Telecom Churn Case Study

Submission by : Mirnalini Beniwal and Soumik Chakraborty

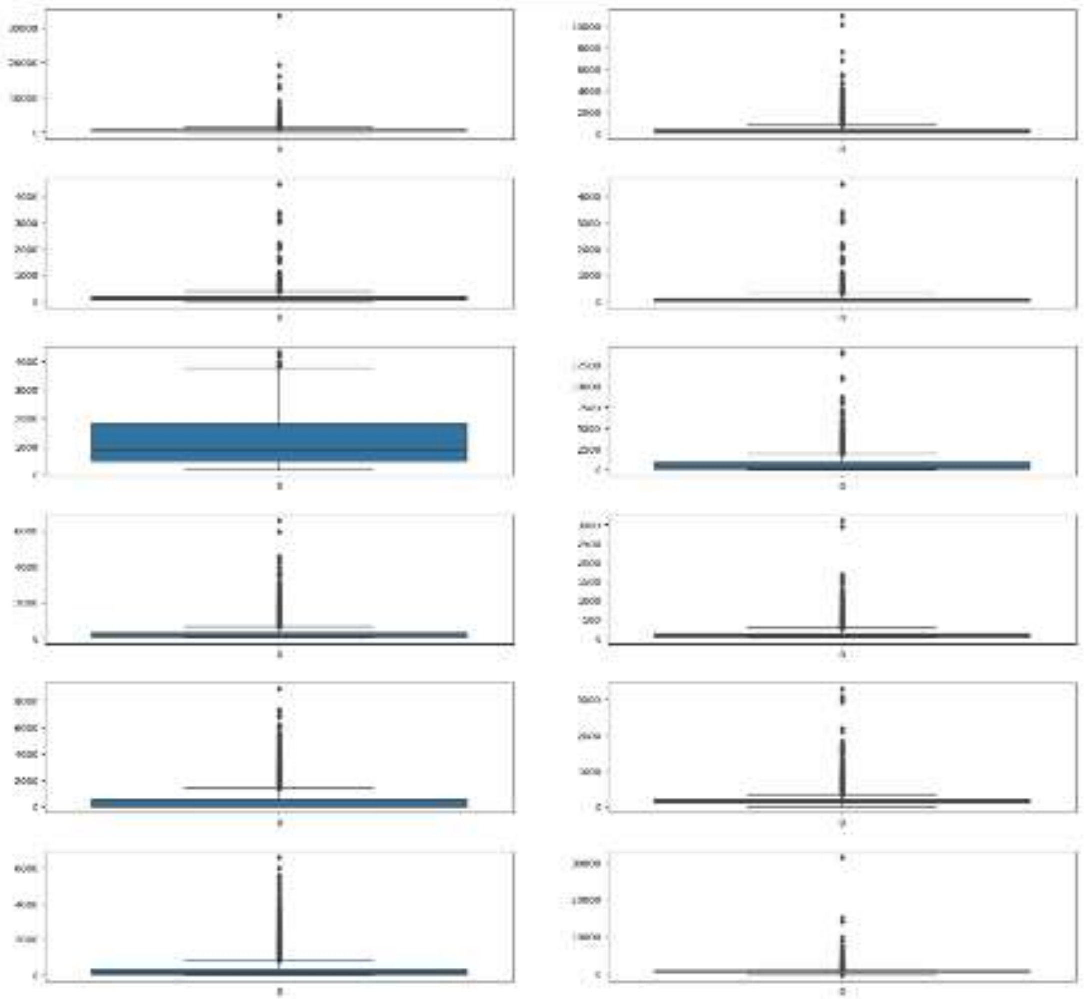
Problem Statement

- In the telecom industry, customers can 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.
- To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.
- Analyse 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.

Steps involved:

- Importing required Python libraries
- Data Exploration: Perform exploratory data analysis (EDA) to understand the distribution of features, identify patterns, and handle missing values or outliers.
- Data Preprocessing: Prepare the data for modeling by handling categorical variables, scaling features, and splitting the dataset into training and testing sets.
- Model Selection: Choose a logistic regression as modeling algorithm. Logistic regression is suitable for binary or multiclass classification problems.
- Model Training: Use the training dataset to train the logistic regression model. The model learns the relationship between the input features and the target variable during this phase.
- Model Evaluation: Assess the performance of the trained model using the testing dataset.
- Feature Engineering: creating new features or transforming existing ones to enhance the model's predictive power.
- Interpretation of Results: Understand the coefficients of the logistic regression model and their impact on the predicted probabilities. Interpretation can provide insights into the relationships between features and the target variable.

Observations:

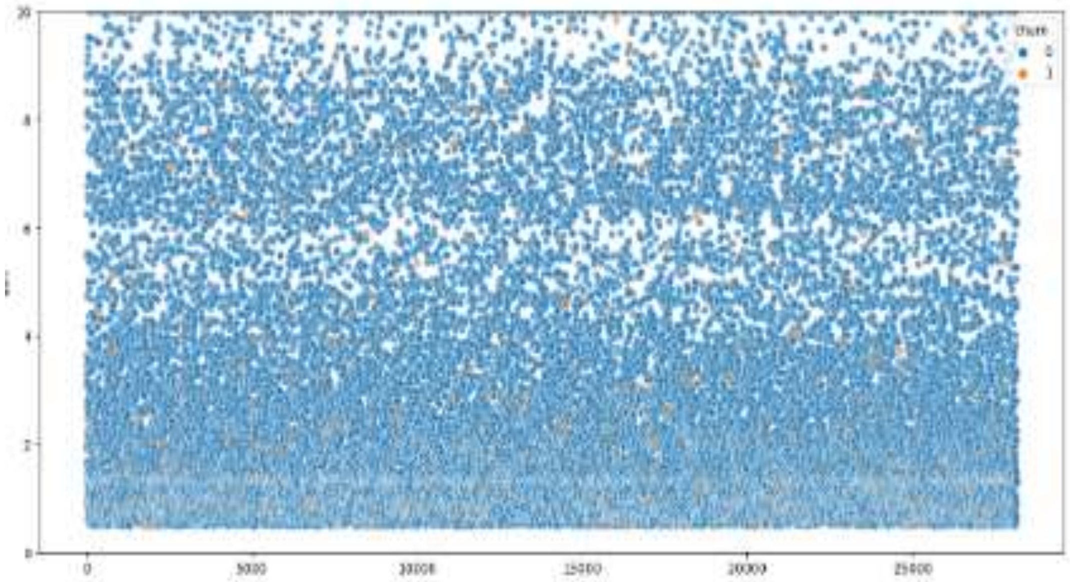


Observations

- From the above plots we can define following upper limits to the sepected variables

| Feature | Value | | --- | --- | | arpu_8 | 7000| | loc_og_mou_8 | 4000| | max_rech_amt_8 | 1000 | | last_day_rch_amt_8 | 1000 | | aon | 3000 | | total_mou_8 | 4000 | | gd_ph_loc_ic_mou | 8000 | | gd_ph_last_day_rch_amt | 1000 | | gd_ph_std_og_mou | 4000 | | gd_ph_max_rech_amt | 1500 | | gd_ph_loc_og_mou | 3000 | | gd_ph_arpu | 7000 |

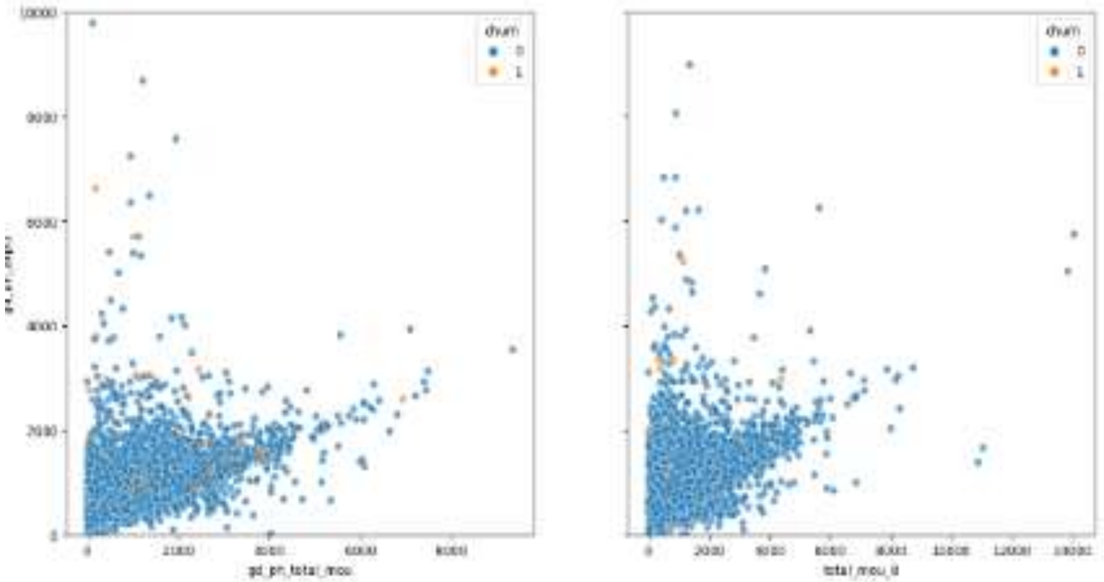
- We will make these changes post exploration of other features



Observation

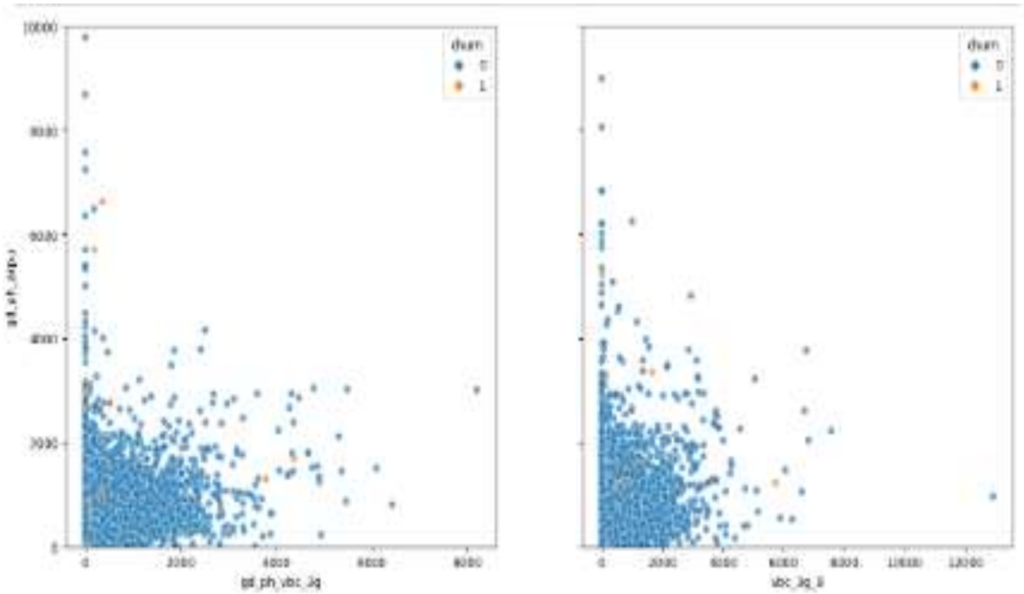
- Though we cannot see a clear pattern here, but we can notice that the majority of churners had a tenure of less than 4 years

Observations:



Observation

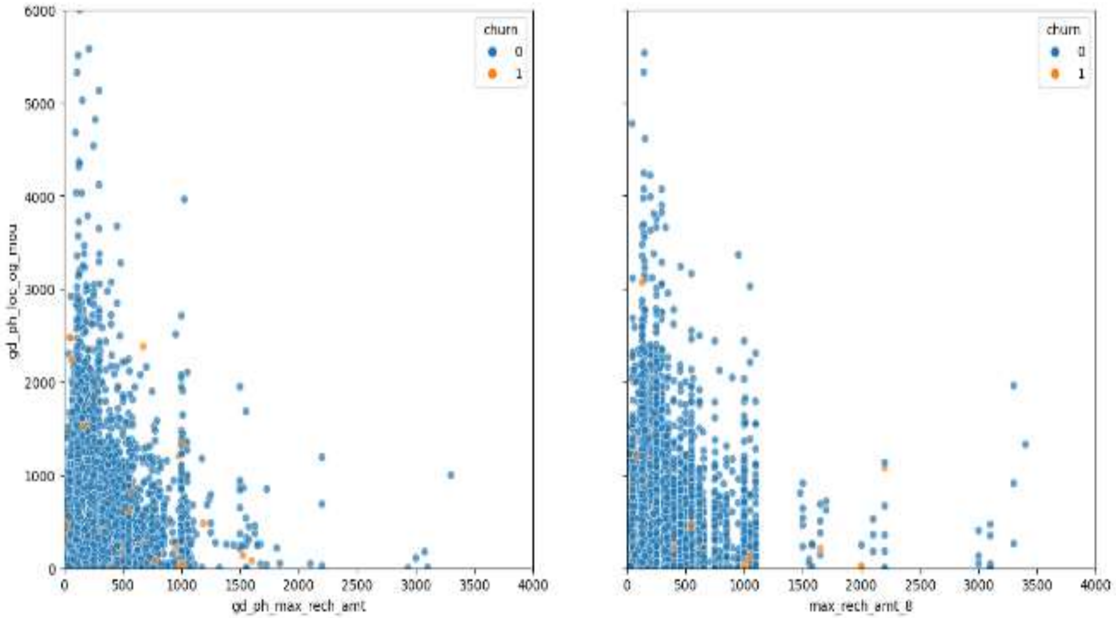
- We can clearly see that MOU have dropped significantly for the churners in the action pahse i.e 8th month, thus hitting the revenue generated from them
- It is also interesting that though the MOU is between 0-2000, the revenue is highest in that region that tells us these users had other services that were boosting the revenue

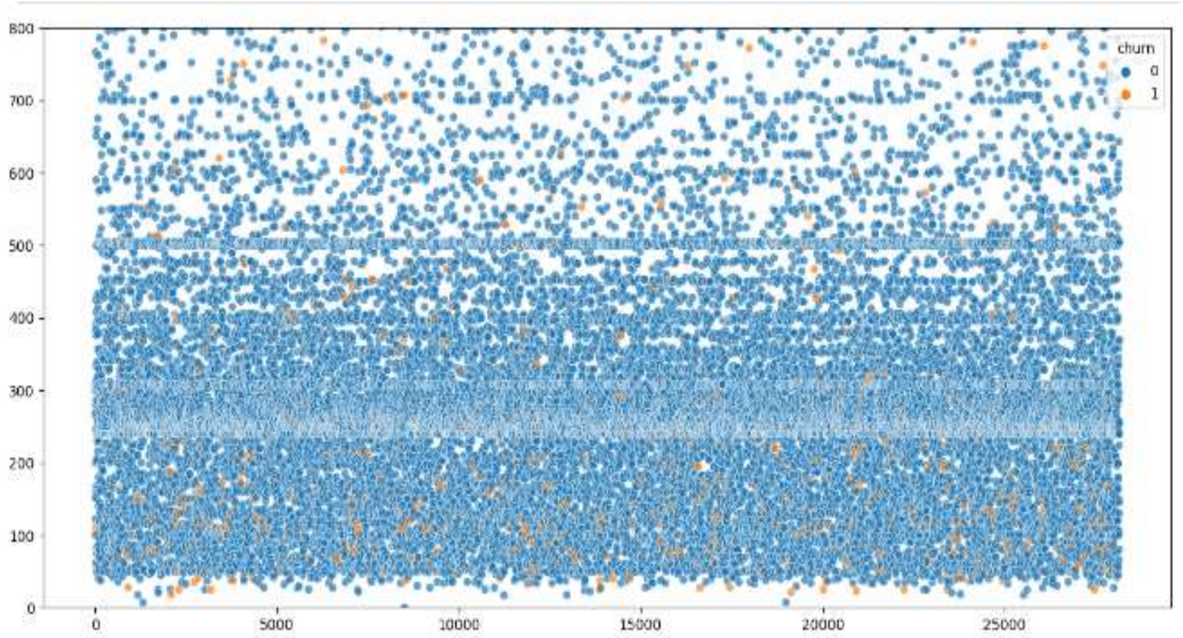


Observation

- 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

Observations:





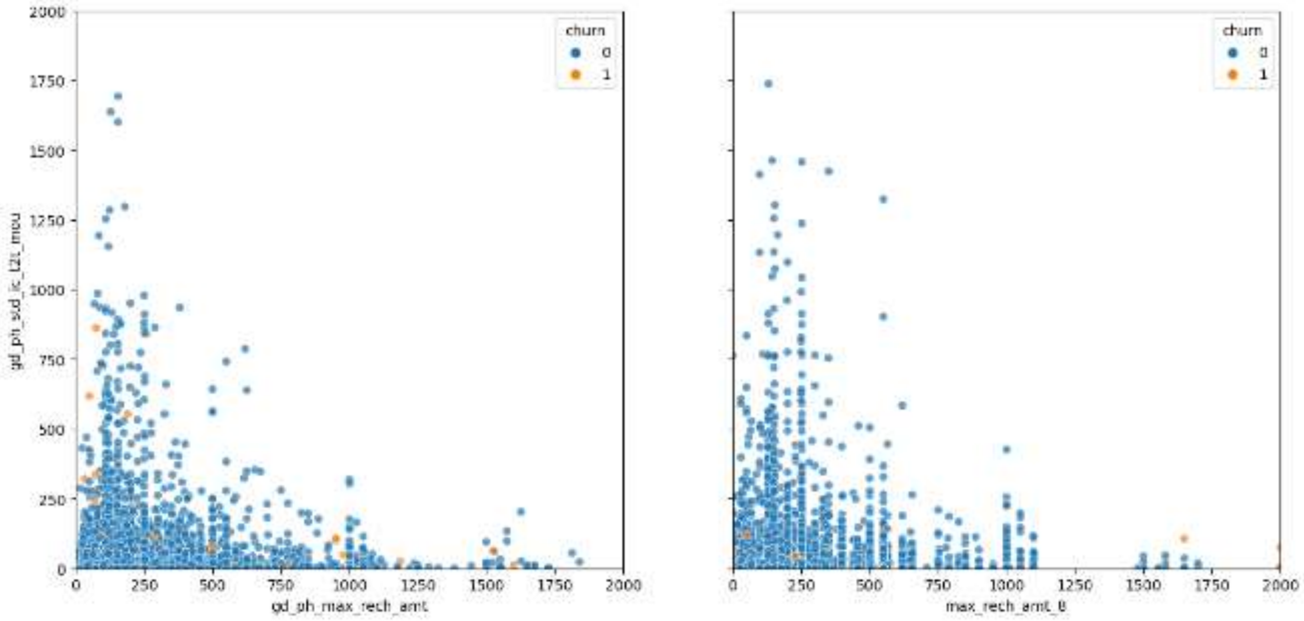
Observations

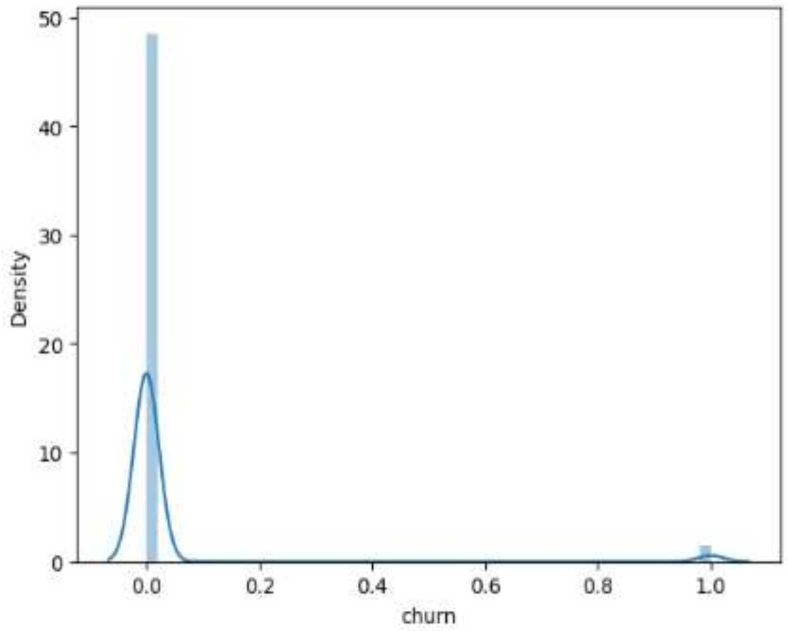
- Users who were recharging with high amounts were using the service for local uses less as compared to user who did lesser amounts of recharge
- Intuitevly people whose max recharge amount as well as local out going were very less even in the good phase churned more

Observation

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

Observations:





Observation

- Users who have max recharge amount on the higher end and still have low incoming call mou during the good pahse, churned out more

Observation

- Though the varible is not skwed it is higly imbalanced, the number of non-churners in the dataset is around 94%
- We will handle this imbalance using SMOTE algorithm

Observations:

Handling Class Imbalance

```
churn_data.churn.value_counts()
```

```
0    27295
1      868
Name: churn, dtype: int64
```

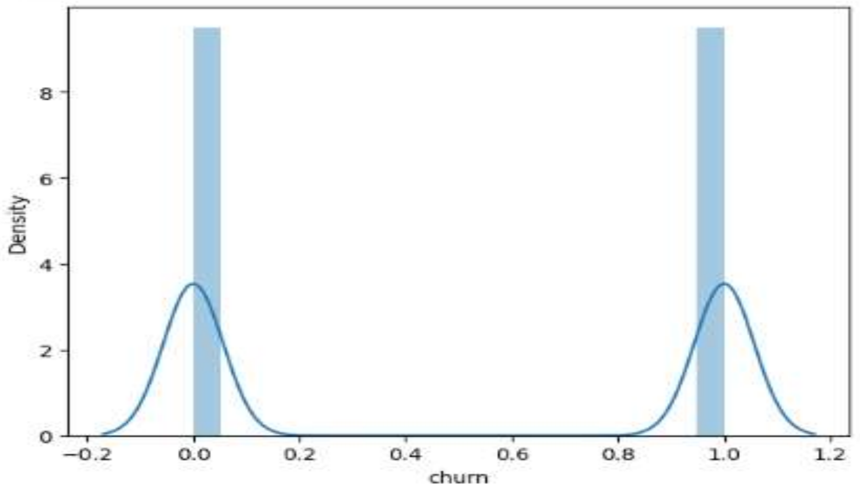
```
# Use SMOTE to take care of class imbalance
from imblearn.over_sampling import SMOTE

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

```
y_res.value_counts()
```

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

```
sns.distplot(y_res)
plt.show()
```



Observations:

- Using Logistic regression we are getting an accuracy of 78.5% on train data and 78.8% on test data
- We can see most of the critical features are form the action phase, which is in line with the business understanding that the action phase needs more attention.
- We are getting an accuracy of 90% on test data, with decision tree.
- Given our business probelm, 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 losing 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 and ada boost are performing the best, which is the highest accuracy along with the highest recall i.e. 95% and 97% respectively. So, we will go with random forest instead of adaboost as that is a comparatively simpler model.
- Users whose maximum recharge amount is less than 200 even in the good pahse, 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

Business Insights:

- Telecom company needs to pay attention to the roaming rates. They need to provide good offers to the customers who are using services from a roaming zone. -- The company needs to focus on the STD and ISD rates. Perhaps, the rates are too high. Provide them with some kind of STD and ISD packages. -- Concentrate on users with 1.27 std devations lower than average incoming calls from fixed line. They are most likely to churn.