DATA MINING CASE STUDY – PROGRESS REPORT

ANALYZING CUSTOMER ATTRITION

Group 15 - Hansika Karkera, Vrunda Shah

Background

Customers are the life force of every business and in this competitive era finding a new customer is as difficult as retaining an old one. Hence analyzing customer data proves essential for a company to understand it's customer's need and reduce customer attrition. Customer attrition is the term used for loss of customers. One of the industries in which analyzing customer attrition proves to be profitable is the telecom industry. In this case we will be evaluating the customer information of Telco industry to analyze their customer defection and the factors that affect it.

Solution Design

Before building the regression model, we begin with data processing. Firstly, we study the summary statistics of the selected data set to check if there are any missing values. On evaluation we observe that the feature total charges have 11 missing values. Since it is a large data set, we will remove all the rows with a missing attribute value. Next, we wrangle the data and prepare it for further analysis by categorizing each attribute in types Yes or No.
Also, replace the numerical entry for tenure into 5 groups – 0-12 months, 12-24 months, 24-48 months, 48-60 months and > 60 months. To check the dependency between the numerical variable we plot a correlation plot. We remove the variables that are strongly correlated to avoid multicollinearity problem. For the categorical data, we plot pie charts to check the distribution of the categories in each attribute. Since we have enough entries for all the categories, there is no need to remove any attribute due to insufficient entries.

Post the above exploratory analysis, we split the data into training set and validation set for regression. Further on, we fit the training data to a generalized linear model since we have factors as our data entry. Using the p-value for the t-test in the summary statistics we figure out the features that are not significantly different from the regression coefficient, these are the most relevant features for the given model.

Algorithm Selection and Implementation

KKN Algorithm – We begin with using one of the simplest algorithms for classification. By analysis the test data set we have, we can classify that under which categories customer comes. And even we can classify using KNN that reasons for customer attrition.

Naïve Bayes – Here in Naïve Bayes algorithm we will use Bayes' Theorem to calculate the probability of the hypothesis obtained by our given dataset. For this case, here we can use naïve bayes algorithm to find the probability of customer leaving Telco industry.

Decision Tree – Using the relevant features from the above model , we implement the decision tree algorithm. Through this method we get a clear picture of the feature importance and feature relations. It would highlight the combination of features that would more likely result in a customer to discontinue services.

Random Forest – To address to the shortcomings of the above regression model we use the random forest algorithm. Initially we create a random forest data which is used to fit the model and rank the features as per importance.