Telecom Churn Case Study

ANKIT MAURYA | RAKESH | VIKAS

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.
- In this project, you will 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.

Definitions of Churn
There are various ways to define churn, such as:

- **Revenue-based churn:** Customers who have not utilised any revenue-generating facilities such as mobile internet, outgoing calls, SMS etc. over a given period of time. One could also use aggregate metrics such as 'customers who have generated less than INR 4 per month in total/average/median revenue'.
- The main shortcoming of this definition is that there are customers who only receive calls/SMSes from their wage-earning counterparts, i.e. they don't generate revenue but use the services. For example, many users in rural areas only receive calls from their wage-earning siblings in urban areas.
- **Usage-based churn:** Customers who have not done any usage, either incoming or outgoing - in terms of calls, internet etc. over a period of time.
- A potential shortcoming of this definition is that when the customer has stopped using the services for a while, it may be too late to take any corrective actions to retain them. For e.g., if we define churn based on a 'two-months zero usage' period, predicting churn could be useless since by that time the customer would have already switched to another operator.
- In this project, we will use the **usage-based** definition to define churn.
- **High-value churn** In the Indian and the southeast Asian market, approximately 80% of revenue comes from the top 20% customers (called high-value customers). Thus, if we can reduce churn of the high-value customers, we will be able to reduce significant revenue leakage.
- In this project, we will define high-value customers based on a certain metric (mentioned later below) and predict churn only on high-value customers

Steps used for solving case study

1. Importing Libraries
2. Loading and reading the data
3. Handling missing Values
4. EDA Univariate and Bivariate analysis
5. Model Building
6. PCA
7. PCA and logistic Regression
8. Evaluation on the test data
9. Hyperparameter Tuning- PCA and logistic regression.
10. Random Forest
11. Feature Importance
12. Extract of Intercept and coefficient from the logistic regression.
13. Final Business Insight.

Understanding the Business Objective and the Data

- 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.

Understanding Customer Behavior During Churn

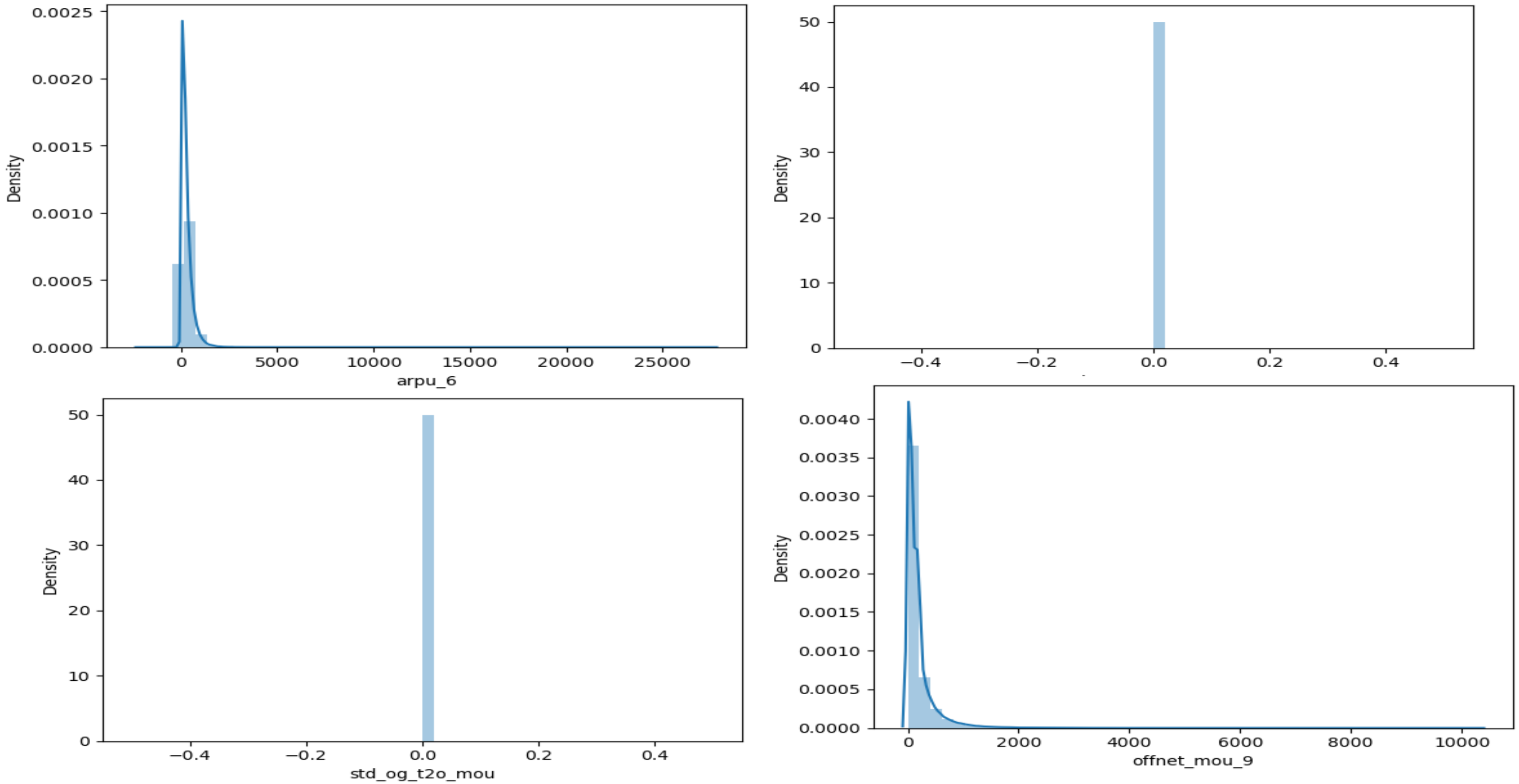
- Customers usually do not decide to switch to another competitor instantly, but rather over a period of time (this is especially applicable to high-value customers). In churn prediction, we assume that there are three phases of customer lifecycle :
- **The ‘good’ phase:** In this phase, the customer is happy with the service and behaves as usual.
- **The ‘action’ phase:** The customer experience starts to sore in this phase, for e.g. he/she gets a compelling offer from a competitor, faces unjust charges, becomes unhappy with service quality etc. In this phase, the customer usually shows different behaviour than the ‘good’ months. Also, it is crucial to identify high-churn-risk customers in this phase, since some corrective actions can be taken at this point (such as matching the competitor’s offer/improving the service quality etc.)
- **The ‘churn’ phase:** In this phase, the customer is said to have churned. We define churn based on this phase. Also, it is important to note that at the time of prediction (i.e. the action months), this data is not available to us for prediction. Thus, after tagging churn as 1/0 based on this phase, we discard all data corresponding to this phase.
- In this case, since we are working over a four-month window, the first two months are the ‘good’ phase, the third month is the ‘action’ phase, while the fourth month is the ‘churn’ phase.

Data Preparation

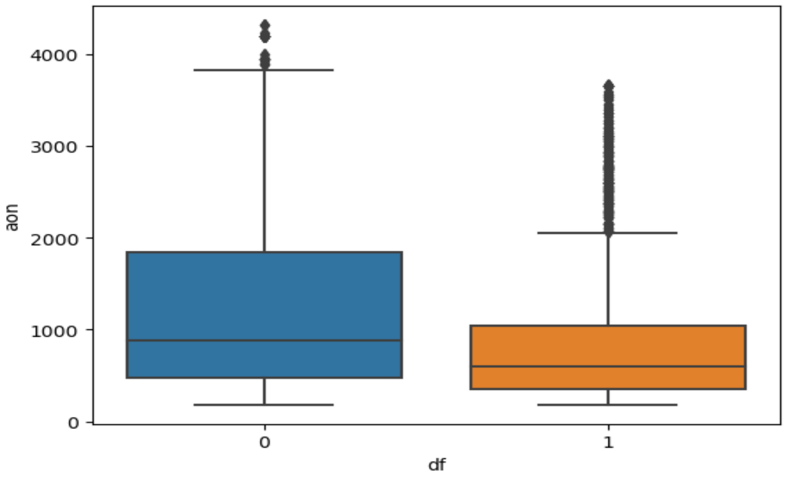
- **Derive new features** This is one of the most important parts of data preparation since good features are often the differentiators between good and bad models. We will use our business understanding to derive features that we think could be important indicators of churn.
- **Filter high-value customers** As mentioned above, we need to predict churn only for the high-value customers. Define high-value customers as follows: Those who have recharged with an amount more than or equal to X, where X is the 70th percentile of the average recharge amount in the first two months (the good phase).
- **Tag churners and remove attributes of the churn phase** Now tag the churned customers (churn=1, else 0) based on the fourth month as follows: Those who have not made any calls (either incoming or outgoing) AND have not used mobile internet even once in the churn phase. The attributes we need to use to tag churners are:
- total_ic_mou_9
- total_og_mou_9
- vol_2g_mb_9
- vol_3g_mb_9
- After tagging churners, we need to remove all the attributes corresponding to the churn phase (all attributes having '_9', etc. in their names).

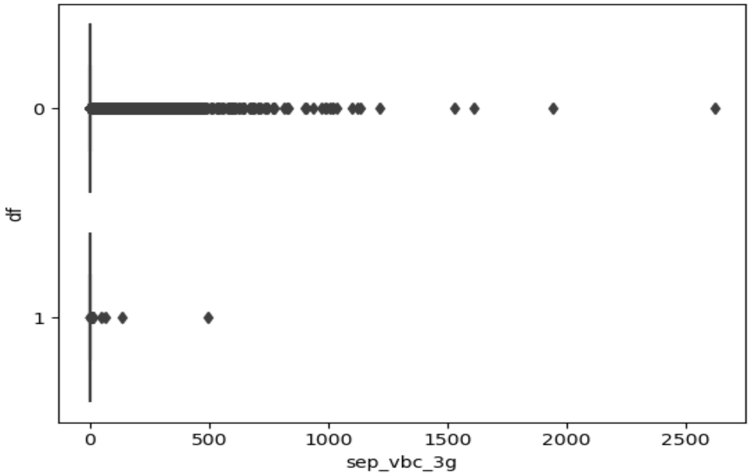
EDA

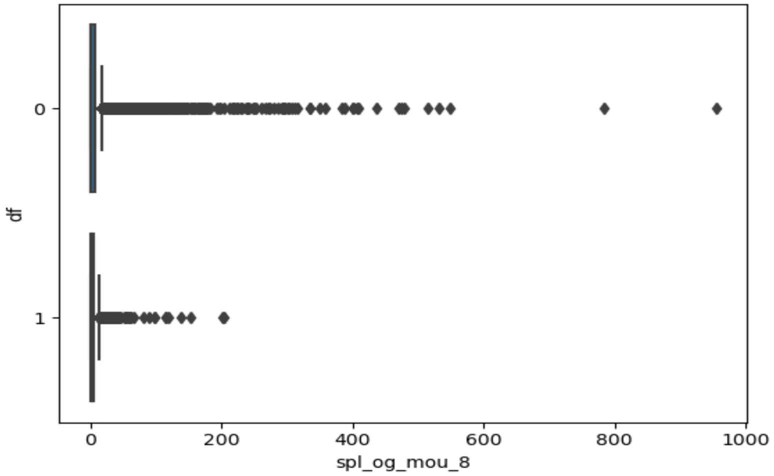
Some of the features



Bivariate analysis
Some of the features







Modelling

- Build models to predict churn. The predictive model that we are going to build will serve two purposes:
- It will be used to predict whether a high-value customer will churn or not, in near future (i.e. churn phase). By knowing this, the company can take action steps such as providing special plans, discounts on recharge etc.
- It will be used to identify important variables that are strong predictors of churn. These variables may also indicate why customers choose to switch to other networks.

Evaluation

Before hyperparameter tunning

Confussion Matrix

```
[[5488 1403]
 [  92  518]]
Sensitivity:    0.85
Specificity:    0.8
AUC:             0.89
```

Evaluation

After hyperparameter tunning

Confussion Matrix

```
[[5763 1128]
 [ 105   505]]
Sensitivity:    0.83
Specificity:    0.84
AUC:             0.9
```

Evaluation

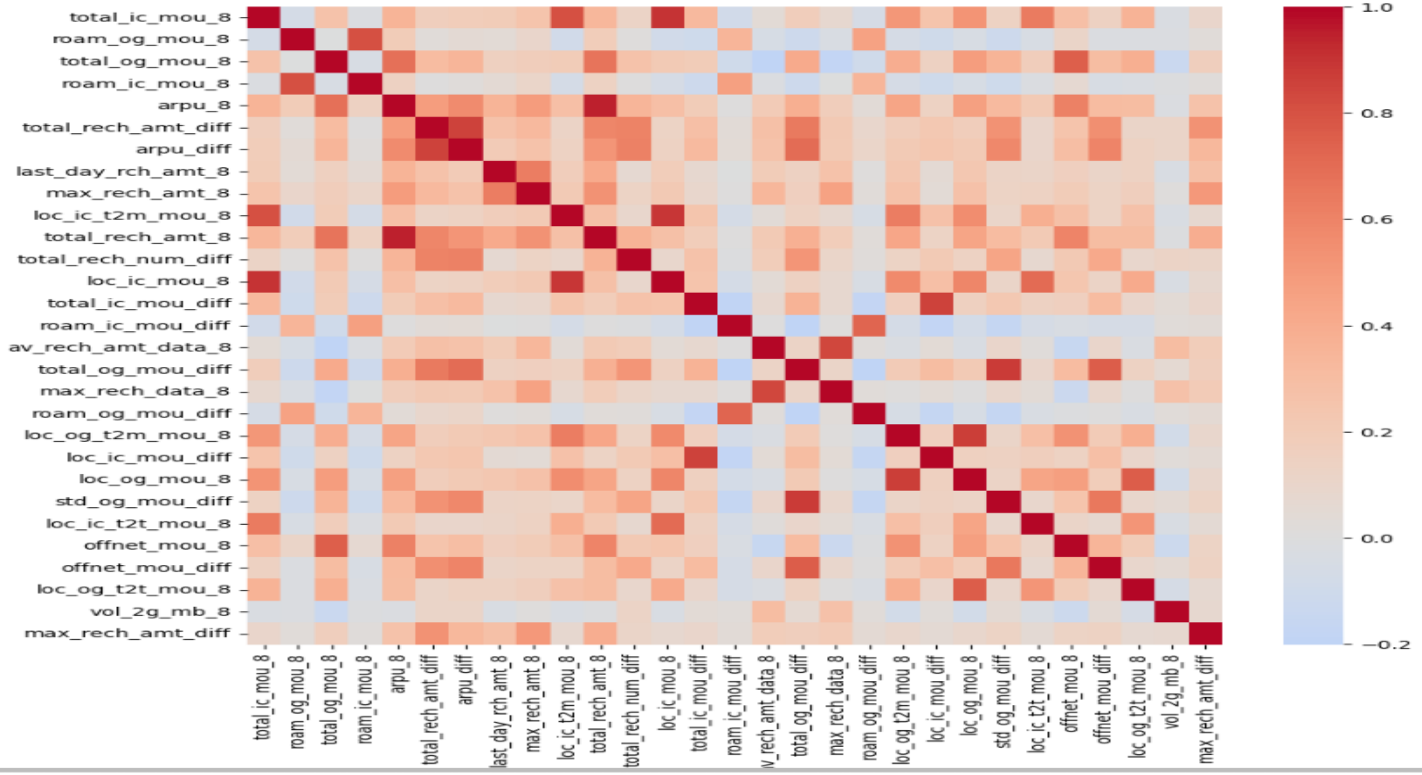
Using Model Random Forest

Confussion Matrix

```
[[6782  109]
 [ 311  299]]
Sensitivity:    0.49
Specificity:    0.98
AUC:             0.93
```

PCA and Logistic Regression

Heat map Of Important features



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.
- * To look into both of the issues stated above, it is desired that the telecom company collects customer query and complaint data and work on their services according to the needs of customers.

End Of Slide