**Capstone Report**

**1. Define**

**Overview:** The goal of this project is to classify real-word e-commerce transactions as fraudulent and non-fraudulent. This is a [Kaggle competition](https://www.kaggle.com/c/ieee-fraud-detection) using data provided by the Vesta Corporation. The competition is also sponsored by the IEEE Computational Intelligence Society (IEEE-CIS) and their researchers who want to improve the accuracy of fraud detection models.

**Problem Statement:** As defined by Kaggle, “In this competition you are predicting the probability that an online transaction is fraudulent, as denoted by the binary target isFraud.” The submission file contains the probability of fraud for a given transaction, not rounded to 0 or 1.

**Metrics:** The submission file will contain the ID for each *test* transaction and the corresponding probability that the transaction is fraudulent. Kaggle then calculates the area under the ROC curve. The y-axis of the ROC curve is the True Positive Rate (TPR) = TP / (TP + FN) and the x-axis is the False Positive Rate (FPR) = FP / (TN + FP). By examining different cutoff thresholds (when to predict 0 or 1), other than just 0.5, the ROC curve can be created. The curve represents the tradeoff between false positives and false negatives. The optimal AUC score would be 1, implying that the model perfectly classifies transactions.

This way of measuring classification tasks is better than simply calculating the accuracy of the predictions. In this data set, only 3.5% of the transactions are fraudulent. Thus, a model that predicts every transaction as NOT fraudulent would have an accuracy of 96.5%, giving the illusion that this is an effective model.

**2. Analyze**

**Data Exploration: “**The data is broken into two files identity and transaction, which are joined by TransactionID. Not all transactions have corresponding identity information.” Thus, both the train and test set have two files (identity and transaction) that need to be merged to create one train table and one test table. The result is a training set of 590,540 examples and a test set of 506,691 examples, both with 432 features.

**Exploratory Visualization:**

**Algorithms & Techniques**

**Benchmark:** The public leaderboard on Kaggle establishes a reference for evaluating the performance of my model.

**3. Implement**

**4. Results**

**5. Conclusions**