**Team 14 Project Proposal**

**Introduction & Background**

Our team has decided to use machine learning techniques to analyze the “Credit Card Fraud Detection” Kaggle dataset. Credit card fraud is a rampant problem. In 2020, the Federal Trade Commision identified credit cards as the prevalent payment method in fraud reports. Also, with a one-year jump of 44.6% between 2019 and 2020 according to the Federal Trade Commission's annual 2020 report, credit card fraud is a growing problem whose root causes should be addressed to mitigate these frauds. Existing analyses of credit card fraud have included techniques including Deep Learning/Entropy, correlation matrices, and confusion matrices. While some of these applications might be useful for our dataset, we must consider its characteristics: few fraudulent entries, no null values, and high false positive/negative rate. We found all this information after preliminary analysis after pre-preprocessing our data.

**Problem Definition**

According to Experian, credit card CVV’s sell on the dark web for as little as $5. Moreover, people’s entire credit information - including name, Social Security number, date of birth, and complete account numbers - sell for as low as $110. In a digitally financial age with an increasingly susceptible aged population, it is more important now than ever to have sophisticated fraud detection techniques to protect people’s livelihoods. Our project focuses on the detection of credit card fraud through analysis of various PCA transformed components.

**Dataset**

After scouring kaggle, we found this dataset to suit our purpose: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>. It contains anonymized credit card transactions labeled as either fraudulent or genuine. It presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced - the positive class (frauds) account for 0.172% of all transactions.

**Methods**

As of now, we plan to detect fraud in a few ways: via (1) GMM clustering, (2) regression analysis, and (3) a multi-layer perceptron neural network. GMM, an unsupervised learning clustering algorithm, results in a soft-clustering output that we can compare against the dataset’s labeled output to determine the model’s accuracy. Regression analysis (either linear, logistic, or both) can be used to identify the correlation between the input parameters and fraudulence. Finally, a well-tuned MLP neural network for classification will likely serve as the basis of comparison.

**Potential Results & Discussion**

From literature review, we discussed that Precision, Recall, F-score, and AUC would be the most suitable metrics for the unbalanced nature of our dataset. Specifically, for GMM, we plan to use the specificity, recall, and Balanced Accuracy of GMM. For regression analysis, we aim to use a straightforward mean squared or root squared error. Lastly, for MLP, we plan to use recall, precision and F-1 score. The area under the precision-recall curve (AUPRC) is also a useful performance metric in our problem setting since we care a lot about finding the positive examples.

**References**

<https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>

<https://www.cardrates.com/advice/credit-card-fraud-statistics/>

<https://mint.intuit.com/blog/planning/credit-card-fraud-statistics/>

<https://statmodeling.stat.columbia.edu/2020/01/10/linear-or-logistic-regression-with-binary-outcomes>

<https://glassboxmedicine.com/2019/03/02/measuring-performance-auprc/#:~:text=The%20area%20under%20the%20precision,about%20finding%20the%20positive%20examples>

**Contribution Table**

|  |  |
| --- | --- |
| **Name** | **Project Task** |
| Mitali | **GMM** |
| Matthew |
| Sunidhi | **MLP** |
| Muskaan | **Data Visualization** |
| Azhan | **Regression Analysis** |

**Timeline**

https://docs.google.com/spreadsheets/d/1po2iJ1vFaG1CmB4J\_Djwdk4p0TJhpWTq/edit#gid=290773380

Table

Description automatically generated