- In this case study, the target variable is the 'Churn' column, which indicates whether a customer will depart from this network.

CASE STUDY APPROACH

- DATA PREPARATION , CLEANING & EDA
- TEST-TRAIN SPLIT AND SCALING
- MODEL BUILDING
- MODEL EVALUATION
- PREDICTION ON DATA SETS
- CONCLUSION
- RECOMMENDATI
ONS

DATA PREPARATION , CLEANING & EDA

- Importing Data
- Analyzing the Data frame
- Data Cleaning
- EDA
- Data preparation

HANDLING MISSING VALUES

- There are various columns which have approximately 75% NAN values
- We observe that over 74% of the values pertaining to recharge-related data are absent or missing.

arpu_3g_7	74.428744
total_rech_data_9	74.077741
count_rech_3g_9	74.077741
fb_user_9	74.077741
max_rech_data_9	74.077741
arpu_3g_9	74.077741
date_of_last_rech_data_9	74.077741
night_pck_user_9	74.077741
arpu_2g_9	74.077741
count_rech_2g_9	74.077741
av_rech_amt_data_9	74.077741
total_rech_data_8	73.660737
arpu_3g_8	73.660737
fb_user_8	73.660737
night_pck_user_8	73.660737
av_rech_amt_data_8	73.660737
max_rech_data_8	73.660737
count_rech_3g_8	73.660737

arpu_3g_6	74.846748
night_pck_user_6	74.846748
total_rech_data_6	74.846748
arpu_2g_6	74.846748
max_rech_data_6	74.846748
fb_user_6	74.846748
av_rech_amt_data_6	74.846748
date_of_last_rech_data_6	74.846748
count_rech_2g_6	74.846748
count_rech_3g_6	74.846748
date_of_last_rech_data_7	74.428744
total_rech_data_7	74.428744
fb_user_7	74.428744
max_rech_data_7	74.428744
night_pck_user_7	74.428744
count_rech_2g_7	74.428744
av_rech_amt_data_7	74.428744
arpu_2g_7	74.428744
count_rech_3g_7	74.428744
arpu_2g_8	73.660737
count_rech_2g_8	73.660737
date_of_last_rech_data_8	73.660737

DATA CLEANING

FOR CUSTOMERS WITH A TOTAL OUTGOING MINUTES OF USAGE (TOTAL_OG_MOU) EQUAL TO 0, WE WILL IMPUTE THE VALUES OF ONNET, OFFNET, ROAM_OG, LOC_OG, STD_OG, ISD_OG, SPL_OG, AND OG_OTHERS AS 0.

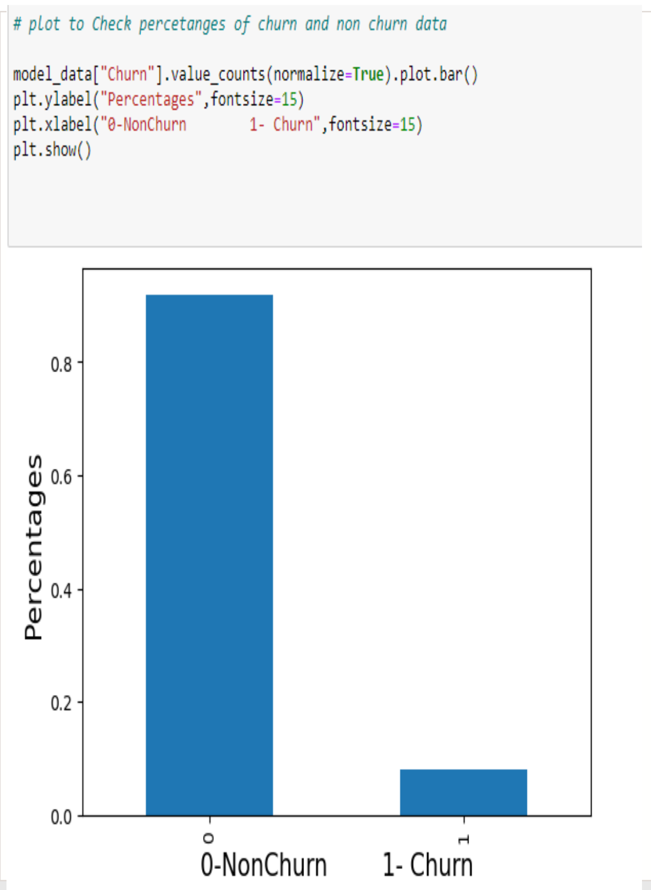
ALSO IMPUTED ROAM_IC, LOC_IC, STD_IC, SPL_IC, ISD_IC, IC_OTHERS AS 0 AS TOTAL_IC_MOU IS 0 FOR CUSTOMER

FILTERED HIGH-VALUE CUSTOMERS

THE CALCULATED PERCENTILE FOR THE AVERAGE RECHARGE AMOUNT IN THE 6TH AND 7TH MONTHS IS 956.0.

UNIVARIATE ANALYSIS

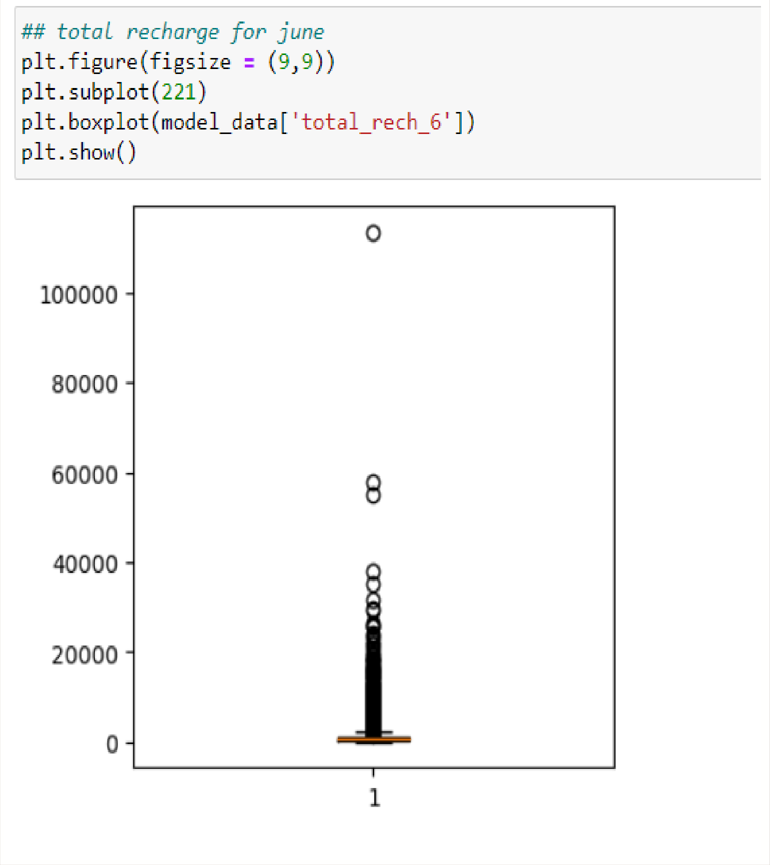
- We have 92% customers belong non-churn and 8% customers belong to Churn type. Clear indication of imbalance data.



UNIVARIATE ANALYSIS

- AS WE CAN SEE THERE ARE OUTLIERS WITH THE VARIABLE “TOTAL_RECH_6”

THE DATA IS NOT EVENLY DISTRIBUTED UNIFORMLY.

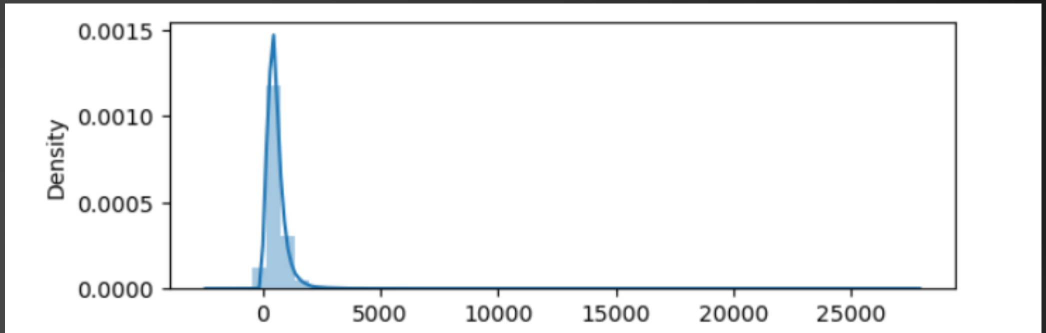


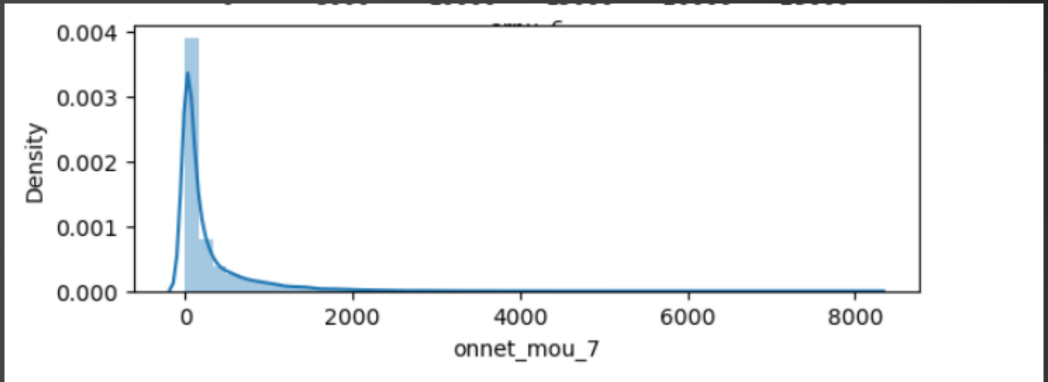
```python
## plotting the graph for average revenue per user
plt.subplot(211)
sns.distplot(model_data.arpu_6)
model_data.arpu_6.describe()
```

- From the graph, we can observe that the highest value for average revenue per user reaches 27731.

```python
## plotting the graph for onnet_mou_7
plt.subplot(212)
sns.distplot(model_data.onnet_mou_7)
model_data.onnet_mou_7.describe()
```

- From the graph, it is evident that the peak value is 8157.78.

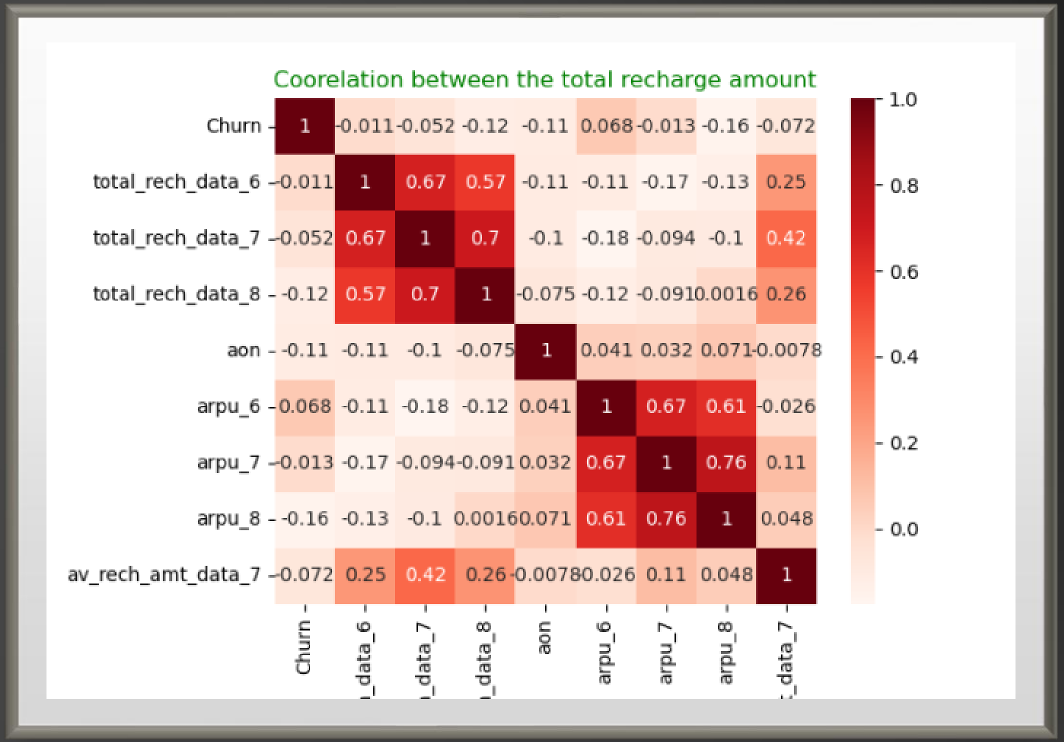




UNIVARIATE ANALYSIS

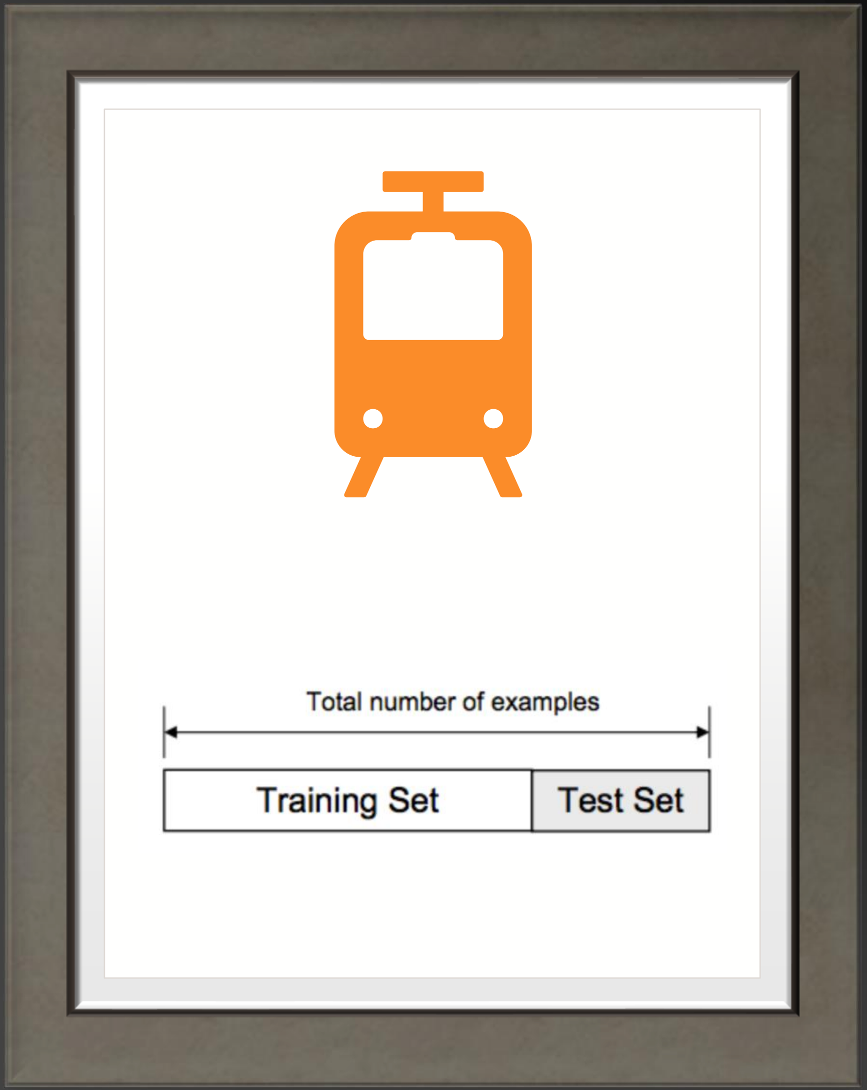
MULTIVARIATE ANALYSIS

- Average unit per user demonstrates a positive correlation with the average recharge amount
- Churn exhibits a positive correlation with the average revenue per user in the 6th month.



TEST-TRAIN SPLIT AND SCALING

Partition the dataset into training and testing data using a 70:30 ratio.



FEATURE SCALING

Feature scaling is done by using StandardScalar function

For training data , fit_transform function is used

For testing data , transform function is used

The allocation ratio between training and testing data may vary depending on the specific models being used.

MODEL BUILLDING

We have constructed multiple models employing the following algorithms:

- We have developed models utilizing **Principal Component Analysis (PCA) and Regression.**
- We have implemented a Logistic Regression model with the **Recursive Feature Elimination (RFE) and Variance Inflation Factor (VIF) techniques.**
- Additionally, we have built a **Decision Tree model** as part of our analysis. In our analysis, we've also incorporated an **ADA Boosting model** in conjunction with Decision Trees
- We have included a **random forest model** in our array of algorithms for analysis

Accuracy

Sensitivity & Specificity

Precision and Recall

EVALUATION METRICS APPLICABLE TO ALL MODELS

PCA with regression	
Precision Test :- 37,	**Recall:- 71.33**
Logistic Regression	
Precision Test:- 40.7,	**Recall :- 71.33**
Decision Tree	
Precision Test:- 73,	**Recall :- 46**
ADA Boosting with DT	
Precision Test:- 69.1,	**Recall :- 52.3**
Random Forests	
Precision Test:- 73 ,	**Recall :- 48.0**

CALCULATING PRECISION AND RECALL ON THE TEST DATASETS FOR DIFFERENT MODELS.

CONCLUSION

We observe that across most models, the values are consistently close to each other, and frequently, there is a trade-off between precision and recall. Recognizing the significance of both metrics, we believe that moving forward with Random Forests is a prudent choice.

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

- THE END