Telecom Churn Case Study Summary

Group partners :

Rakesh

Shawn

Tirtha

Analysis Approach:

- 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 become even more important than customer acquisition.
- Here we are given with 4 months of data related to customer usage. In this case study, we 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.
- Churn is predicted using two approaches. Usage based churn and Revenue based churn. Usage based churn:
- Customers who have zero usage, either incoming or outgoing - in terms of calls, internet etc. over a period of time.
- This case study only considers usage based 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. Hence, this case study focuses on high value customers only.
- 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.
- This is a classification problem, where we need to predict whether the customers is about to churn or not. We have carried out Baseline Logistic Regression, then Logistic Regression with PCA, Decision Tree, ADA Boosting, Random Forest.

Analysis Steps

Data Cleaning and EDA

1. We have started with importing Necessary packages and libraries.
2. We have loaded the dataset into a dataframe.
3. We have checked the number of columns, their data types, Null count and unique value_value_count to get some understanding about data and to check if the columns are under correct data-type.
4. Checking for duplicate records (rows) in the data. There were no duplicates.
5. Since 'mobile_number' is the unique identifier available, we have made it our index to retain the identity.
6. Have found some columns that donot follow the naming standard, we have renamed those columns to make sure all the variables follow the same naming convention.
7. Following with column renaming, we have dealt with converting the columns into their respective data types. Here, we have evaluated all the columns which are having less

than or equal to 29 unique values as catrgorical columns and rest as contineous columns.

8. The date columns were having 'object' as their data type, we have converted to the proper datetime format.
9. Since, our analysis is focused on the HVC(High value customers), we have filtered for high value customers to carry out the further analysis. The metric of this filtering of HVC is such that all the customers whose 'Average_rech_amt' of months 6 and 7 greater than or equal to 70th percentile of the 'Average_rech_amt' are considered as High Value Customers.
10. Checked for missing values.
11. Dropped all the columns with missing values greater than 50%.
12. We have been given 4 months data. Since each months revenue and usage data is not related to other, we did month-wise drill down on missing values.
13. Some columns had similar range of missing values. So, we have looked at their related columns and checked if these might be imputed with zero.
14. We have found that 'last_date_of_the_month' had some misisng values, so this is very meaningful and we have imputed the last date based on the month.
15. We have found some columns with only one unique value, so it is of no use for the analysis, hence we have dropped those columns.
16. Once after checking all the data preparation tasks, tagged the Churn variable(which is our target variable).
17. After imputing, we have dropped churn phase columns (Columns belonging to month - 9).
18. After all the above processing, we have retained 30,011 rows and 126 columns.
19. Exploratory Data Analysis

- The telecom company has many users with negative average revenues in both phases. These users are likely to churn.
- Most customers prefer the plans of '0' category.
- The customers with lesser 'aon' are more likely to Churn when compared to the Customers with higer 'aon'.
- Revenue generated by the Customers who are about to churn is very unstable.
- The Customers whose arpu decreases in 7th month are more likely to churn when compared to ones with increase in arpu.
- The Customers with high total_og_mou in 6th month and lower total_og_mou in 7th month are more likely to churn compared to the rest.
- The Customers with decrease in rate of total_ic_mou in 7th month are more likely to churn, compared to the rest.
- Customers with stable usage of 2g volume throughout 6 and 7 months are less likely to churn.
- Customers with fall in usage of 2g volume in 7th month are more likely to Churn.
- Customers with stable usage of 3g volume throughout 6 and 7 months are less likely to churn.
- Customers with fall in consumption of 3g volume in 7th month are more likely to Churn.
- The customers with lower total_og_mou in 6th and 8th months are more likely to Churn compared to the ones with higher total_og_mou.
- The customers with lesser total_og_mou_8 and aon are more likely to churn compared to the one with higher total_og_mou_8 and aon.
- The customers with less total_ic_mou_8 are more likely to churn irrespective of aon.
- The customers with total_ic_mou_8 > 2000 are very less likely to churn.

1. Correlation analysis has been performed.
2. We have created the derived variables and then removed the variables that were used to derive new ones.
3. Outlier treatment has been performed. We have looked at the quantiles to understand the spread of Data.
4. We have dropped unwanted columns which did not add values to modelling.
5. We have checked categorical variables and contribution of classes in those variables. The classes with less contribution are grouped into 'Others'.
6. Dummy Variables were created.

Pre-processing Steps

1. Train-Test Split has been performed.
2. PCA with modelling, PCA evaluation.
3. Logistic Regression with RFE and VIF

Modelling

1) At first Total 5 Models have been created to brought the P values less than .005.
2) Checked VIF value , there we got VIF a very high VIF 12.73 thereby dropped the variable " local_any_8"
3) AT model 6 we got P values less than .005 and VIF less than 5.
4) With, Cut-off values of 0.1 following values obtained:
   a) Sensitivity of the Model: 0.30
   b) Specificity of the Model: .84
   c) Precision of the Model: .73

5) Decision Tree:

a) we increase the value of max_depth, both training and test score increase till about max-depth = 4, after which the test score gradually reduces. Note that the scores are average accuracies across the 5-folds.

b) We have see that at values > 125, the test and train score starts to converge and hence the model starts becoming more stable.

c) We notice that with values > 125, the train and test accuracy start to converge hence making the model more stable and less complex.

6) After ADA Boosting

a) We have observed accuracy score of .94.

7) Random forest:

a) We can get accuracy of 0.9440476190476191 using {'max_depth': 16, 'max_features': 10 , 'min_samples_leaf': 10, 'min_samples_split': 10, 'n_estimators': 10}

Recommendations

We notice that the following 5 factors affect the churn rate considerably -

1) Total Incoming Minutes of usage in the August

2) Total Incoming Minutes of usage in the July

3) 2G data pack

4) Roaming

5) Sachet 2g

Also these metrics are inversely proportion to churn which means that we need to come up with campaigns that would keep people engaged either via calls (incoming) or on internet. One interesting thing to note here is that we see that a lot of people are hooked on 2G and hence are not churning. It presents us with a great opportunity that if shift these people from 2g to 3g then we have a greater chance of these people not churning. Hence discounts on 3G pack can be one of the popular marketing campaigns.