Midterm Project Report

1st Colsen Murray
Tickle College of Engineering
University of Tennessee
Knoxville, United States
jmurra47@vols.utk.edu

2nd Ashton Dy
Tickle College of Engineering
University of Tennessee
Knoxville, United States
ady@vols.utk.edu

3rd Yousif Abdulhussein
Tickle College of Engineering
University of Tennessee
Knoxville, United States
yabdulhu@vols.utk.edu

4th Ulugbek Kahramonov
Tickle College of Engineering
University of Tennessee
Knoxville, United States
ukahramo@vols.utk.edu

***Abstract*—The Noname2 group wrote this paper to present a method of detecting credit card fraud. The research builds an ensemble machine learning model to predict fraud cases. Our model combines multiple classification algorithms and data preprocessing to ensure the accuracy of our results. The experimental results show substantial improvements in fraud detection performance through F1 scores that exceed industry benchmarks. This research helps protect financial institutions and consumers from digital-economy fraud through its contributions to ongoing security efforts.**

***Index Terms*—Fraud Detection, Machine Learning, Ensemble Model, Classification Models**

I. Introduction

A. Background and Motivation

Credit cards are a crux of the modern financial world, spanning industries from retail to banking. The ease of tapping a card to a register and worrying about payment later makes shopping easier for the average consumer, providing an incentive to spend more money [1]. It also provides a certain amount of security to purchases. Because the money does not need to leave the purchaser's account until the end of the month, fraudulent purchases can be flagged and dealt with before they impact the card holder.

However, as credit card uses expand, so does the risk of fraudulent transactions. Most security threats are the result of negligence on the part of a human. For example, a credit card is much more likely to be stolen through a phishing attack than a data leak. However, both attacks increase the likelihood of a fraudulent transaction. These purchases can cause headaches for card holders, concurrently, resulting in large losses of money for banks and financial institutions. In 2024 alone, 449,032 cases of credit card fraud were reported [2]. As a result of this, an entire industry has developed around the issue of preventing it.

B. Significance of Addressing the Problem

Fraudulent charges can cause economic damage to both banks and consumers simultaneously, resulting in a loss of reputation for the financial institution. When their fraud flags fail to trigger a scam, people lose trust in that bank. As a result, the problem of fraud detection is very relevant to the bottom line for many businesses. Aside from this, failing to prevent fraud and having to pay back cardholders can cost banks significant amounts of money, requiring interest raises or lowering the revenue for lenders.

Credit card fraud reports by year

Year	Credit Card Fraud Reports
2019	277,739
2020	399,727
2021	395,391
2022	448,443
2023	425,988
2024	458,538

Data source: Federal Trade Commission (2025).

Fig. 1. Fraud Cases by Year

C. Our Solution

The aim of this project is to create a system that can detect fraudulent credit card transactions using an ensemble of machine learning techniques. Traditional methods rely on predefined rules and struggle to keep up with the ever-evolving nature of fraud. By combining multiple models, we can take advantage of their individual strengths and build a more robust solution. An ensemble model allows us to reduce false positives while still catching suspicious behavior that might otherwise go unnoticed. The ultimate goal is to provide a tool that not only detects fraud with greater accuracy, but also adapts to new patterns and reduces financial risk for both consumers and institutions.

D. Structure of the Paper

The remainder of this paper is organized as follows. First, we present a description of the dataset used for this project. This section outlines the features provided, any cleaning or normalization that was necessary, and the trends we observed through exploratory data analysis. Following that, we introduce a baseline solution to the fraud detection problem. We discuss existing approaches, justify our choice of baseline, and explain how it was implemented. Finally, we explore potential extensions of the project. These include ideas for

new feature engineering, improvements to the baseline model, and additional ways to visualize and interpret the results.

II. Dataset

A. Structure of the Dataset

This project uses a publicly available credit card transaction dataset sourced from Kaggle [3]. The dataset contains a wide range of features related both to the transaction and to the card holder. These include details such as transaction date and time, merchant name and type, purchase amount, and location-based information such as the city and state of the card holder, as well as the latitude and longitude of both the merchant and the purchase. Additional features provide a demographic context, including the population of the card holder's city, their occupation, and their date of birth. Each purchase is also assigned a unique transaction number, and most importantly, a binary label indicating whether or not the transaction was fraudulent.

In total, the dataset includes over 14,000 individual entries and 14 distinct features. However, the distribution of these entries is heavily imbalanced, with only around 13 percent of the transactions labeled as fraudulent. This skewed class distribution poses a challenge for most machine learning models, as they tend to favor the majority class unless specific techniques are applied to mitigate the imbalance. Understanding the nature of this data is a critical first step in building an effective fraud detection model, as it helps guide preprocessing steps, feature selection, and evaluation strategies.

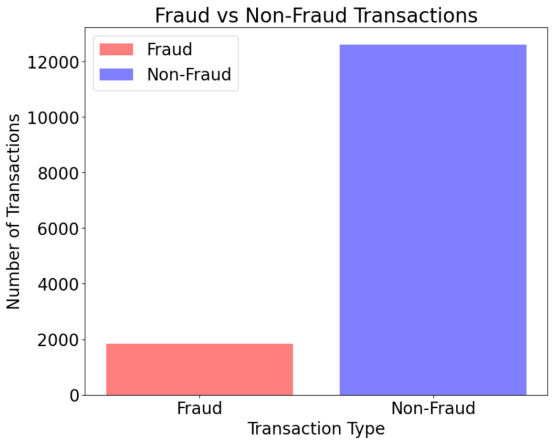

Fig. 2. Number of Fraudulent versus Non-Fraudulent Transactions

B. Data Cleaning

We engineered certain features to take advantage of the given features that were word-based. For example, we created a feature representing merchant frequency by counting how often each merchant appeared in the dataset. We then transformed this into a binary variable, indicating whether a merchant appeared fewer than 30 times. This required grouping the data by merchant name and filtering based on count. The direct correlation with fraud was weak, but we used this as an input to the models.

We then applied the same logic to geographic data. We grouped transactions by city and state, computed their frequencies, and created a binary feature for low-frequency locations (fewer than 30 appearances). This process involved combining city and state fields, aggregating counts, and merging the result back into the main dataset. Like the first feature, this provided a minimal but distinct signal for classification.

The third engineered feature involved a more complete data preprocessing pipeline. We began by removing rows with missing or null values to ensure consistency. Next, we retained all available features and applied one-hot encoding to categorical columns such as job title and merchant type. These steps produced a clean, model-ready dataset with fully numeric inputs, suitable for training a more comprehensive model.

C. Trends in the Data

In analyzing the geographic distribution of transactions, we noticed that most states had a fairly even split between fraudulent and non-fraudulent cases. However, a few states deviated from this pattern. Specifically, Alaska, Nebraska, and Oregon showed a disproportionately high number of fraud cases relative to their total transaction counts. This outlier behavior suggests that regional characteristics could influence the likelihood of fraud. Identifying these anomalies helped inform our consideration of geographic features in model design.

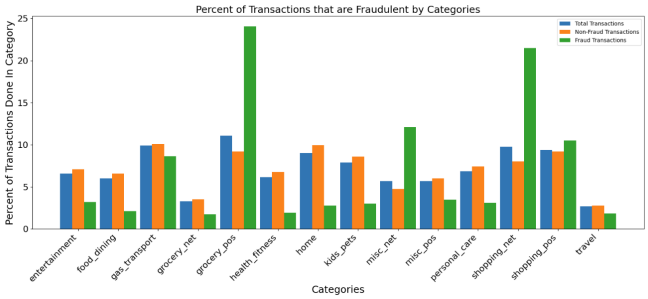

Fig. 3. Percent of Transactions that are Fraudulent by Categories

Age also revealed clear trends in fraud patterns. While most non-fraudulent transactions were made by individuals aged 30 to 60, fraudulent transactions tended to cluster in the upper end of that range, particularly between ages 50 and 60. Interestingly, the 30 to 50 age range accounted for a large volume of purchases but significantly fewer instances of fraud. This contrast suggests that older card holders may be more vulnerable to fraudulent activity, which could be a valuable feature for model training or risk profiling.

III. Baseline Solution

A. Existing Approaches

A variety of solutions have been developed to detect credit card fraud, ranging from traditional rule-based systems to modern machine learning models. Rule-based systems rely on

predefined patterns, such as flagging unusually large purchases or transactions from foreign countries, but they often struggle to adapt to new fraud tactics [4]. Statistical methods, like anomaly detection and regression analysis, have also been used to identify outliers in spending behavior. More recently, machine learning approaches have become popular due to their ability to learn complex patterns from data. Models like logistic regression, decision trees, and random forests are commonly used to classify fraudulent transactions. Some systems also use unsupervised learning to spot unusual activity without needing labeled data. In many cases, these techniques are combined with data balancing methods, such as oversampling rare fraud cases or generating synthetic examples, to improve accuracy.

B. Our Solution

Our solution involves an ensembling approach, which aims to surpass the above solutions by combining several different models into one system. Instead of relying on a single model that might perform well in some situations but poorly in others, we allow multiple learners to contribute to the final prediction. Each individual model brings its own perspective—some may focus on age, others on location or transaction patterns. This diversity helps balance out individual weaknesses. For example, if two learners incorrectly label a transaction as non-fraud, but five others correctly flag it, the ensemble can still make the right call. By letting the majority guide the decision, we reduce the risk of bias from any single model and improve overall accuracy. This approach not only helps catch more fraud, but also avoids flagging too many legitimate purchases, which is equally important for maintaining trust in the system.

For the sake of this project, our baseline models can be considered the standard models we used, with our improved version being the voting classifier ensemble, which combines the functions of all the other models.

C. How We Implemented It

We began by testing many different models. First, we tried regression models. Our linear regression was our overall worst performing model; however, this is to be expected considering that our task is labeling and linear regression is not intended for that purpose. We then created a logistic regression model. This one performed better, with a 94% accuracy. The issue with this model was its less than stellar recall score of 0.63. This means it has an excess of false negatives, which is undesirable for fraud detection.

Next, we tried more ambiguous models with more complex bases. We tried a random forest and a decision stump model. The first generates a forest of prestructured decision trees, which it then fills with features. The second does the same, but each tree only has a depth of one. Both of these are simple ensembles in and of themselves. The forest had an accuracy of 97.6% using K-fold CV, with generally high metrics. The decision stump had an accuracy of 94.6%, with significantly lower recall, precision, and f1-score. Because of its high performance, we included the random forest in our ensemble. Finally, we tried a gradient boosting model. This model resulted in a K-fold accuracy of 97.6%, with very high metrics as well.

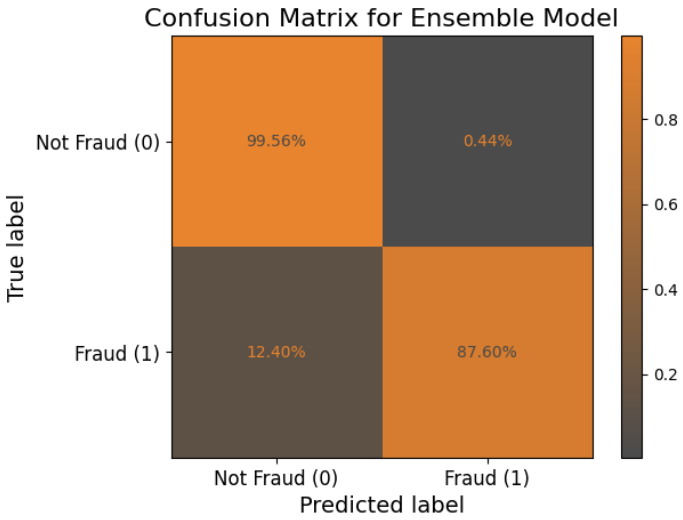

Fig. 4. Confusion Matrix for the Ensemble Model

Our final implementation relies on several separate learners operating in parallel, each trained separately. These learners run independently and generate their own predictions based on the patterns they recognize. We then connected these models through a group voting mechanism. This voting system takes the predictions from each learner and uses the majority outcome to decide whether a transaction fraud. The idea is that even if one model makes an error, the rest can correct it through consensus. Based on their individually high performances, we selected the logistic regression, random forest, and gradient boosting models for use in the ensemble. The result was a K-fold accuracy of 98% with very high metrics.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.94	0.93	0.82	0.86
Random Forest	0.98	0.97	0.94	0.95
Decision Stump	0.95	0.89	0.87	0.88
Gradient Boosting	0.98	0.97	0.93	0.95
Ensemble	0.98	0.97	0.93	0.95

TABLE I
CLASSIFICATION METRICS FOR FRAUD DETECTION MODELS

IV. RESULTS AND DISCUSSION

The individual results gathered from the various models that we trained appeared to be extremely promising. Each model demonstrated remarkably high accuracy, all achieving accuracy well above 90%. Most importantly, our final ensemble voting classifier combined all of the strengths of our previous models to yield and extraordinary 99.9% accuracy. This results initially appeared to be a resounding success.

Model	Accuracy With Time	Accuracy Without Time
Logistic Regression	0.943	0.944
Random Forest	0.999	0.977
Gradient Boosted	0.998	0.976
Ensemble	0.999	0.978

TABLE II
MODEL PERFORMANCE

Although our results were remarkable by nature, we decided to continue our research with a degree of caution. While our models clearly fit well to the dataset, these extremely high metrics, prompted us to question the reliability and generalizability of our model. As a result, we took a systematic approach to investigate different aspects of our project. This included assessing the models we used, the reviewing our data transformations and training methods, and most importantly, an in-depth analysis of the dataset and its features to ensure it represents the entire hypothesis space.

The root of our deceitful success lied in the features of the dataset itself. We employed a multitude of methods to clearly identify and visualize the issues within the dataset. We created a correlation heat map of all of the features to expose any underlying connections between features. The only significant correlations observed were those we anticipated, such as among temporal features like year, month, and day, as well as between geographical coordinates and their corresponding states. We observed a correlation between the time stamps and the likelihood of the sample being fraudulent. Upon further investigation, we identified a clear error in how the data was collected. The data was cherry picked where only one category was collected in a given time period.

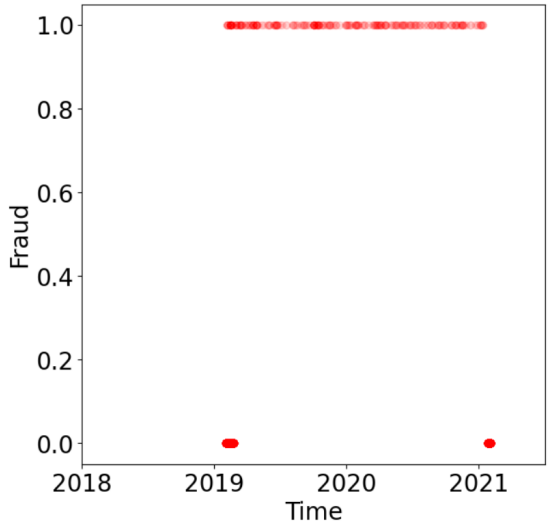

Fig. 5. Each Sample was plotted on a timeline against its fraud feature

This error in data collection likely led our models learn to look at the transaction year, month, and day instead of wholistically taking into account all of the features to determine if the transaction was fraudulent. The accuracy we got was very misleading for us as our model would likely not generalize well to the unseen hypothesis space.

We retrained our models without the time stamps and still got high but definitely more realistic results. Our ensemble classifier fell from 99.9% to 97.8% with a confidence accuracy interval 95% between 97.1%-98.3%. While we were able to identify the error in the transaction times, we are still apprehensive to accept the reliability these new accuracies given that other features may been cherry picked as well.

For future works, validating the collection of the dataset will be evermore important as important features such as transaction date and time can end up working against us. Additionally, the dataset spanned two years during the Covid-19 Pandemic which could further skew the data points that were collected. While data is important for machine learning, not all data is good data, in this case it permitted misleading results and a model that does not generalize well.

V. DISTRIBUTION OF WORK

Yousif Abdulhussein was responsible for the preprocessing of the dataset, including data analysis, cleaning, and organization to make it compatible with the AI model. Also, he created and produced visualizations to help in the understanding and explanation of the data.

Colsen Murray was responsible for documentation and writing the midterm report, analyzing the existing project code and recording it for the paper, while transferring it to an IEEE-ready format. He also provided the K-fold cross validation and performed the reformatting and commenting of the code for readability.

Ashton Dy was responsible for much of the data analysis. He handled the splitting and scaling of the current data. In addition, he set up the logistic regression model and decision tree with metrics to enable easy analysis of their performances. He also performed the analysis to find the date/time error.

Ulugbek Kahramonov initially implemented some of the simpler models and later created our ensemble model using a Voting Classifier from the aforementioned models.

VI. CONCLUSION

Credit card fraud continues to be a growing problem with serious consequences for both consumers and financial institutions. Through this project, we are exploring an ensemble-based machine learning approach to improve fraud detection by combining weak learners, each focusing on different aspects of the data. We began by examining the structure and trends within our dataset, identifying key patterns in geography and age that informed our model design. Our initial learners used simple, interpretable features, while our more advanced model worked with the full cleaned dataset and additional dummy features. As we continue, we plan to expand the ensemble with new learners to better capture the complexity of fraud. Ultimately, our goal is to build a flexible and reliable system that can adapt to the ever-changing nature of fraudulent activity.

REFERENCES

[1] D. Prelec, S. Banker, "How credit cards activate the reward center of our brains and drive spending," MIT Sloan Management School, 9 Jun. 2021. [Online]. Available: https://mitsloan.mit.edu/experts/how-credit-cards-activate-reward-center-our-brains-and-drive-spending

[2] J. Caporal, "Identity theft and credit card fraud statistics for 2025," Motley Fool Money, Mar. 13, 2025. [Online]. Available: https://www.fool.com/money/research/identity-theft-credit-card-fraud-statistics/

[3] N. R. Choudhury, "Credit card fraud data," Kaggle, Oct. 2024. [Online]. Available: https://www.kaggle.com/datasets/neharoychoudhury/credit-card-fraud-data

[4] M. Bhati, "Rules Based Fraud Detection Approach, Types and Benefits," Nected, 24 Dec. 2024. [Online]. Available: https://www.nected.ai/blog/rules-based-fraud-detection