STOR 565 Proposal

Team: Liam deVerteuil, Xuhesheng Chen and Anthony Hu

Goal of Project

Banks are required to report elder financial abuse but, unless a customer reports fraud and files a claim, financial abuse can go undetected and repeated fraud via digital payments can continue without the banks' knowledge. Without detection models, a large amount of fraud is left unreported by consumers and older adult populations (defined as those who are , which are more susceptible to losing savings to fraudulent payments. Hence, we seek to build classification models, to help identify those who are suffering from elder fraud, and whether victims of elder fraud share similar financial and transactional patterns.

Variables in Dataset

The variables in this set consist of various user financial and demographic information, such as IDs, location, various transaction information (time, amount, device), and account information. Some of these variables may be redundant, so we can drop some of these to simplify our analysis.

Data Processing

Currently, Wells Fargo provides a dataset with 14000 user records with fraud labels, found here. To deal with this problem, we need to split the original dataset into the training set and the testing set.

Since Wells Fargo has offered 14000 samples dataset with fraud labels, in which 4164 samples are labeled fraud, we need to split the original dataset into two datasets –training dataset and test dataset. To preserve the ratio between fraud-labeled data and non fraud-labeled data, the training set will contain 80% of all observations in the data (11200 observations), and the testing set will contain the remainder of the observations (2800 observations).

We have decided that this project will primarily be focusing on supervised learning models, but we have not decided which specific learning method to apply. Viable candidates include logistic regression, SVMs, Naive Bayes, and Random Forest. In general, we seek to create a model to predict fraudulent transactions, based on patterns of various features in the dataset.

We plan to adopt four predictive accuracy measures for assessing trained classification models: Precision, Recall, F1-measure and Accuracy, as well as 10-fold cross validation in the evaluation process.