Final Project: Proposal

Bradford Simkins

DS740: Data Mining

12/1/2018

**My Proposal**

For my final project, I am going to use the Customer Churn Prediction Analysis data set (Huzaif, 2018). The data set contains 91,698 observations, 1 response variable (CHURN), 34 predictor variables, and a column for customer ID number. The idea is to predict churn, which is customer attrition, or the loss of a customer using the 34 predictor variables, or a subset of predictor variables. Exploring the data most of the predictor variables are not even close to being normally distributed. This rules out several methods. I will be using the following two methods to analyze this data set:

* **Decision trees**. I would try bagging with decision trees to try to reduce the variance in the estimates, but first I checked for correlations. There are significant correlations in several of the predictor variables, such as Hist\_Visits with Total\_Sale, W3\_Sale with W3\_Max\_Sale and W3\_Min\_Sale, and W4\_Sale with W4\_Min\_Sale. If the predictor variables are highly correlated bagging will cause high variance, so instead I will use **random forests**. I will then use 10-fold cross-validation to tune the mtry value to choose the number of randomly sampled predictor variables.
* **Artificial neural networks (ANN)**. ANNs are best when used on large data sets with a lot of nonlinear relationships. This is the case with this data set, so I will use it here. I will use 10-fold cross-validation to tune the number of hidden nodes and the weight decay parameter.

If there is a low error rate in the predictive ability of this model it would be very useful for retail businesses to predict which customers will stay and which will be lost in the churn. They can then focus advertising and customer satisfaction surveys in a more targeted way. For instance, they might not want to waste resources sending advertising to customers the model predicts will be lost. They might however want to try to get those potentially lost customers to answer some survey questions to try to get an idea about what is causing customers to stop coming back.

**Reference**

Huzaif, Tila (2018). Customer Churn Prediction Analysis [csv data file]. Retrieved

from <https://www.kaggle.com/huzaiftila/customer-churn-prediction-analysis/>