Telecom Churn Case Study – Advanced Machine Learning

By Priya Bakhtani and Arpita Patil

Content

Business Problem Understanding

- 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 telecommunication experiences an average of 15-25 % annual churn rate.

- 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.

- 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 the prepaid model, customers who want to switch to another network can simply stop using the services without any notice and its hard to know whether someone has actually churned or is simply not using the services temporarily which is the most common model in India and Southeast Asia.

Objective

- In the Indian and Southeast Asian markets, approximately 80% of revenues comes from the top 20% customers. Thus we have reduced the churn of high – value customers, we will be able to reduce significant revenue leakage.

- The dataset contains customer level information for a span of four consecutive months – June, 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.

- Since you are working over a four – month window, the first 2 months are the good phase the third month is the action phase and the fourth month is the churn phase.

- For the retaining high profitable customers, we need to analyze customer – level data of a leading telecom firm, build predictive models to identify customers at a high risk churn and identify the main indicators of churn.

Solution Methodology

- **Data Cleaning and data manipulation**.
  1. Check and handle duplicate data.
  2. Check and handle NA values and missing values.
  3. Drop Columns, if it contains large amount of missing values and useful for the analysis.
  4. Imputation of the values, if necessary.
  5. Check and handle outliers in data.

- **EDA**
  1.1. Univariate data analysis: value count, distribution of variable etc.
  2.2. Bivariate data analysis: Correlation coefficient and pattern between the variables etc.

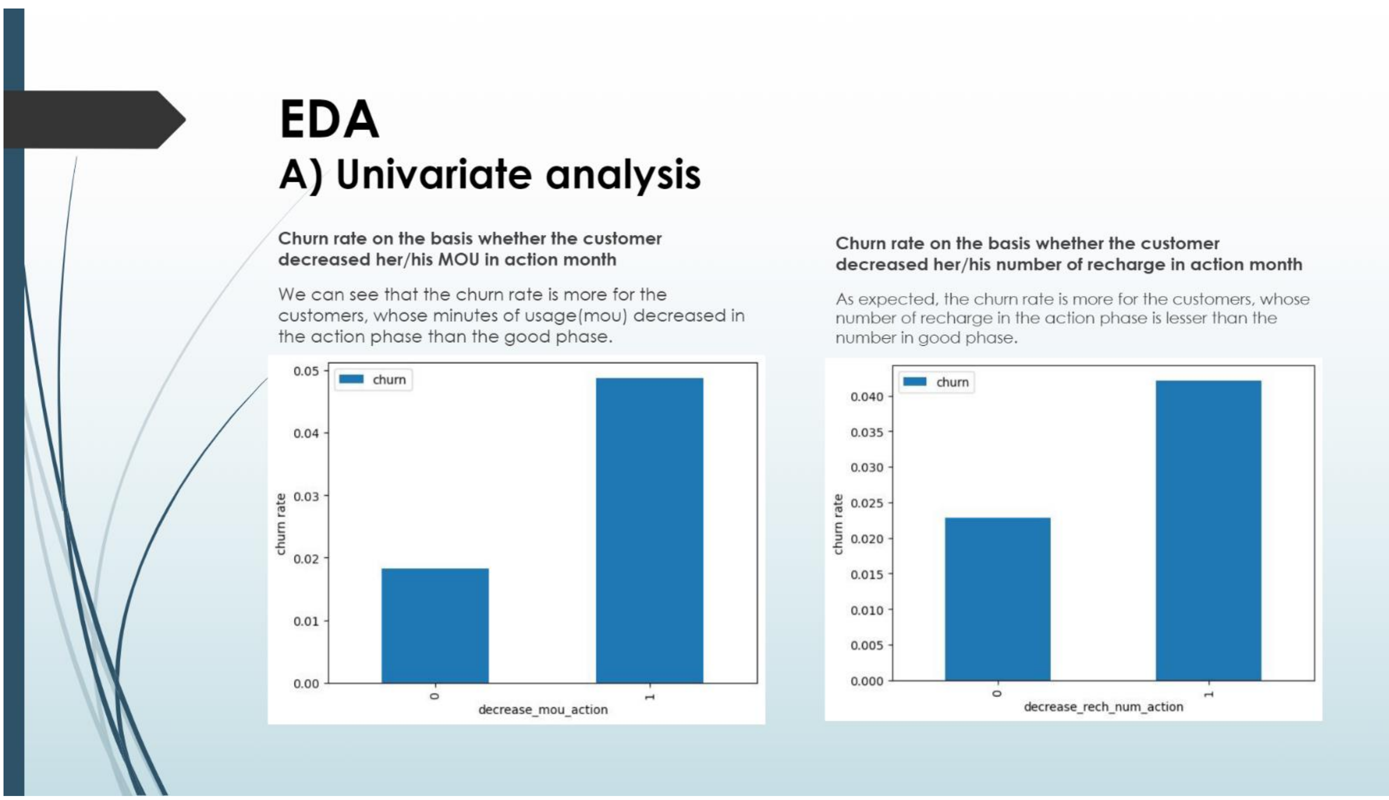
- Dealing with data imbalance, features scaling & dummy variables, encoding of the data,
- Classification technique: logistic regression, decision tree and random forest are used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusion and recommendations.

EDA

A) Univariate Analysis:

Churn rate on the basis whether the customer decreased her/his MOU in action month

We can see that the churn rate is more for the customers, whose minutes of usage (mou) decreased in the action phase than the good phase.

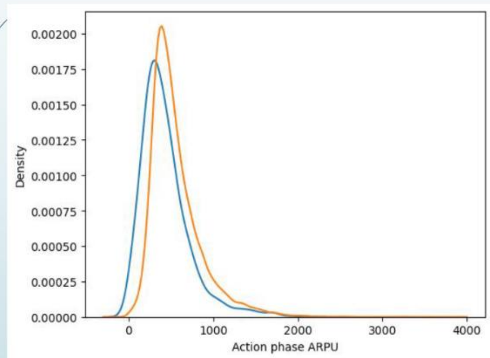


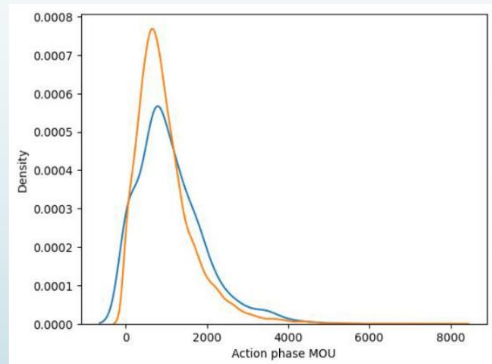
Analysis of the average revenue per customer (churn and not churn) in the action phase :

Below, Average revenue per user (ARPU) for the churned customers is mostly dense on the 0 to 900. The higher ARPU customers are less likely to be churned.

ARPU for the not churned customers is mostly dense on the 0 to 1000.

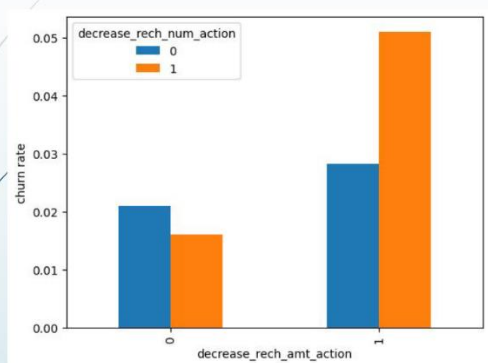
Below, Minutes of usage(MOU) of the churn customers is mostly populated on the 0 to 2500 range. Higher the MOU, lesser the churn probability.

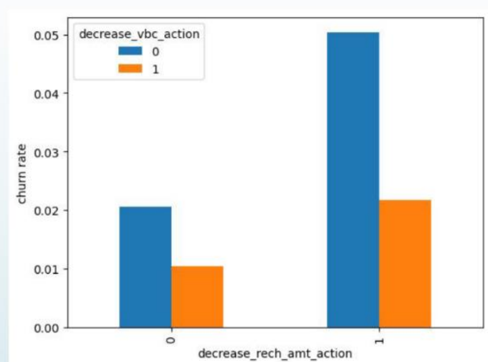




B) Bivariate analysis:

Analysis of churn rate by the decreasing recharge amount and number of recharge in the action phase





We can see from the above plot, that the churn rate is more for the customers, whose recharge amount as well as number of recharge have decreased in the action phase than the good phase.

Here, also we can see that the churn rate is more for the customers, whose recharge amount is decreased along with the volume based cost is increased in the action month.

Business Recommendations

Top Predictors:

- Besides are few top variables selected in the logistic regression model.
- We can see most of the top variables have negative coefficient. That means, the variables are inversely correlated with the churn probability.
- Eg:

If the local incoming minutes of usage (loc_ic_mou_8) is lesser in the month of August than any other month, then there is higher chance that the customer is likely to churn.

Recommendations

- Target the customers, whose minutes of usage of the incoming local calls and outgoing ISD calls are less in the action phase.
- Target the customers whose outgoing others charge in July and incoming others on August are less.
- Also, the customers having value-based cost in the action phase increased are likely to churn than the other customers. Hence, these customers may be a good target to provide offer.

- Customers, whose month 3G recharge in August is more likely to be churned.
- Customers have decreased STD minutes of usage for operators T to fixed lines of T for the month of August are more likely to churn.
- Customers decreasing monthly 2G usage of August are more probable to churn.