Telecom Churn case Study

- Trupti Gokhale

Contents

Problem Statement & Strategic Objective

Problem Statement:

In the telecommunications industry, customers can easily choose between various providers and often switch from one to another. This results in a high annual churn rate, typically ranging from 15% to 25%. With the cost of acquiring a new customer being 5 to 10 times greater than retaining an existing one, prioritizing customer retention has become more critical than acquisition. For many telecom operators, retaining valuable customers is their top priority. To mitigate churn, telecom companies need to identify customers who are at a high risk of leaving.

Strategic Objective:

The dataset includes customer data for four consecutive months—June, July, August, and September, encoded as months 6, 7, 8, and 9.The aim is to forecast customer churn for September using data from the first three months. Understanding typical patterns of customer behavior associated with churn will enhance the accuracy of this prediction.

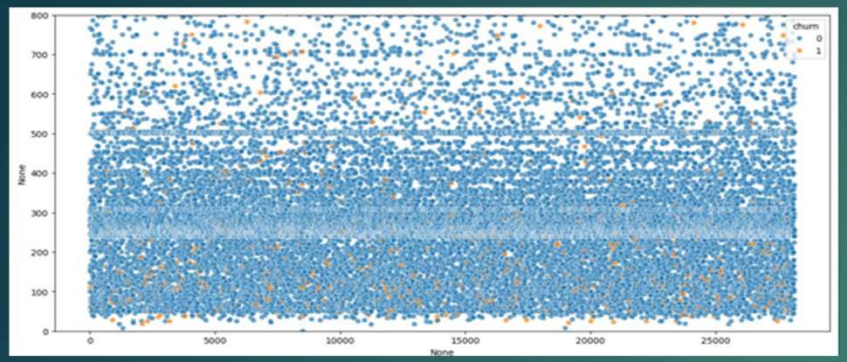
Steps to be followed

Data Cleaning:
Identify and manage duplicate records.
Address and manage missing or NA values.
Remove columns with excessive missing values that do not contribute to the analysis.
Perform imputation to fill in missing values, if required.
Detect and manage outliers in the data.

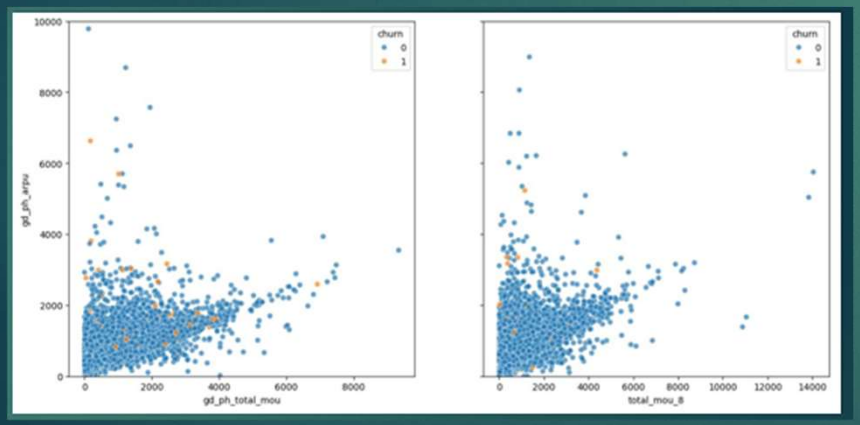
Exploratory Data Analysis:
Univariate Analysis: Examine value counts and the distribution of individual variables.
Bivariate Analysis: Analyze correlations and patterns between pairs of variables.

Data Preparation:
Standardize data.
Address class imbalance issues. Apply Principal Component Analysis (PCA) if needed.
Model Selection: Evaluate and choose the most suitable classification model, such as Logistic Regression, Decision Tree, or Random Forest.
Model Validation: Assess and validate the performance of the selected model.

Univariate and Multivariate Analysis

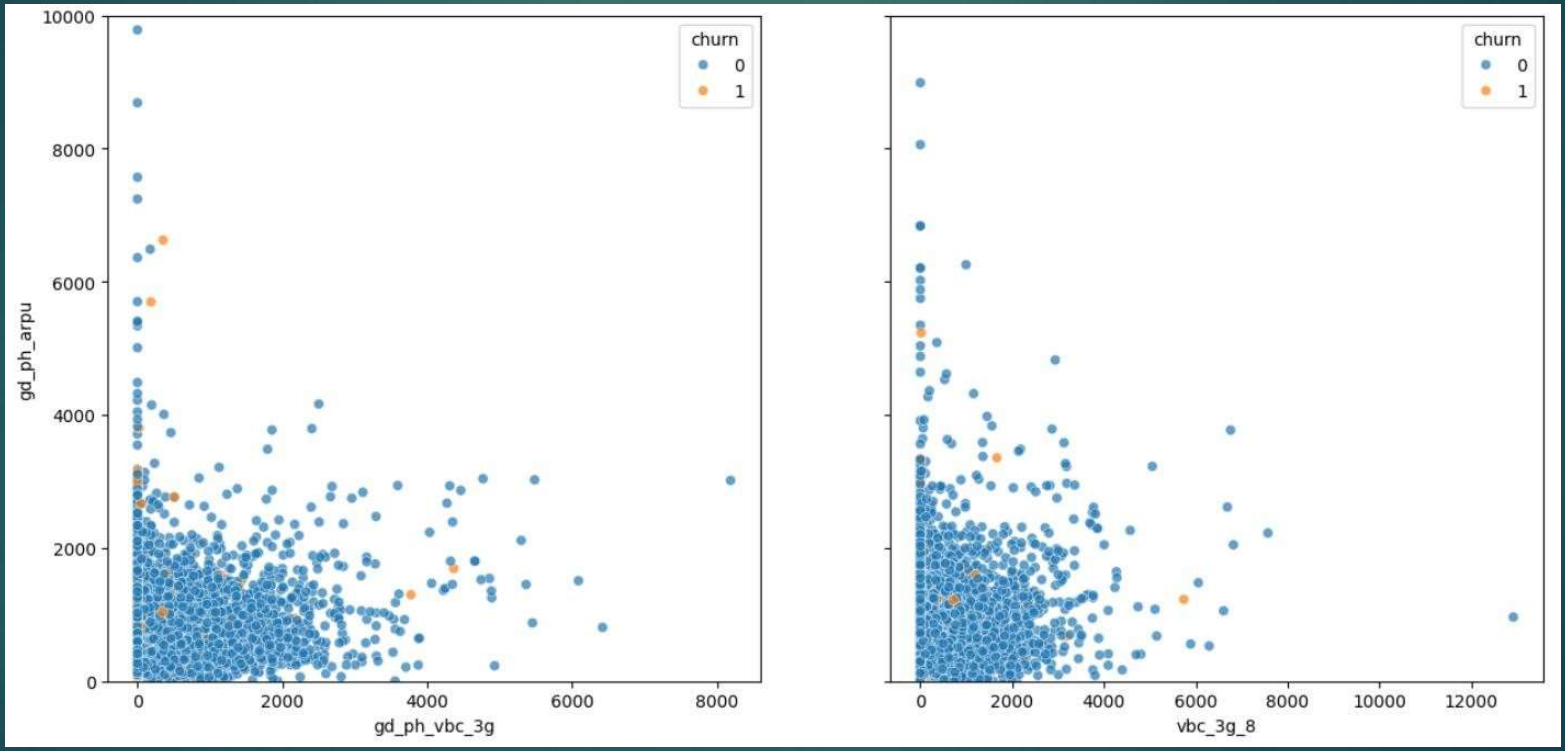


- It is evident that users with a maximum recharge amount below 200 have a higher churn rate.

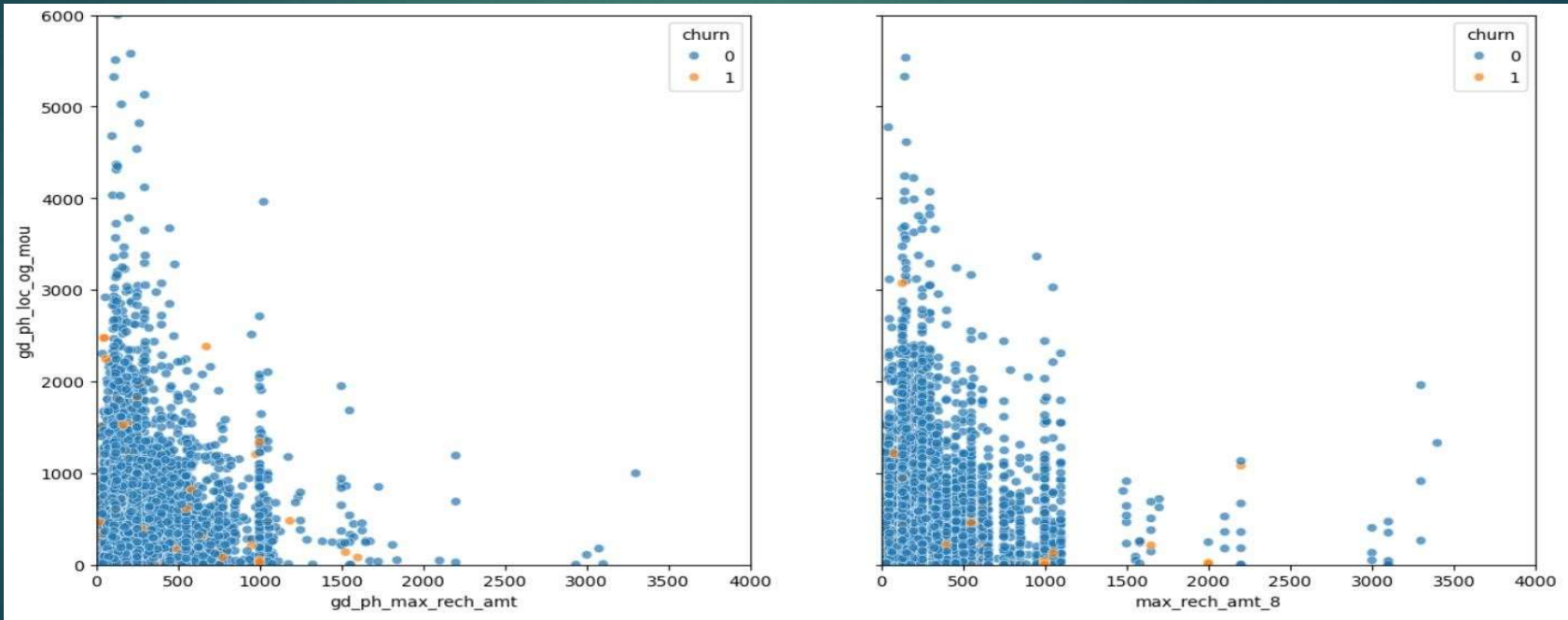


- It is apparent that the Minutes of Use (MOU) for churners decreased significantly during the action phase (8th month), leading to a reduction in the revenue generated from these users.
- It is also noteworthy that despite the MOU between 0 to 2000, the revenue is highest within this range. This suggests that these users likely had additional services that contributed to the increased revenue.

- We can see that users who generated significant revenue while using minimal VBC data were more likely to churn.
- Moreover, it is evident that higher revenue correlates with lower data consumption.



- Users who made larger recharge amounts were using the service for local calls less frequently compared to those who recharged with smaller amounts.
- It appears that users who had low recharge amounts and minimal local outgoing calls, even during stable periods, tended to churn more.



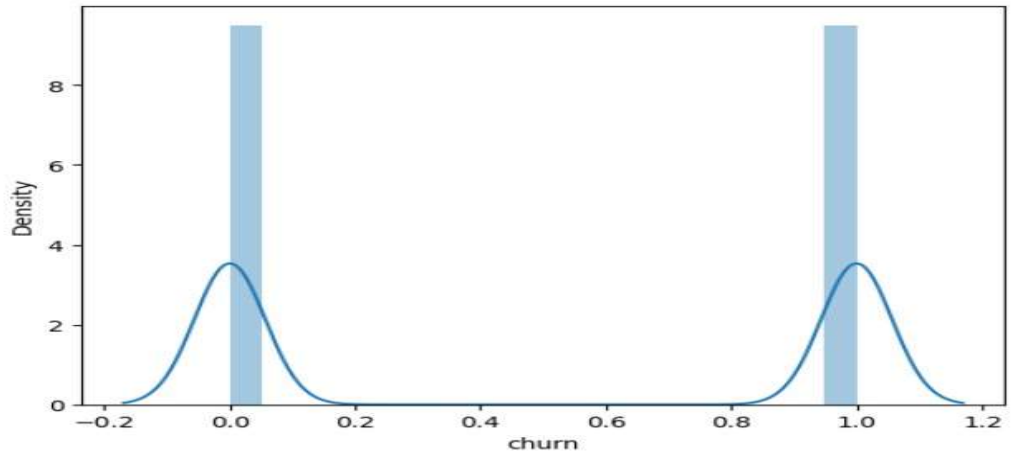
Handling Class imbalance

```
[56] # Using SMOTE to take care of class imbalance

     from imblearn.over_sampling import SMOTE

     sm = SMOTE(random_state = 42, k_neighbors = 5)
     X_res, y_res = sm.fit_resample(X, y)
```

```
[57] y_res.value_counts()
```

```
churn
1    27295
0    27295
Name: count, dtype: int64
```

```
[58] sns.distplot(y_res)
     plt.show()
```



PCA (Principal Component Analysis)

7.3 PCA

```
[59] X.shape
```

```
(28163, 55)
```

```
[60] from sklearn.decomposition import PCA

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

```
(54590, 25)
```

Model Building

- Since the dependent variable is categorical, the appropriate approach is to use a classification model.

- The classification methods considered include Logistic Regression, Decision Tree, and Random Forest.

- All three models have been developed and evaluated based on various metrics such as accuracy, precision, and ROC.

- After a thorough analysis, Random Forest emerged as the most effective model among the three.

Conclusion

Given our business problem, increasing customer retention requires a higher recall rate. This is because offering incentives to users who are at risk of churning is more cost-effective than losing a customer and acquiring a new one. Therefore, it is crucial to accurately identify those who are likely to churn, which is why recall is a key metric.

Upon comparing the trained models, the tuned Random Forest model demonstrates the best performance, achieving both the highest accuracy and a recall rate of 95%. Thus, we will proceed with the Random Forest model.

Final Model

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

```
[156] 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       0.99      0.98      0.99     19080
                 1       0.98      0.99      0.99     19133

          accuracy                           0.99     38213
         macro avg       0.99      0.99      0.99     38213
      weighted avg       0.99      0.99      0.99     38213

      Report on test data
                    precision    recall  f1-score   support

                 0       0.97      0.93      0.95      8215
                 1       0.93      0.97      0.95      8162

          accuracy                           0.95     16377
         macro avg       0.95      0.95      0.95     16377
      weighted avg       0.95      0.95      0.95     16377
```

Observations & insights

The top 10 predictors are :

Features

---

1. loc_og_mou_8
2. total_rech_num_8
3. monthly_3g_8
4. monthly_2g_8
5. gd_ph_loc_og_mou
6. gd_ph_total_rech_num
7. last_day_rch_amt_8
8. std_ic_t2t_mou_8
9. sachet_2g_8
10. aon

- Most of the top predictors come from the action phase, where there is a noticeable drop in engagement.

- During EDA, we observed several factors that can be associated with these insights:

✓ ***Users with a maximum recharge amount under 200 during the good phase should be flagged and monitored regularly, as they have a higher likelihood of churning.***

✓ ***Users who have been with the network for less than 4 years should be periodically monitored, as data indicates they have a higher tendency to churn.***

✓ ***MOU is a key factor to consider, but data usage, particularly VBC (Volume Based Charging is also crucial, especially if the user is not using a data pack.***

Thank You !