ISyE 6740 - Spring 2021
Project Proposal

Team Member Names: Rodrigo Silva Casarrubias

Project Title: Estimating churn in telecommunications industry

Problem Statement

The churn rate is defined as the percentage rate at which customers stop subscribing to a service within a given period. One of the main goals of businesses is to maintain and increase their active subscribers.

According to an article found in The European Business Review, telecommunications companies usually experience high rates of customer churn, which translates into companies incurring in high costs of losing customers constantly. Identifying which services are more valuable for customers in order to offer them better packages/services could help telecommunications companies reduce their costs.

(Optional) Data Source

The data source is one of the datasets found in **kaggle.com** found here. It contains information (22 variables) for 5986 users. Most of these variables are categorical variables, such as $PaymentMethod$, $PhoneService$, $InternetService$. The only non-categorical variables found in the dataset are $customerID$, $Tenure$, $MonthlyCharge$, and $TotalCharges$. A description of each variable is found on Table 1

Methodology

The dataset has a column called $Churn$, which identifies whether there was a churn or not. This means that a classification supervised method would be appropiate to analyze this dataset. The first step for this analysis will be to split the data between training and test set.

Having done the split, I am interested on performing one of SVM, naive Bayes, random forest, neural network or logistic regression models on this data, after performing feature selection by the information theoretic metric.

The reasoning behind the choice of models is that, according to a paper wrote by Adnan Amin et al. (2017), there have been several previous studies on churn prediction in telecommunications industry where the technique performed was one of those models. Also, performing feature selection before creating the models will help us to simplify the model, reduce overfitting, and maybe also to increase the accuracy of the models

Evaluation and Final Results

The evaluation of the models will be done comparing the confusion matrix and the accuracy of each of the models.

Table 1: Description of variables

Variable	**Description**
customerID	customer id
gender	client gender (male / female)
SeniorCitizen	is the client retired (1, 0)
Partner	is the client married (Yes, No)
tenure	how many months a person has been a client of the company
PhoneService	is the telephone service connected (Yes, No)
MultipleLines	are multiple phone lines connected (Yes, No, No phone service)
InternetService	client's Internet service provider (DSL, Fiber optic, No)
OnlineSecurity	is the online security service connected (Yes, No, No internet service)
OnlineBackup	is the online backup service activated (Yes, No, No internet service)
DeviceProtection	does the client have equipment insurance (Yes, No, No internet service)
TechSupport	is the technical support service connected (Yes, No, No internet service)
StreamingTV	is the streaming TV service connected (Yes, No, No internet service)
StreamingMovies	is the streaming cinema service activated (Yes, No, No internet service)
Contract	type of customer contract (Month-to-month, One year, Two year)
PaperlessBilling	whether the client uses paperless billing (Yes, No)
PaymentMethod	payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
MonthlyCharges	current monthly payment
TotalCharges	the total amount that the client paid for the services for the entire time
StreamingMovies	is the streaming cinema service activated (Yes, No, No internet service)
Churn	whether there was a churn (Yes or No)