TELECOM CHURN CASE STUDY

Prepared By,

Dipjit Basak

Dipen Jaysukh Prajapati

Harshal Sapkade

Batch : BUSINESS ANALYTICS – DATA SCIENCE C60 – IIIT B(UPGRAD)-2023-24

PROBLEM STATEMENT & BUSINESS OBJECTIVE

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.

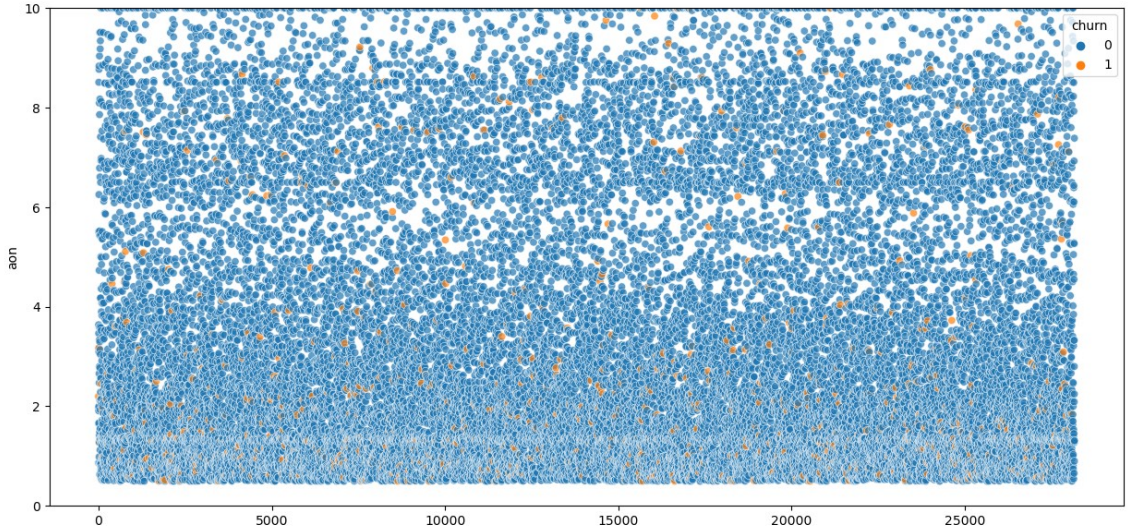
BUSINESS OBJECTIVE

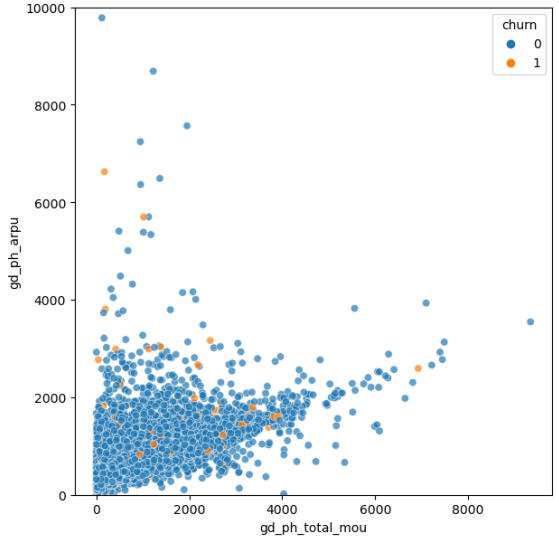
- The dataset contains customer-level information for a span of four consecutive months - June, July, August and September. The months are encoded as 6, 7, 8 and 9, respectively.
- The business objective is to predict the churn in the last (i.e. the ninth) month using the data (features) from the first three months. To do this task well, understanding the typical customer behavior during churn will be helpful.

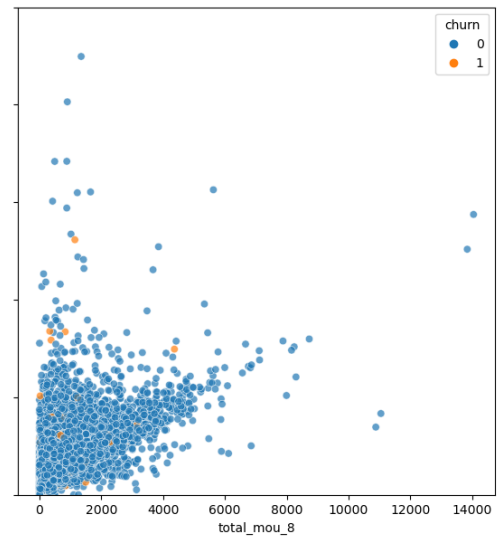
METHODOLOGY

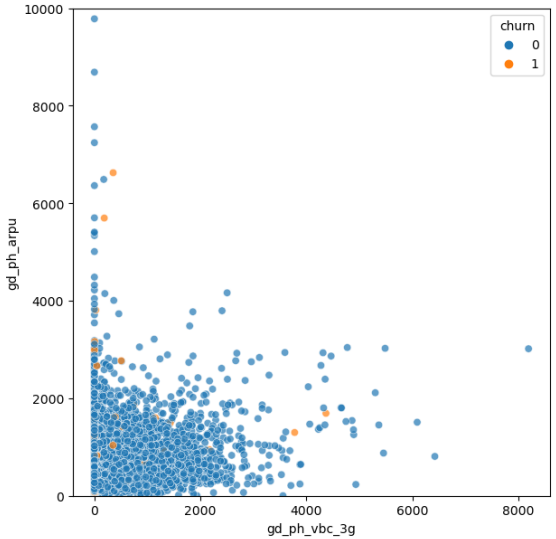
- **Data Cleaning**
  - Check and handle duplicate data.
  - Check and handle N/A values and missing values.
  - Drop columns, if it contains large amount of missing values and not useful for the analysis.
  - Imputation of the values if necessary.
  - Check and handle outliers in data.
- **Exploratory Data Analysis**
  - Univariate data analysis: value count, distribution of variable etc.
  - Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- **Data preparation, Standardization, Handling Class Imbalance, Principal Component Analysis(PCA)**
- **Selecting the best classification model:** Logistic regression, Decision Tree, Random Forest
- **Validation of the best model.**

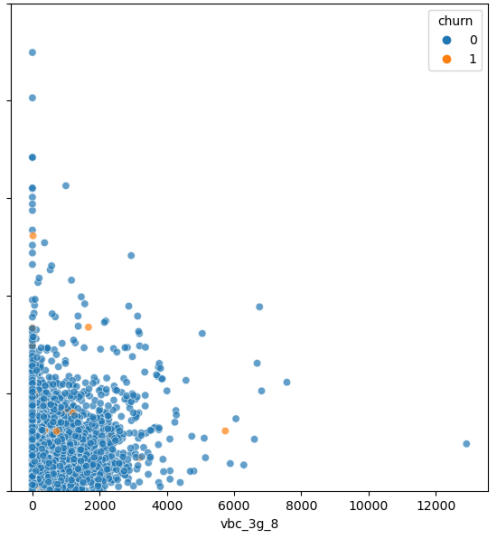
UNIVARIATE/MULTIVARIATE ANALYSIS 1





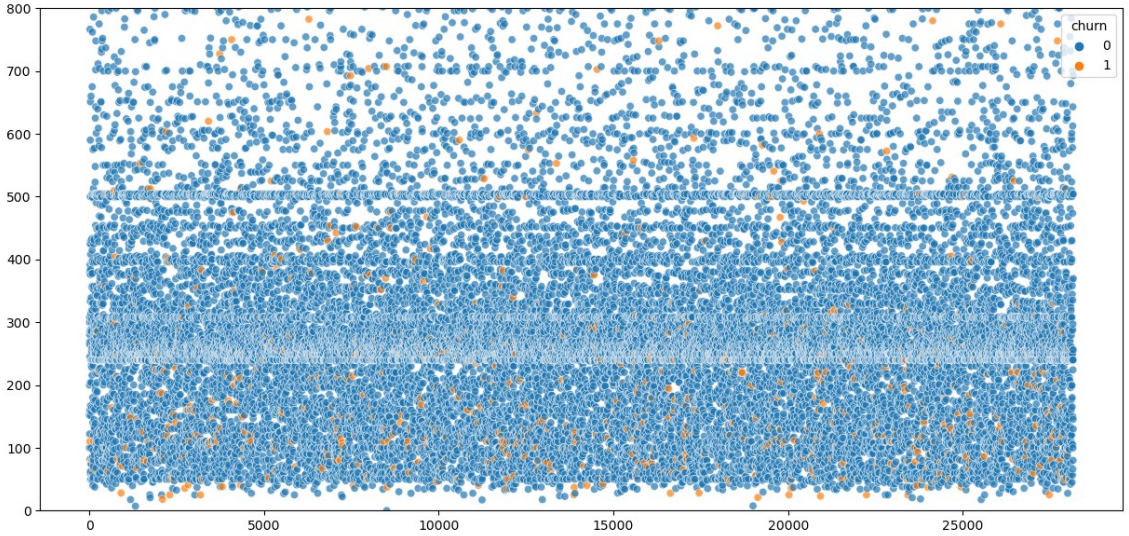


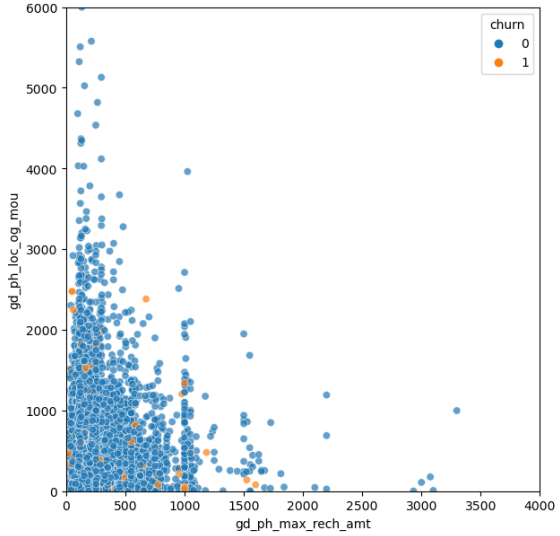


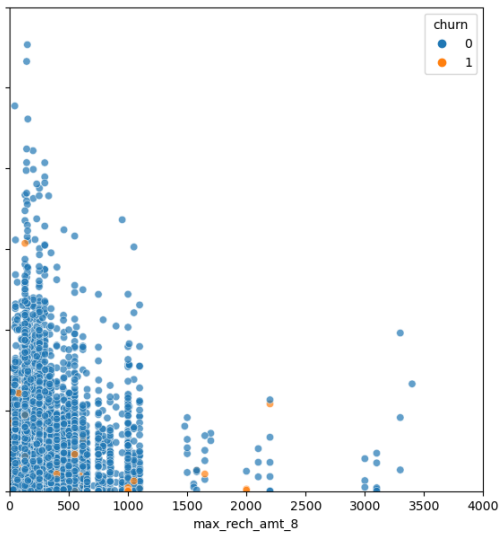


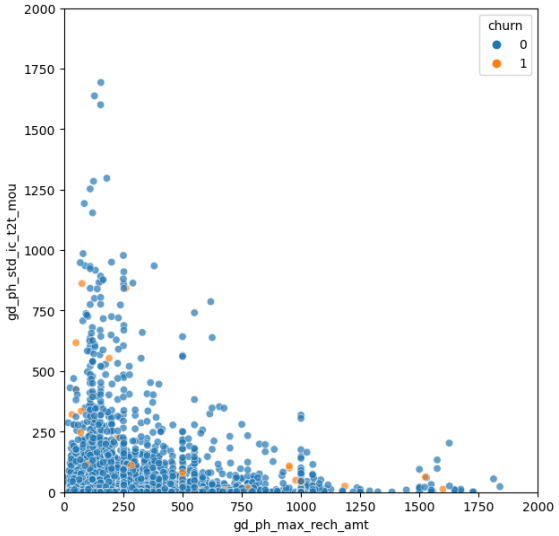
- Though we cannot see a clear pattern here, but we can notice that the majority of churners had a tenure of less than 4 years
- MOU have dropped significantly for the churners in the action phase i.e 8th month, thus hitting the revenue generated from them
- It is also interesting that though the MOU is between 0-2000, the revenue is highest in that region that tells us these users had other services that were boosting the revenue
- Users who were using very less amount of VBC data and yet were generating high revenue churned
- Revenue is higher towards the lesser consumption side

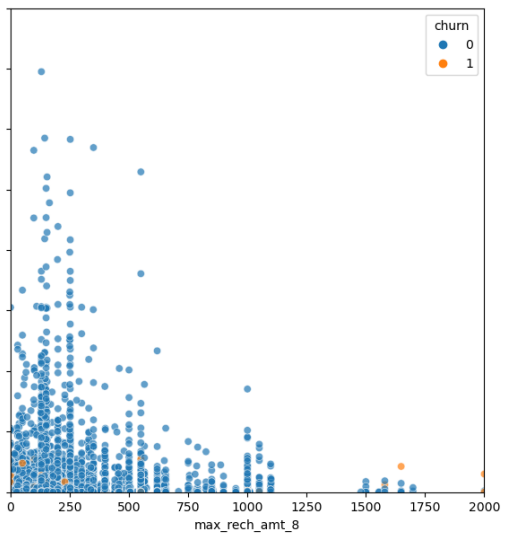
UNIVARIATE/MULTIVARIATE ANALYSIS 2





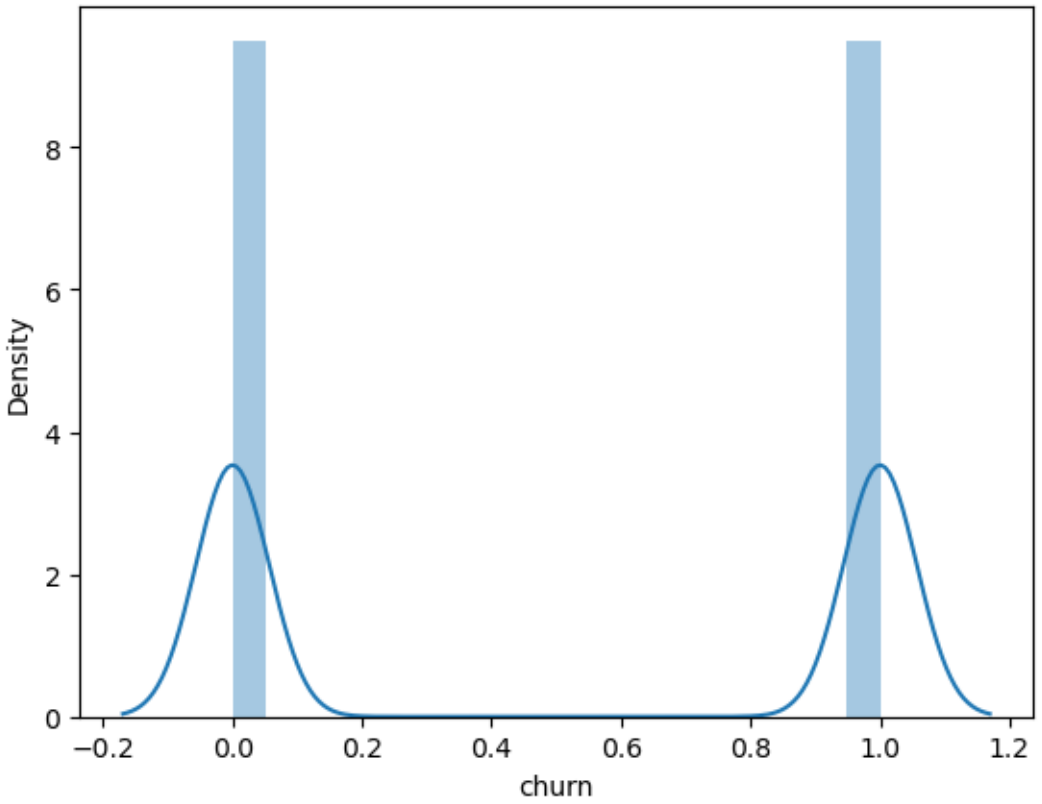






- Users who were recharging with high amounts were using the service for local uses less as compared to user who did lesser amounts of recharge
- Intuitively people whose max recharge amount as well as local out going were very less even in the good phase churned more
- We can see that users who had the max recharge amount less than 200 churned more
- Users who have max recharge amount on the higher end and still have low incoming call mou during the good phase, churned out more

HANDLING CLASS IMBALANCE & PRINCIPAL COMPONENT ANALYSIS



PCA

In [49]:
```
X.shape
```

Out[49]: (28163, 55)

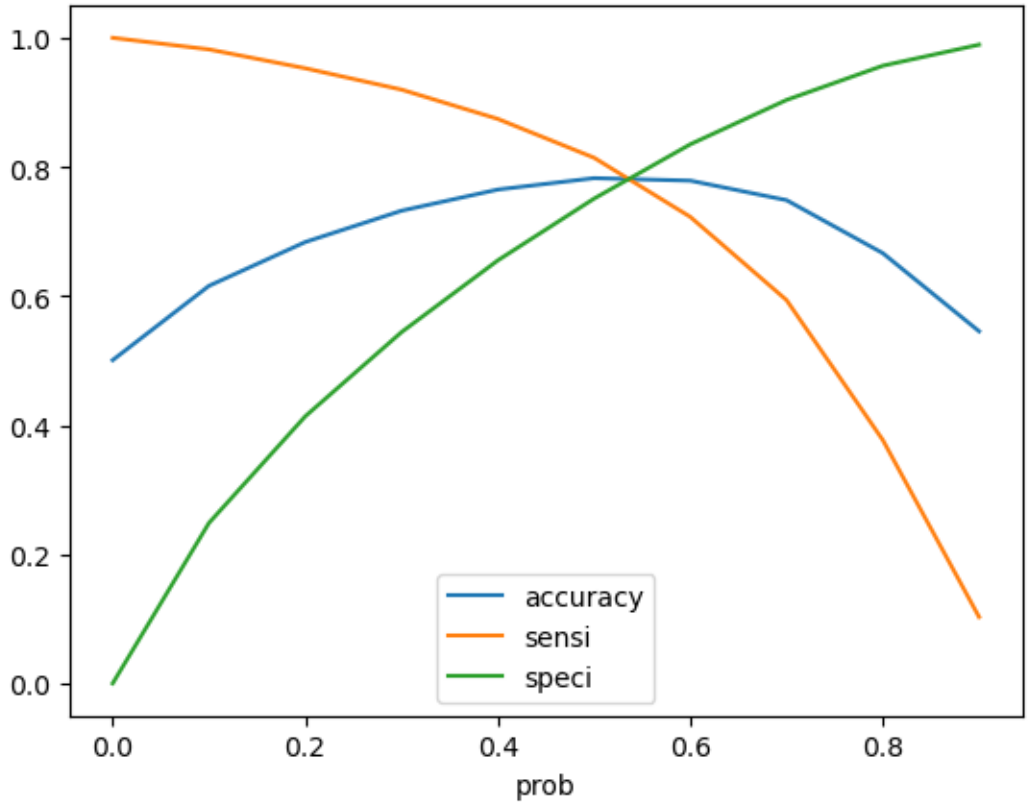
In [50]:
```
from sklearn.decomposition import PCA

pca = PCA(n_components=25)
X_pca = pca.fit_transform(X_res)
X_pca.shape
```

Out[50]: (54590, 25)

MODEL BUILDING

- As the dependent variable is categorical hence the general model is classification model.
- Now classification taught are- Logistic Regression, Decision Tree and Random Forest.
- Hence, all three models have been made and tested on various parameters and results like accuracy, precision, ROC.
- After analyzing all, the three models, the best model came out to be Random Forest.

ROC CURVE & DECISION TREE



Train accuracy : 0.8772145604898857

	precision	recall	f1-score	support
0	0.88	0.79	0.83	8215
1	0.81	0.89	0.85	8162
accuracy			0.84	16377
macro avg	0.85	0.84	0.84	16377
weighted avg	0.85	0.84	0.84	16377

CONCLUSION & STRATEGIES TO MANAGE CUSTOMER

- Given our business problem, to retain their customers, we need higher recall. As giving an offer to an user not going to churn will cost less as compared to loosing a customer and bring new customer, we need to have high rate of correctly identifying the true positives, hence recall.
- When we compare the models trained we can see the tuned random forest is performing the best, which is highest accuracy along with highest recall i.e. 95%. So, we will go with random forest.

Some of the factors we noticed while performing EDA which can be clubbed with these insights are:

- 1. Users whose maximum recharge amount is less than 200 even in the good phase, should have a tag and re-evaluated time to time as they are more likely to churn
- 2. Users that have been with the network less than 4 years, should be monitored time to time, as from data we can see that users who have been associated with the network for less than 4 years tend to churn more
- 3. MOU is one of the major factors, but data especially VBC if the user is not using a data pack if another factor to look out

THANKS