**Summary**

Our recommendation is <best model> with <data set>. The classification accuracy was <best accuracy>. Based on receiving new data with these same categories, the business will be able to predict with approximately a 95% accuracy whether the transaction will result in money lost (0) or gained (1). Out of the 5% error rate in prediction, 4.07% were false positive results and 6.56% were false negative results. The 95% accuracy can be summarized as

*(Add this to the code we turn in, not sure where, so putting it here:*

*False Positive Rate-781/19202=.0407, False Negative Rate-839/12792*

*True Positive Rate-94% (shown on output), True Negative Rate-96%.*

***Adjust numbers and calculations from actual decision!)***

**Our Process**

The objective is to suggest/recommend Machine Learning models that would best classify the outcome ‘y.’ Our process was as follows:

1. Clean Data
2. Feature Select
3. Model
4. Analyze
5. Evaluate

**Clean Data**

The raw data set was unlabeled. It contained a total of 160,000 records and 51 features. Using a Python library called ‘Pandas\_Profiling’, we uncovered the following:

1. Datatypes
   1. Numeric: 45
   2. Categorical: 5
   3. Boolean: 1
2. High Correlation
   1. x41 – Dropped x41 from dataset
   2. x6 – Dropped x6 from dataset
3. High Cardinality
   1. x37 – The values had incorrect datatype – causing an incorrect assesessment of high cardinality. To clean it, we removed the commas and dollar sign, and replaced the parenthesis with a minus symbol. Converted datatype to numeric.
4. Data augmentation
   1. Misspellings were present in x24. Corrected ‘euorpe’ to ‘europe’
   2. Standardize values
      1. x29 – Replaced all month abbreviations to full spelling (i.e. Aug to August)
      2. x30 – Corrected misspelling of ‘thurday’ to ‘thursday’
5. Missing values, NaN:
   1. There are 20-40 missing values in each column.
   2. Most of the missing values appears to be from Asia. It was decided that these data can be replaced with the mean of the available data from Asia.
   3. The observations with missing values in the column with different continent names were dropped.
6. Data Distribution was normal. No data transformation was required.

**Feature Select**

This portion took a bulk of our time. The process of feature selection was a continuous cycle between ‘feature select’ and ‘modeling.’ The first pass was to use the complete dataset with each subsequent iteration removing features and/or records.

We attempted to reduce the number of features required through PCA. We found that 8 principal components can explain for 99.9% of the variance.

Because this was a classification assignment, we selected a few classification models to find the best model for the data set (refer to ‘Model’ section for more details). In the end, we decided that all the categorical features made little to no impact on the accuracy scores – therefore, dropped. Using Random Forest, we found the following features to be of importance:

**Model**

Logistic Regression

Decision Tree

K-NN

SVM

Naïve Bayes

**Analyze**

**Evaluate**