Churn Rate

Churn rates are of interest to the financial sector and commercial businesses that rely on subscriptions and sale of services. Churning means cancellation of a subscription which is bad for the company. All companies would like to minimize their customers’ churn rates to maximize profit. Churn is a binary response with a customer either churned or not churned. The dataset has 31 features and 27000 observations. Given the data, the analysis tries to minimize churn rates by determining which feature contributes heavily to the customers cancelling their subscriptions.

Methods

In this analysis, logistic regression, decision tree, discrete, real, and gentle AdaBoosting along with random forest were used. R/Rstudio was used to perform the analysis. For the boosting technique, the data was divided into 2/3 training set and 1/3 test set with cross validation being used. For the random forest, the data was split 60% training set and 40% test set. tuneRF was a function used to determine how many features were optimal for the model which turned out to be 5 with the minimal OOB error. For decision tree, cross validation was used to grow a large tree, then 1SE rule was applied to prune the tree into the best subtree.