CAPSTONE PROJECT REPORT

ON

CHURN PREDICTION

Submitted by

Ritesh Mishra
Chandan Nishad

ABSTRACT:-

Customer churn is the focal concern of most companies due to its associated costs and telecommunications industry can be considered in the top of the list with an approximate annual churn rate of almost 30%. Most companies favor the creation and nurturing of long-term relationships with customers because retaining customers is more profitable than acquiring new ones. As the market is very competitive, it is inevitable that companies proactively tackle the defection of their customers by determining behaviors that might ultimately lead to churn. Churn prediction is a predictive analytics technique to identify churning customers ahead of their departure and enable organizations to take action to keep them. Our brief report explains an efficient model on customer churn prediction for telecom services. The model with a decent accuracy will predict customers who will change and turn to another provider for the same or similar service. Sample dataset we have used for our project has been compiled by a Telecommunication company. The result shows that modeling on Generalized Linear Model(GLM) algorithm has good prediction effect.

INTRODUCTION:-

Among industries that are mostly hit by churn, Mobile Telecommunication comes on top. The causes are among others: the technological development of telecommunications, the liberalization, the globalization and fierce competition. An increasing churn rate is considered as the plague of all carriers because losing a high valued client means a loss of future incomes. Also, the cost of winning a new customer is 5 to 10 times higher than the cost of satisfying and keeping an existing customer. Therefore, customer retention has become the prime objective of marketing campaigns. It's more profitable for mobile telecom operators to invest in those customers that already have an experience with the service by renewing their trust, rather than constantly trying to attract new customers characterized by a higher churn rate.

Tremendous volume of data is being generated from mobile telecom networks; however, customer's data is complex and under privacy regulation, whereby very limited information about the subscriber or his experience with the service is available. Our predictive model uses Logistic Regression with a real world dataset.

DATASET:-

The dataset contain CSVs for pay, data, usage spread across ten months. Besides, there were CSVs having account related, request information and disconnection data. The original customer names were masked in order to ensure privacy.

METHODOLOGY:-

The methodology we used to develop a churn prediction model is the generalized linear model using logistic regression. The data exploration gave us a good idea about how the data looked. We pre-processed the data to remove outliers, impute missing values and got cues towards feature engineering. We used R for data exploration, pre-processing and model building. Hadoop / Spark was used for fast computation .

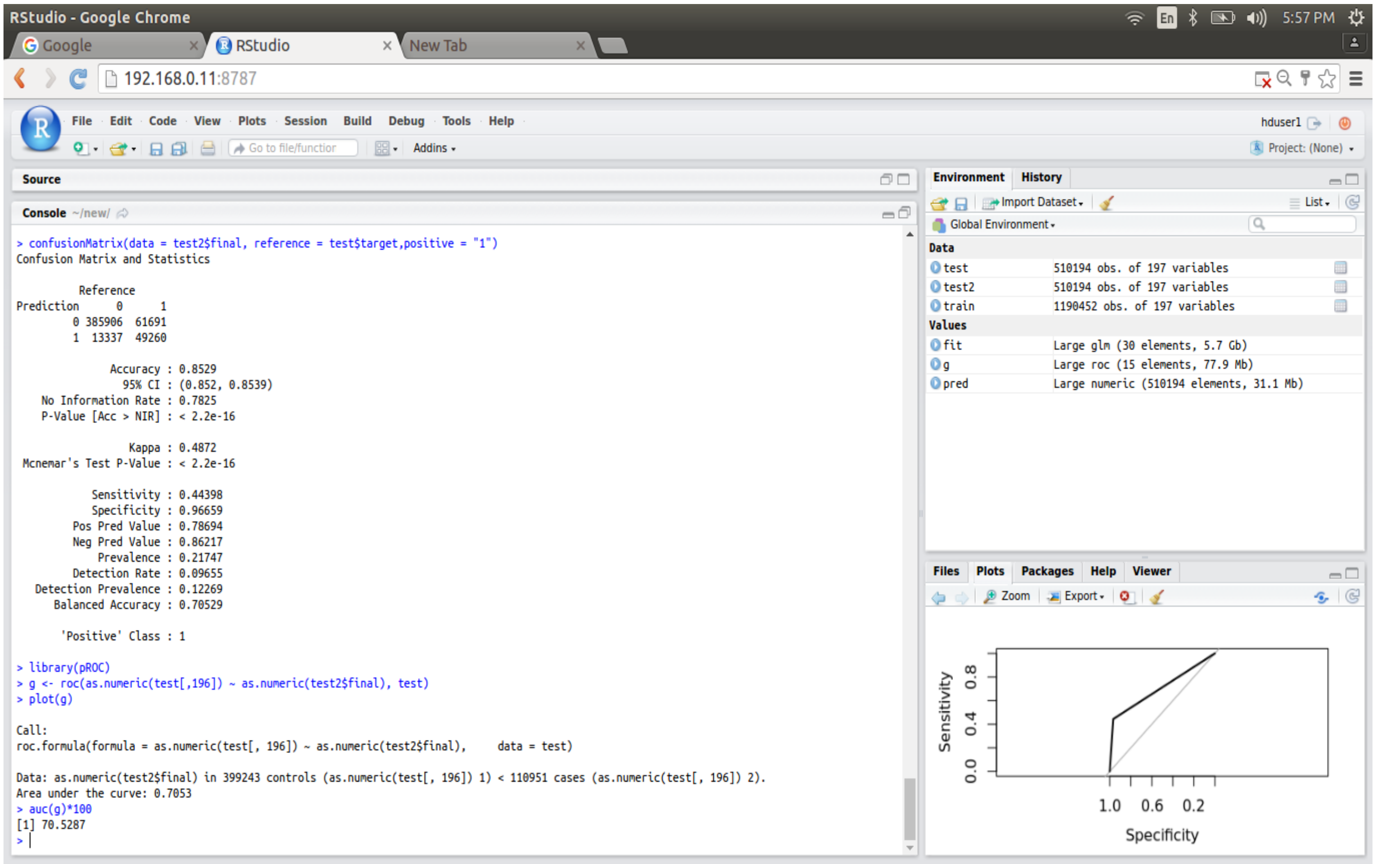
The different steps taken are as follows:-

1) For the model to draw a pattern, continuous data was required. Hence we took five months of continuous data.
2) We stored the data i.e. different CSVs in Hadoop distributed file system(HDFS) and each month's data was fetched and imported to R.
3) For July data we read the pay, usage and data CSVs into R. We merged these three files by the common column "id".
4) If all are NAs in a particular row, we deleted that row because of lack of data for our convenience.
5) After that we removed the duplicate records in the July file(as there were ids getting repeated)
6) We found out the outliers using the Q3+1.5*IQR, Q1-1.5*IQR and imputed them with NA.
7) NAs in every column were replaced with median, mean and mode for numerical data, integer data and categorical data respectively.
8) New variables were formed from the existing variables. The NaNs were replaced with 0.
9) Next we changed the date format in the activation date and imputing all the NA in this column with the first date of July.
10) The august data was fetched and the same process was repeated. Finally July and August data were merged.
11) There were two activation date columns. So we transferred the activation dates of August to July and dropped the activation date column of August.
12) We fetched September data, did all the steps of preprocessing and merged with merged July and august data.
13) After transferring the activation column of September to merged July and August data, we dropped the activation column of September.
14) We repeated the same process for October and November. We dropped the rows with all NAs and imputed the remaining NAs with 0.
15) Then we read the disconnection file. We assigned 1 to churn and merged with our five months data. The NAs after merge were imputed with 0.
16) There was a date column in the disconnection file. After merge, the NAs of this column was replaced with the last date of November as this date is the last date for our consideration.

17) We removed all the rows with NAs and calculated the no of days the customer is associated with the system.
18) After dropping few irrelevant columns, we had the single view of our data for developing the model.
19) We divided the data 70% and 30% into training(1190452 rows and 197 columns) and test data with 510194 rows respectively.
20) We used logistic regression and developed the model. We ran the code in Spark cluster for fast computation.

RESULT:-

The result is as follows:-

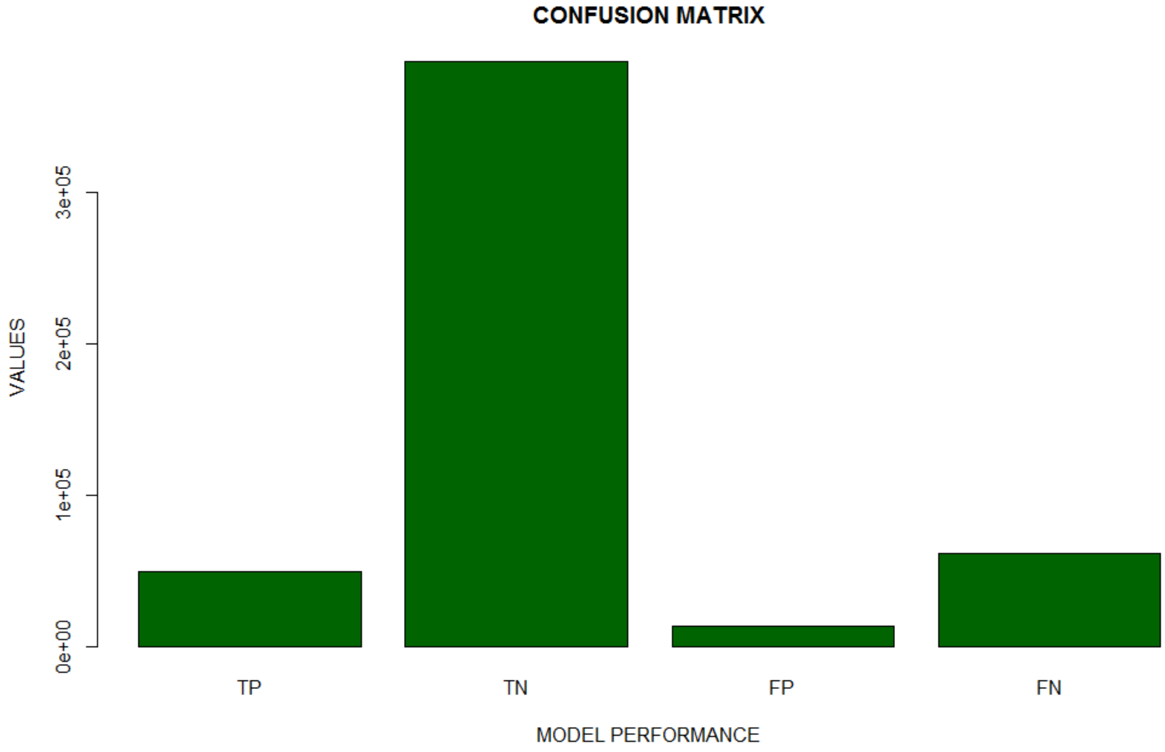


```
> confusionMatrix(data = test2$final, reference = test$target,positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction      0      1
         0 385906  61691
         1  13337  49260

               Accuracy : 0.8529
                 95% CI : (0.852, 0.8539)
    No Information Rate : 0.7825
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4872
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.44398
            Specificity : 0.96659
         Pos Pred Value : 0.78694
         Neg Pred Value : 0.86217
             Prevalence : 0.21747
         Detection Rate : 0.09655
   Detection Prevalence : 0.12269
      Balanced Accuracy : 0.70529

       'Positive' Class : 1

> library(pROC)
> g <- roc(as.numeric(test[,196]) ~ as.numeric(test2$final), test)
> plot(g)

Call:
roc.formula(formula = as.numeric(test[, 196]) ~ as.numeric(test2$final),     data = test)

Data: as.numeric(test2$final) in 399243 controls (as.numeric(test[, 196]) 1) < 110951 cases (as.numeric(test[, 196]) 2).
Area under the curve: 0.7053
> auc(g)*100
[1] 70.5287
> |
```

```
Data
test     510194 obs. of 197 variables
test2    510194 obs. of 197 variables
train    1190452 obs. of 197 variables
Values
fit      Large glm (30 elements, 5.7 Gb)
g        Large roc (15 elements, 77.9 Mb)
pred     Large numeric (510194 elements, 31.1 Mb)
```

We got a decent accuracy of 85.29% with Area under the curve as 70.53

RESULT PLOT:-



CONCLUSION:-

This project presented a predictive model for Churn in telecommunication companies that predict customers with high propensity to churn. In comparison with other classifiers, this model predicts with a decent accuracy. Using this model, companies can arrest churn by devising new and personalized marketing campaigns pertaining to the benefit of a customer.