ucc

Coláiste na hOllscoile Corcaigh, Éire
University College Cork, Ireland

CS 6405: DATA MINING

PROJECT REPORT

Team Members:

118221508 - Freda Pinto

118220940-Rohit Badgujar

Course: MSc Computing Science 2018-19

1. Introduction

Customer retention is a challenge in the ultra-competitive mobile phone industry. A mobile phone company is studying factors related to customer churn, a term used for customers who have moved to another service provider.

The sample data set is of telecom industry consists of 4000 customer records. The target variable of interest is the column called Churn, which takes two values: 1: The customer has moved to another service provider. 0: The customer still using the same service. The churn has two types of features they are numeric features and categorical features and together they comprise of 21 features.

The categorical features possess only two categories as they are binary values. And there are in total 2 such features. The remaining features are having numeric values which can be any integer value. For the given dataset, it has been divided into holdout and train set respectively. And, the objective is to train the model using train data which has 400 customers and test on holdout instances which are 1000 customers. The test analysis performed on holdout will give us predictions from which we can conclude which model is better. the goal of the model is to predict whether the customer is likely to churn or not. Before finalizing the method, we ran data through different regression and classification algorithm will help in better mining and understanding the data.

2. **Methods**

The algorithms used are Logistics Regression, Decision Tree Classifier, and Random Forest Classifier. As the data is highly unbalanced in terms of customers retention and customer churn.

We tuned the model by changing the value of input parameter random_state for Logistic Regression and *n_estimators, random_ state, max_depth* for Random Forest Classifier. *n_estimators* represents the number of trees in the forest. Usually the higher the number of trees the better to learn the data. However, adding a lot of trees can slow down the training process considerably, therefore we do a parameter search to find the sweet spot between 50-60. The parameter *max_depth* decides the maximum depth of the tree. We have tried values between 10-20 for max_depth. To understand the performance of the model we have plotted the ROC-AUC Curve. ROC is a probability curve and AUC represent degree or measure of separability. ROC curves illustrate the performance of a classifier by considering the true positives and false positives at various cut-offs. The Area Under Curve (AUC) measures the model's ability to correctly classify those who churned and those who did not. The higher the AUC the better.

In Machine Learning, Data preprocessing is a vital step as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn. Therefore, it is extremely important that we preprocess our data before feeding it into our model. Pre-processing essentially focuses on two problems: (1) data quality and (2) data representation. Data preprocessing helps us to understand inconsistencies or discrepancies in our data. Initially, we identified the missing or null values in the data set. In order to decide the features of data-set

to be included in our prediction model, we used heatmap to examine the correlation between churn and each customer feature. We have discarded some features that showed a strong correlation for example,total_intl_charge and total_night_charge. Also, the unique feature phone_number is discarded as it does not provide us the information we can learn from. The major problem with the customer churn dataset is that the many features have completely different ranges, for example, *international_plan* has value {0,1} and *total_intl_charge* has value {0 to hundreds}. Such a difference may cause a problem in some model. To account this we have applied Standardization technique to our dataset.

Proper predictive models' evaluation is important because we want our model to have the same predictive ability across many different data sets. In other words, the results need to be comparable, measurable and reproducible, which are important factors. We have estimated predictive performance is estimated by using the confusion matrix for each model. Confusion matrix can give a better idea of what classification model is getting right and type of errors it is making.

The best model is decided on the basis of the ROC/AUC score generated as it is used for calculating the model's performance metric. We also tuned the model in order to achieve better performance and accuracy.

3. Results and Discussion

Estimates of predictive test performance of different models

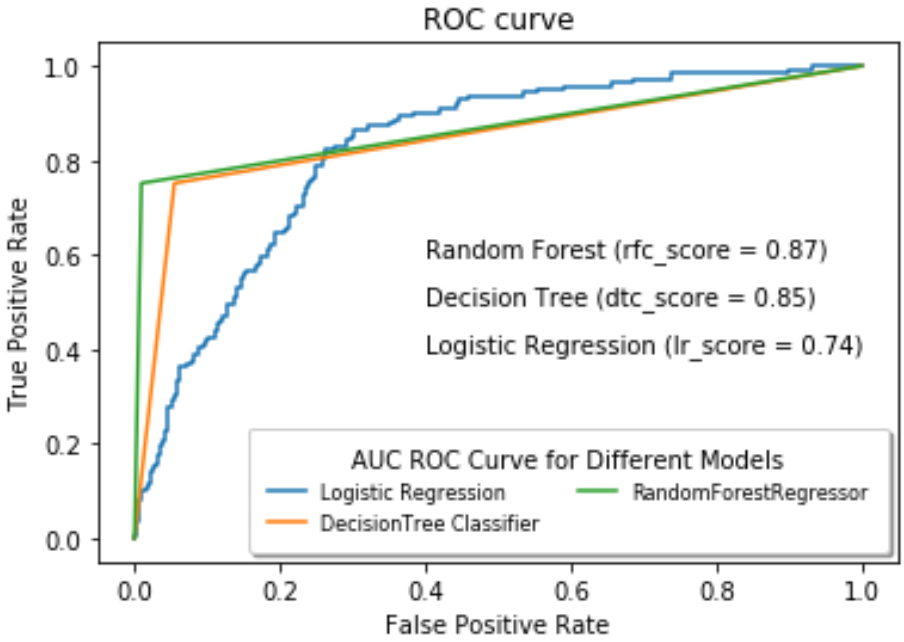

Figure 1 - AUC ROC Curve for different Models

3.1 Comparison of models

Logistic Regression is useful when data to be predicted falls under the classification category. It basically uses a logistics function to produce a binary output. With random_state set to 0, we have got the AUC-ROC score as 74%. The advantage of logistic regression over other models is it does not require any tuning. It is highly interpretable and does not require many computational resources.

On the other hand, Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. Accuracy score obtained for customer churn data set is 85%. The data type is not constrained in the Decision tree i.e. it can handle both categorical and numerical data. The disadvantage of using the decision tree is overfitting. It creates a complex

that does not generalize the data well. The decision tree is also unstable as a small variation in data might result in a completely different tree. It can get complex when we have many columns in the data set. Random forests solve the problem of overfitting. Random forests are considered as a highly accurate and robust method because of the number of decision trees participating in the process. For the given data set ROC-AUC score obtained for Random forest classifier is 87%, which is the best score obtained when compared with Logistic Regression and Decision Tree Classifier. Random Forest comprises of decision trees and the majority decision of forest is chosen as predicted output. Random Forest model is computationally slow and hard to interpret when compared with Decision Tree Model.

We have chosen Random Forest Classifier on the holdout data as it showed better accuracy score of 87% and performance metric when compared with another model. We have tuned the parameter (n_estimators = 70, random_state = 0,max_depth=20) to obtain best accuracy score.

This analysis shows the vital information about the various features and how this feature affects customer churn. With the help of this analysis, we can help the telecom industry to bring necessary changes, which will retain the customer. A large and high-quality set of data can help to build strong analysis. Proper Exploratory Data Analysis (EDA) and data preprocessing techniques can help to build a perfect training model. Appropriate model tuning can considerably improve the accuracy score.

The interesting thing to note about the analysis done is the rate of customer churn is highly dependent on the categorical values such as i*nternational_plan, voice_mail_plan,* etc. As shown in figure 2, high churn rate is observed among customers with international plans. Customers

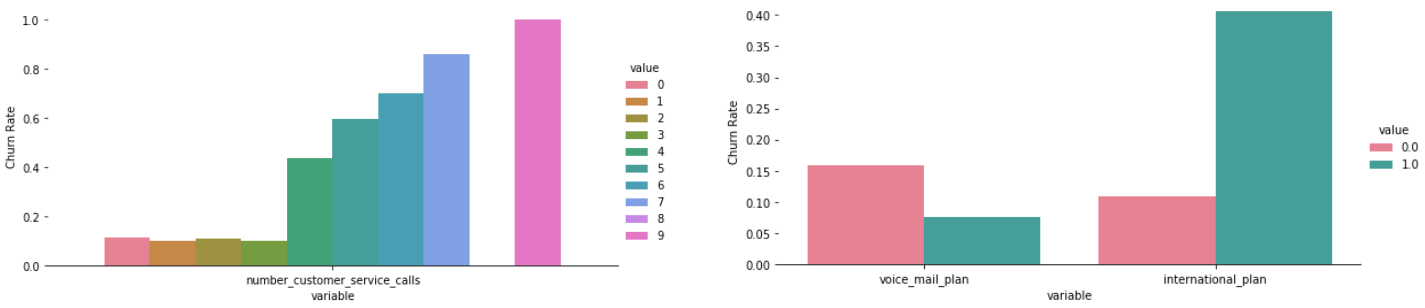

Figure 2 Churn Rate for number of customer service calls/ voicemail plan and international plan

with four or more service calls are more likely to leave the company. Companies should improve their service call centers to resolve customer issues in fewer than three calls.

3.2 Limitation

Customer behavior is dynamic, it may change over time. The model should be validated periodically to assess changes in customer behavior. Churn insights are based on correlating behaviors, correlations do not necessarily indicate causative relations. Interventions in customer behavior should be empirically tested.

4. Conclusion and Future Work

In this project we have studied the customer churn data set. The main aim of the analysis was to find out the pattern and features which lead to customer churn. We achieve these we have initially found out the important features and their relationship. We discarded the features which were not useful for the analysis. We preprocessed the data by removing the inconsistencies before further analysis. We split the data into train and test data frame. We evaluated performance model on three models. We also tuned the model parameter to achieve accurate

results. We obtained the ROC-AUC as shown in figure 1. With the best performance score of 87% we choose Random Forest Classifier as our predictive model for holdout data set. We then passed the holdout dataset as test data to Random forest classifier to predict the weather customer will churn or not with respective probabilities.