Telecom Case Study

Presented by

Rina Dinda

Sowparni R

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, we 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.

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 behaviour during churn will be helpful.

Data Preparation

- **Derive new features**
- **Filter high-value customers**
- **Tag churners and remove attributes of the churn phase**

- total_ic_mou_9
- total_og_mou_9
- vol_2g_mb_9
- vol_3g_mb_9

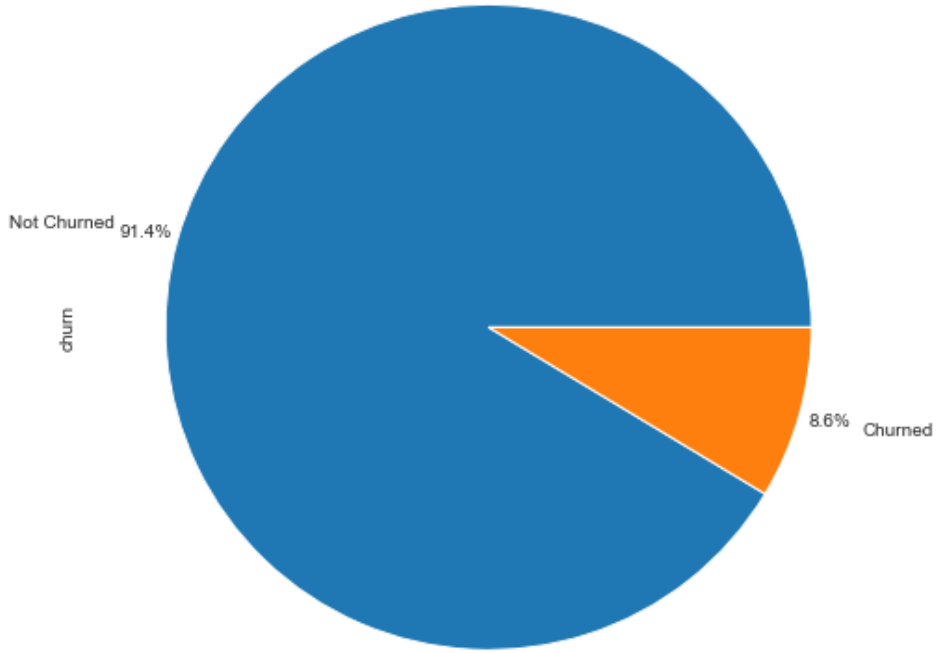
Modelling

- Build models to predict churn. The predictive model that we are going to build will serve two purposes:

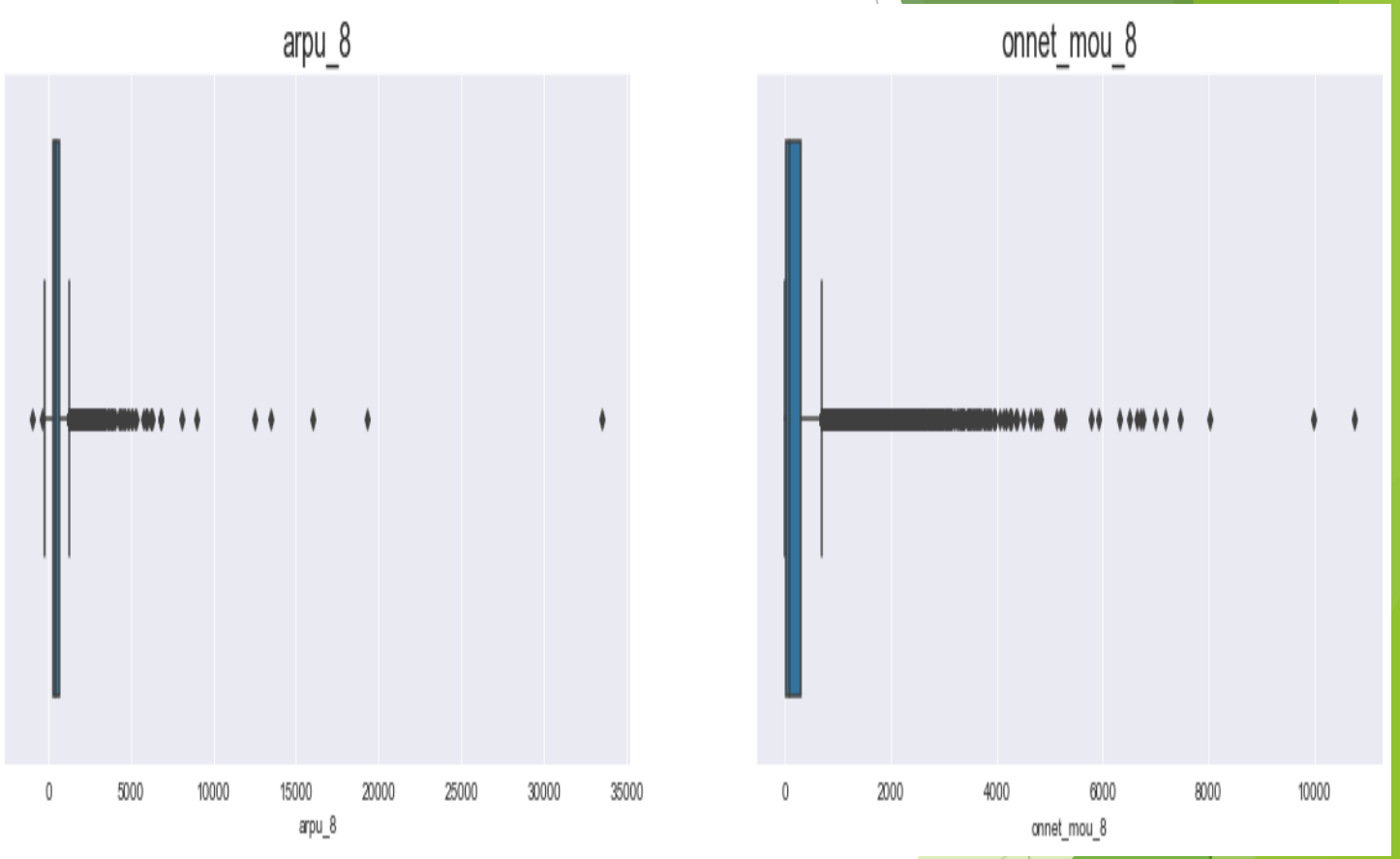
1. It will be used to predict whether a high-value customer will churn or not, in near future (i.e. churn phase). By knowing this, the company can take action steps such as providing special plans, discounts on recharge etc.
2. It will be used to identify important variables that are strong predictors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

"Churn" feature is highly skewed. To balance this column, will use "class_weight" command during the modeling process instead of using Undersampling/Oversampling.

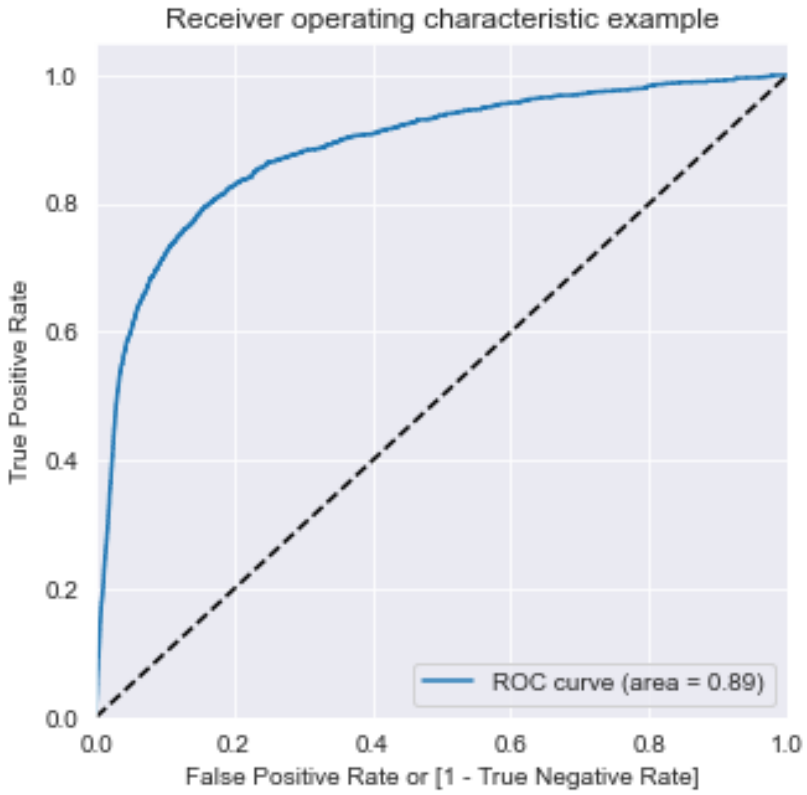
For easier exploratory analysis, deriving some new features. Taking the average of the first 2 months i.e. Month 6 and Month 7, and will derive a new column for each feature using that. The new feature will be that of the good phase.



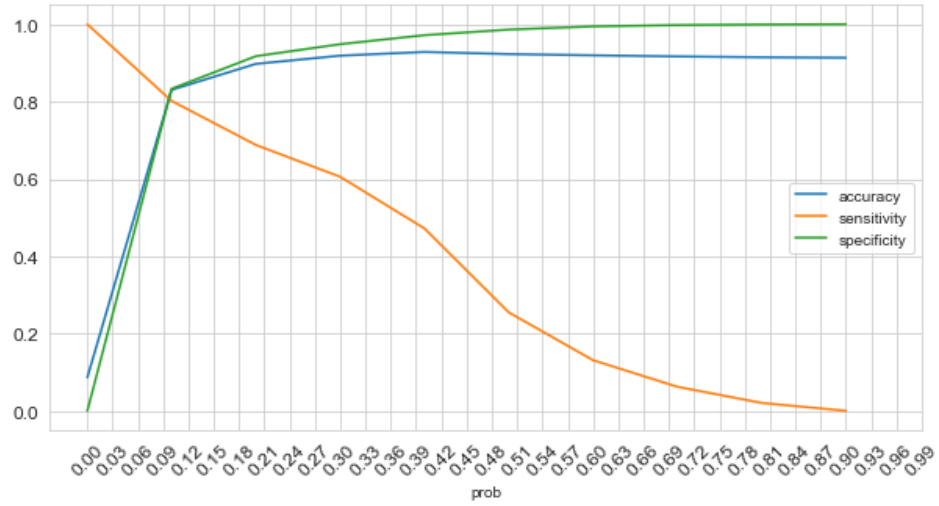
It seems every column has outliers on the higher side. So instead of removing these rows, capping the outliers to the 95th percentile value.



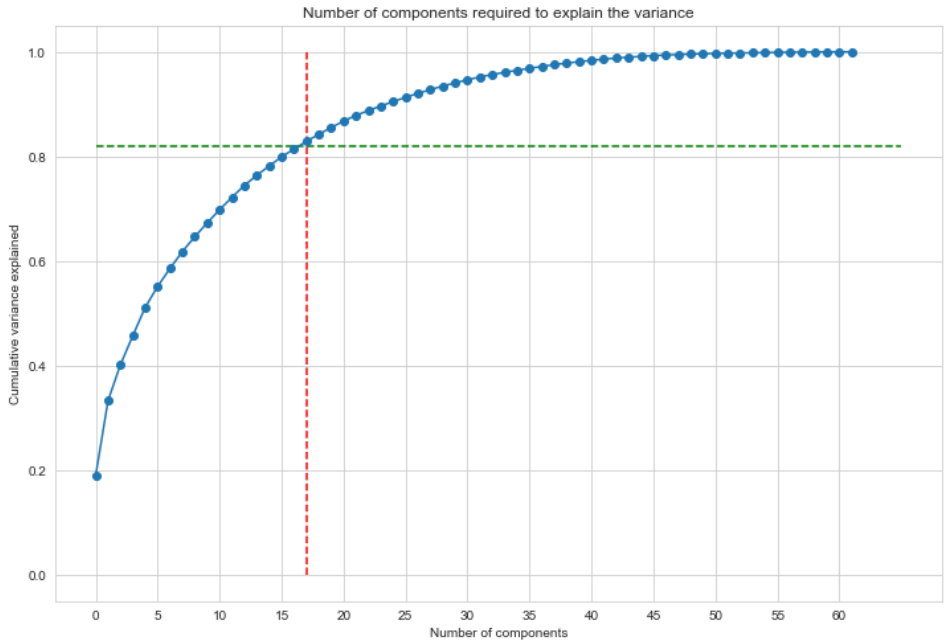
Area under the curve is 0.89, which is very good. Sensitivity is really low, so will now find the optimal cut-off point to get better accuracy

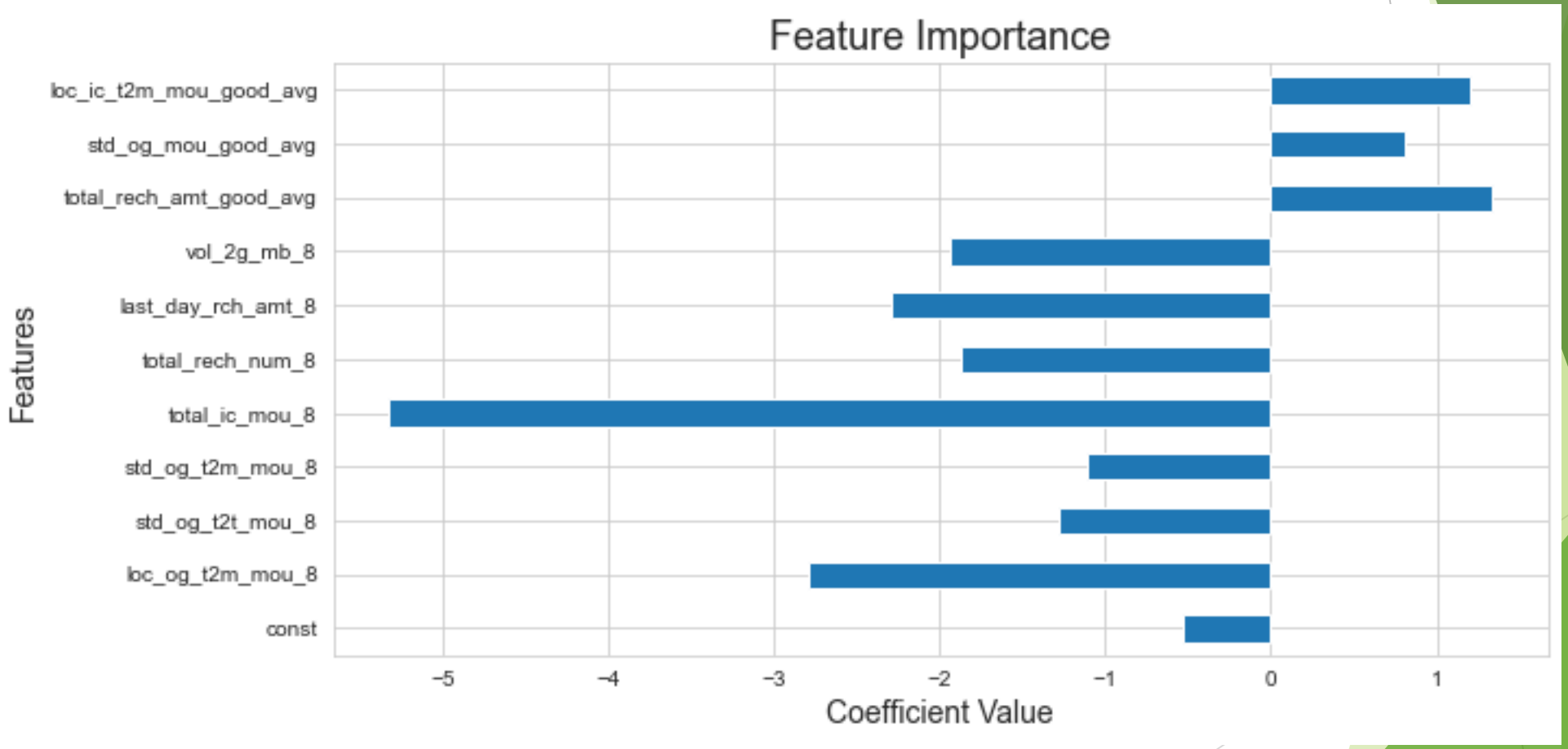


From above graph, it can be observed that the optimal cut-off point seems to be 0.1



So the optimal number of PCA is 17.




Feature Importance
loc_ic_t2m_mou_good_avg
std_og_mou_good_avg
total_rech_amt_good_avg
vol_2g_mb_8
last_day_rch_amt_8
total_rech_num_8
total_ic_mou_8
std_og_t2m_mou_8
std_og_t2t_mou_8
loc_og_t2m_mou_8
const
Features
−5
−4
−3
−2
−1
0
1
Coefficient Value

Conclusion and Final Remarks

- 1 interpretable model was created to see what factors are the most important for a customer to decide if they'd like to churn or not.
- 3 high performance models were created using: Logistic Regression, Random Forest, and XGBoost. All the 3 models gave fairly high accuracy for training and test datasets.
- ***Random Forest Model*** works best on the data given and will be best to predict the future customers who could possibly churn. This model was **95% accurate** on the training set, and on the unseen test dataset accurate of **91%** was received, which is very high and good. Random Forest was also less computationally less expensive.
- From the **interpretable model** , the top 3 important features:
  - total_ic_mou_8 : -5.3295
  - loc_og_t2m_mou_8 : -2.7948
  - last_day_rch_amt_8 : -2.2981