Telecom Churn Case Study

- Mukul Dutt Bharadwaj
- Naveen Rajpal
- Narra Siva Sai Kumar

Agenda

- Problem Statement
- Methodology
- Exploratory Data Analysis
- Data Manipulation
- Model Building
- Final Model
- Key Insights

Problem Statement

In the telecom Industry customers are able to choose from multiple service providers and activity switch from one operator to another. In this highly competitive market, the telecommunication industry experience an average of 15-25% annual churn rate. Given the fact that it cost 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.

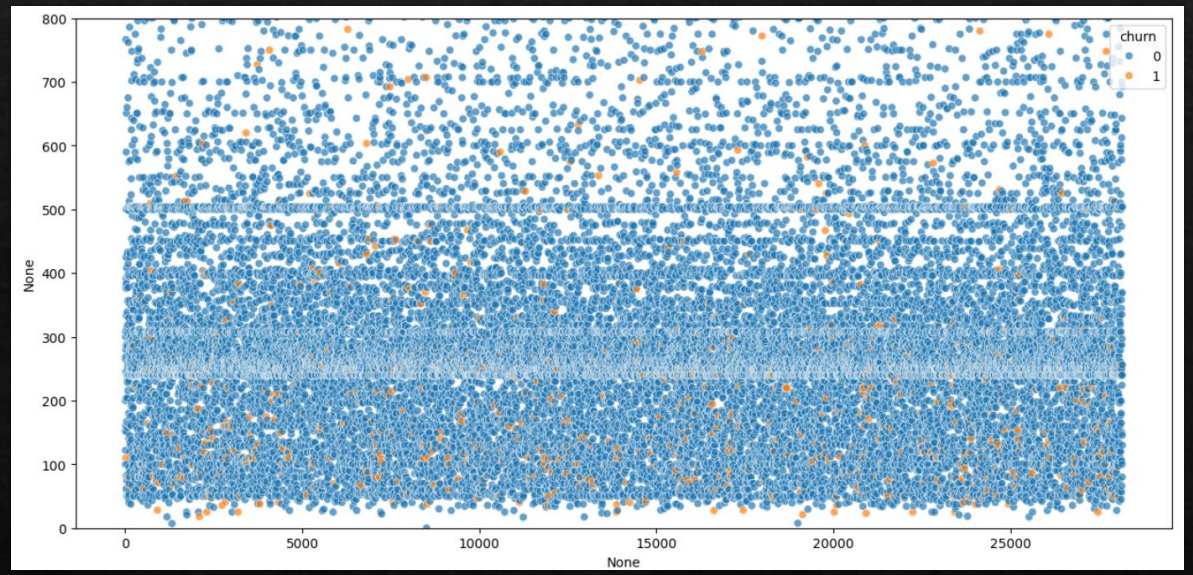
Business Objective

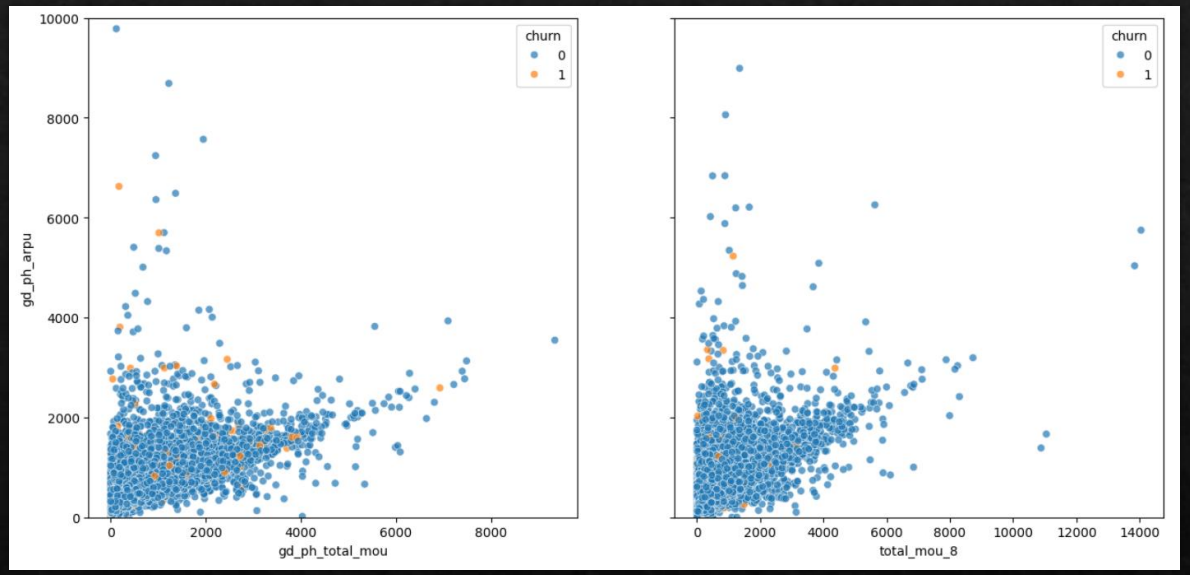
- The dataset contains customers-level information for a span of four consecutive months – Jun, Jul, Aug and Sep. The months are encoded as 6, 7, 8 and 9 respectively.
- The business objective is to predict the churn in the last (i.e. ninth month) using the data (features) from the first three months. 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

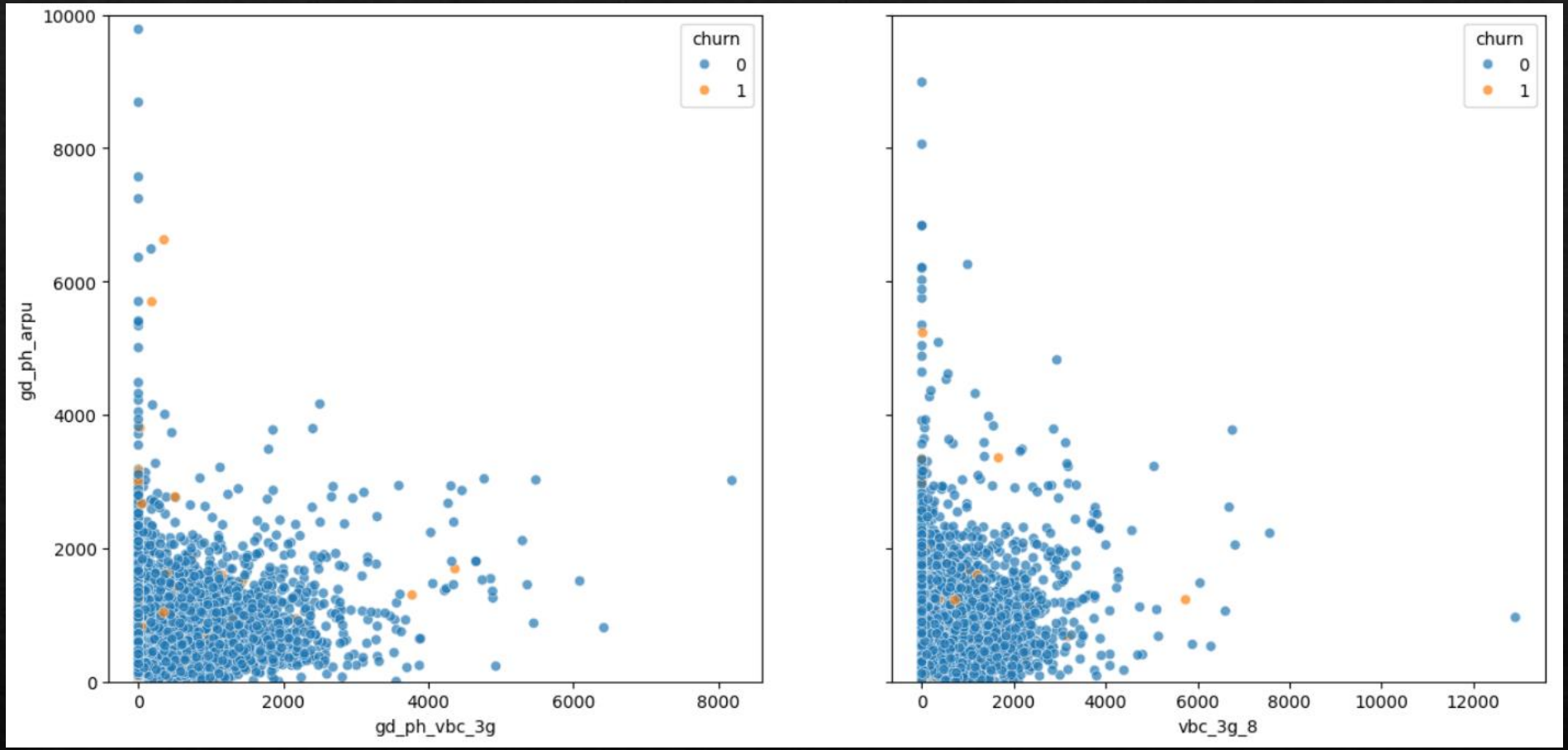
Univariate / Multivariate Analysis



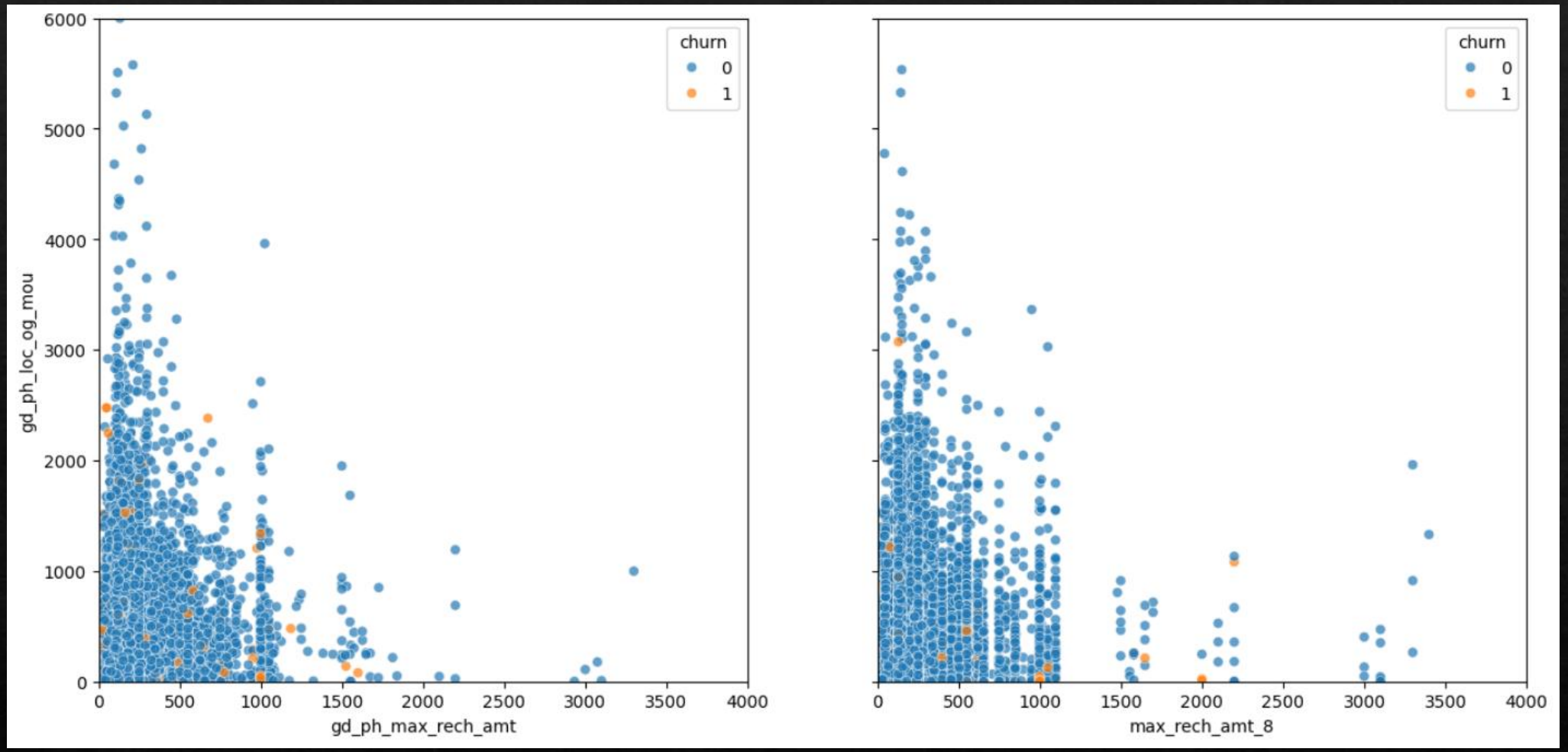


- We can see that users who had the max recharge amount less than 200 churned more

- We can clearly see that 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



- We can see that the users who were using very less amount of VBC data and yet were generating high revenue churned
- Yet again we see that the revenue is higher towards the lesser consumption side



- 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

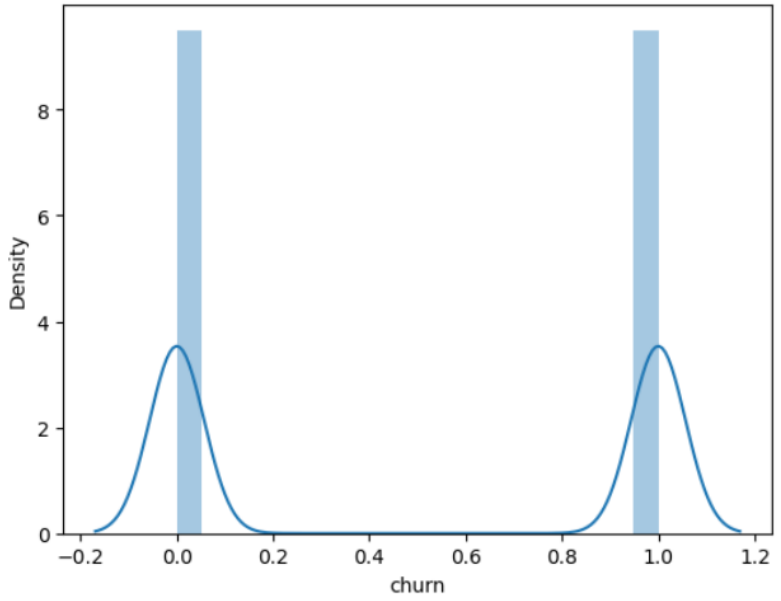
Handling Class Imbalance

```
[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()
```



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

- As the dependent variable is categorical hence the general model is a 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.

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

```
[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
```

Key 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

- 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:
- 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
- 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
- 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

Thank you,