Telecom Churn Case Study

ANUBHA MISHRA

Agenda

Problem Statement

Methodology

Data Manipulation

Exploratory Data Manipulation

Model Building

Final Model

Key Insights

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

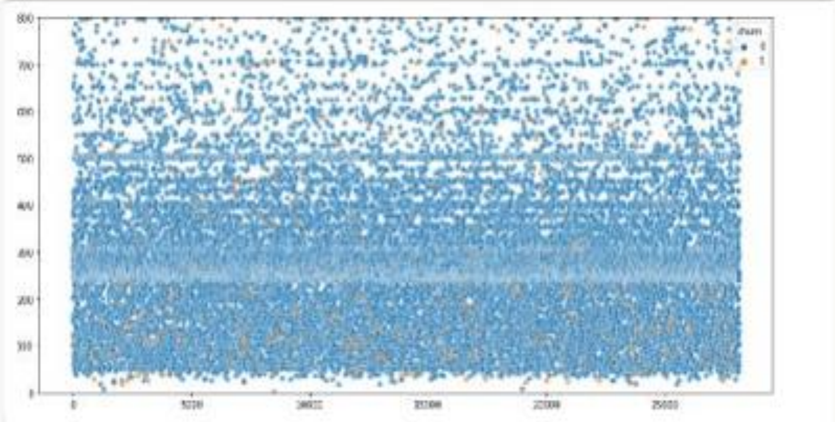
- Check and handle duplicate data.
- Check and handle NA values and missing values.
- Drop columns, if it contains large amount of missing values and not useful for the analysis.
- Imputation of the values, if necessary.
- Check and handle outliers in data.

❑ Exploratory data Analysis

- Univariate data analysis: value count, distribution of variable etc.
- Bivariate data analysis: correlation coefficients and pattern between the variables etc.

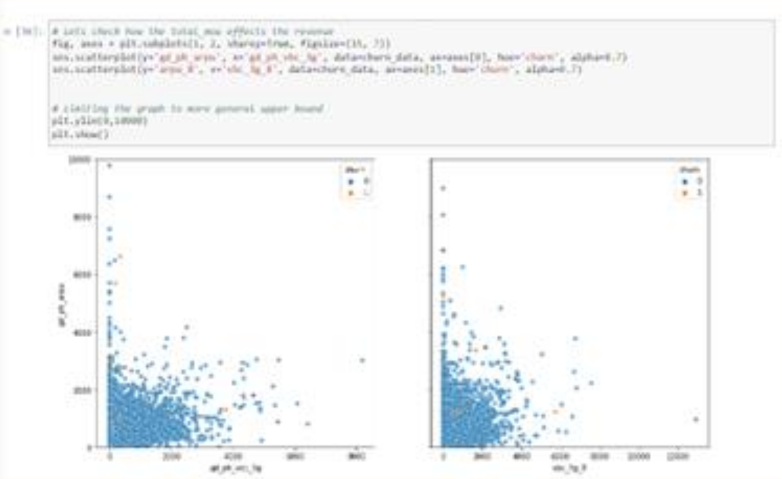
- ❑ Data preparation, Standardization, Handling Class Imbalance, Principal Component Analysis(PCA)
- ❑ Selecting the best classification model: Logistic regression, Decision Tree, Random Forest
- ❑ Validation of the best model.

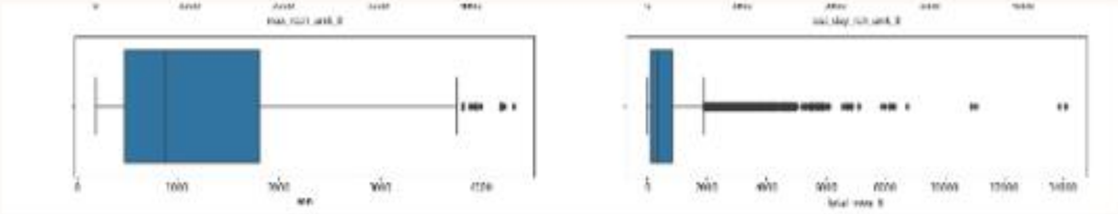
Univariate/ Multivariate Analysis



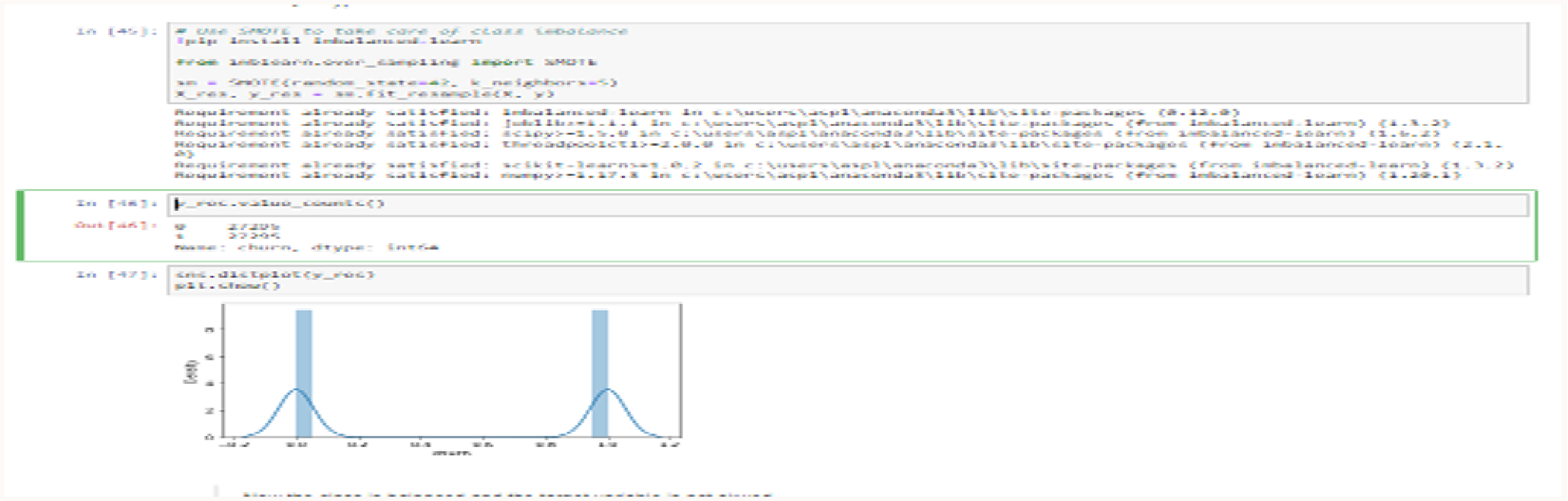
Observation

- We can see that users who had the max recharge amount less the 200 churned more





Handling Class Imbalance



Principal component Analysis

PCA

In [48]:
```
X.shape
```
Out[48]:
```
(28163, 55)
```

In [49]:
```
from sklearn.decomposition import PCA

pca = PCA(n_components=25)
X_pca = pca.fit_transform(X_res)
X_pca.shape
```
Out[49]:
```
(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 analysing all, the three models, the best model came out to be Random Forest

Conclusion

- 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.

Final Model

```
In [123]: final_model = RandomForestClassifier(max_depth=30, min_samples_leaf=5, n_jobs=-1,
                                               random_state=25)
```

```
In [124]: y_train_pred = rf_best.predict(X_train)
          y_test_pred = rf_best.predict(X_test)



          # Print the report
          print("Report on train data")
          print(metrics.classification_report(y_train, y_train_pred))

          print("Report on test data")
          print(metrics.classification_report(y_test, y_test_pred))
```

```
Report on train data
              precision    recall  f1-score   support

           0       1.00      0.99      1.00     19080
           1       0.99      1.00      1.00     19133

    accuracy                           1.00     38213
   macro avg       1.00      1.00      1.00     38213
weighted avg       1.00      1.00      1.00     38213

Report on test data
              precision    recall  f1-score   support

           0       0.93      0.87      0.89      8215
           1       0.87      0.93      0.90      8162

    accuracy                           0.90     16377
   macro avg       0.90      0.90      0.90     16377
weighted avg       0.90      0.90      0.90     16377
```

Key Insights

Strategies to Manage Customer Churn

The top 10 predictors are :

Features
loc_og_mou_8
total_rech_num_8
monthly_3g_8
monthly_2g_8
gd_ph_loc_og_mou
gd_ph_total_rech_num
last_day_rch_amt_8
std_ic_t2t_mou_8
sachet_2g_8
aon

- We can see most of the top predictors are from the action phase, as the drop in engagement is prominent in that phase

- 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