STATS 765 – Project Milestone 1

Customer Churn Prediction

1 Objective

The objective is to predict the propensity of an existing customer to leave (churn) given business data, and to find the variables that are the most useful for predicting churn. In other words, to answer the question, “Is this customer predicted to churn, and if so, why?”

2 Data

The dataset comes from Orange, which was a telecoms operator during the 1990’s and 2000’s. The dataset was used in a KDD competition in 2009. There is a small and a large dataset. The small dataset has 230 variables and the large dataset has 15000 variables (1.88 GB). The large dataset has 50,000 training instances. Some variables are numerical, and some are categorical. Each dataset has one header line with the variable names, one line per instance, and a separator tabulation between the values. There are missing values (consecutive tabulations). All training data, including the labels, and some of the test data, are available for the large dataset for customer churn. Clicking on some of the data links for the small dataset gives an error. For example, the churn labels for the small dataset are not available. This project will use the large dataset.

3 Exploratory

Explore the data first. Identify some descriptive statistics for the data. Generate some suitable graphical displays for the data. Is the data unbalanced? If so, to what extent, and will the data need to be balanced. The labels are +1 or -1. The description of the data at the source is ambiguous as to which of these labels denotes a churning customer versus a loyal customer. It is assumed from the description that +1 refers to a churning customer. This will need to be confirmed through further investigation of the documentation and data. Can some columns be disregarded before the analysis?

4 Approach

This is a supervised binary classification problem. If the data requires balancing, identify a suitable method given the nature of the data. Identify a good classifier that will also give insight into why churn is happening for Orange. Not only will the machine learning model need to predict well, but also give insight into why it is predicting the way it is. Some models may predict well but give no insight into why the model is predicting for any given instance.

5 Challenges

Challenges include a large dataset (1.88 GB for the large dataset) so suitable tools will need to be used to deal with the size of the dataset. Also, how much number crunching will be required? There are missing values which will have to be dealt with. Some categories have values that only show up in the train or test data.