**The Study of Financial Fraud Detection Using Classification, Clustering, and Anomaly Detection**

**Meng Wang, Jiahui Wu, and Jinjie Liu**

**Progress Report**

**Data Set:**

Synthetic Financial Datasets for Fraud Detection

<https://www.kaggle.com/ntnu-testimon/paysim1>

It is a synthetic dataset generated using the simulator called PaySim. PaySim uses aggregated data from the private dataset to generate a synthetic dataset that resembles the normal operation of transactions and injects malicious behaviour to later evaluate the performance of fraud detection methods.

***Collection Method:***

PaySim simulates mobile money transactions based on a sample of real transactions extracted from one month of financial logs from a mobile money service implemented in an African country. The original logs were provided by a multinational company, who is the provider of the mobile financial service which is currently running in more than 14 countries all around the world.This synthetic dataset is scaled down 1/4 of the original dataset and it is created just for Kaggle.

***Features:***

**step** - maps a unit of time in the real world. In this case 1 step is 1 hour of time. Total steps 744 (30 days simulation).

**type** - CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.

**amount**- amount of the transaction in local currency.

**nameOrig** - customer who started the transaction

**oldbalanceOrg** - initial balance before the transaction

**newbalanceOrig** - new balance after the transaction

**nameDest**- customer who is the recipient of the transaction

**oldbalanceDest** - initial balance recipient before the transaction. Note that there is not information for customers that start with M (Merchants).

**newbalanceDest** - new balance recipient after the transaction. Note that there is not information for customers that start with M (Merchants).

**isFlaggedFraud**- The business model aims to control massive transfers from one account to another and flags illegal attempts. An illegal attempt in this dataset is an attempt to transfer more than 200,000 in a single transaction.

***Target:***

**isFraud**- This is the transactions made by the fraudulent agents inside the simulation. In this specific dataset the fraudulent behavior of the agents aims to profit by taking control of customers’ accounts and try to empty the funds by transferring to another account and then cashing out of the system.

**Preprocessing:**

1. The dataset has 6,362,620 records, 10 features, and 1 class feature. We checked missing data and there were no null values. There are 5 transaction types, but we only focused on “Transfer” and “Cash out” transaction types, as they had been identified to have fraud transactions. Therefore, records of unrelated transaction types were deleted, which finally produced 2,770,409 records.
2. The ratio of fraud transactions of the filtered dataset was only 0.2965%. This is a very skewed dataset. We sampled it to have a balanced data. We sampled the dataset to create a 50-50 ratio by randomly selecting the *p* amount of records from the majority class, with *p* representing the total number of records in the minority class. It produced 16,242 records.
3. The relevant features to fraud transaction were “type”, “amount”, “oldbalanceOrg”, “newbalanceOrig”, “oldbalanceDest”, and “newbalanceDest”. We thus deleted the unrelated features. We binarized the values of “type” feature from string into binary numeric. All relevant features were used as predictors. Feature selection will be refined at the later stage.
4. We standardized the continuous features—“amount”, “oldbalanceOrg”, “newbalanceOrig”, “oldbalanceDest”, and “newbalanceDest”—through boxcox transformation.

**Analytic Platform:**

**1) Supervised Learning:**

1. Logistic Regression (**Finished**)
2. Decision Tree Classifier (CART), Random Forest, and Gradient Boosting (**Finished**)
3. K-nearest Neighbors (**Finished**)
4. Naïve Baysian (**Finished**)
5. Supervised Neural Network (**In progress**)
6. Support Vector Machine (**In progress**)

The things we keep consistent across all supervised learning algorithms are the followings:

1. Filtered/sampled balanced data were used for supervised learning.
2. Training and testing sets were spitted at ratio 7:3, sampling used the same seed (0 in this case) to keep the results comparable.
3. Cross validation procedures were performed to optimize the parameters with CV Fold = 10.
4. For the model evaluation, we are more inclined to increase the recall score rather than the precision score or the accuracy score in order to capture the most fraudulent transactions. This is because if we predict a normal transaction to be fraudulent, it is not as serious as the opposite situation.

The things that vary across all supervised learning algorithms are the followings:

1. Specific transformation of predictors (continuous variables in this project) varies across classifiers.
2. The model parameters as well as the range of sensible values vary across classifiers.

**2) Unsupervised Learning**

1. Clustering
2. Anomaly Detection
3. Neural network (Unsupervised)

We will use the original dataset with 2,770,409 records without sampling.

**Model Evaluation:**

We will compare the recall scores of the results from different classifiers, try different combination of parameter settings and elaborate why one classifier is better over the other.

**Resources and Environment:**

Following Data Analytic Tools are used:

Python 2.7, Java, R, and SQL Server.