Detection of Credit Card Fraud with Artificial Neural Networks and Multiple Classifier Systems

Matthew Laluzerne
mjlaluzerne@wisc.edu

Rui Feng
rfeng34@wisc.edu

Ran Zhao
rzhao99@wisc.edu

ME/ECE/CS 539 3/4/23

Overview

The proposed research will analyze the use of Artificial Neural Networks (ANNs) and Multiple Classifier Systems to better predict credit card fraud. The analysis will compare the classification accuracy of one and multiple ANN systems using Credit Card Fraud datasets from Kaggle.

Background

Ever since the appearance of credit cards in 1950, the number of people preferring credit cards for payments has been rising every year. A report from Federal Reserve Bank of San Francisco shows that 51% of retail transactions are done using credit or debit cards, a lot higher than cash which has only 26% [1]. The rapid growth in the number of credit card transactions has led to a severe problem: the rise of fraudulent activities [2]. Many people have been in a credit card fraud case. 151 million American card holders have been the fraud victims. That's 65% of card holders and the number is still rising compared to 58% last year [1]. Credit Card Fraud greatly harms our economy. The Nilson Report, monitors the payment industry, released a forecast in December 2022, indicating that $165.1 billion will be lost due to card fraud over the next 10 years [3].

The poor protection against credit card fraud has driven us to develop a solution better than keeping a good habit when using credit cards. Machine learning has proved to be effective at anomaly detection, which means it is powerful for detecting unusual patterns such as fraud detection. As credit card fraud has become a main part of fraudulent activities, it has drawn great attention from machine-learning areas. The following techniques have been applied for fraud detection: Decision Tree, Neural networks, Logistic Regression, Genetic algorithms, clustering techniques, and outlier detection [4]. Currently, the study shows that, among all the techniques in machine learning, KNN, Logistic Regression Classifier, Random Forest, and Bayes have the highest accuracy rates in the field [5]. Furthermore, more research has focused on how to improve the performance of these techniques. In order to deal with imbalance data problem, a variety of data augmentation technique have been compared and a new data augmentation model, K-CGAN, has been proposed [6]. We would like to find further optimization probability based on the materials we learn from this course.

Datasets

The dataset we plan to use for the project is about Credit Card Fraud from Kaggle. It contains 1000000 samples (rows) and 8 features (columns) in total with a size of 76.28 MB (Mary Beth). Feature Explanation:

1. distance_from_home: the distance from home where the transaction happened.

2. distance_from_last_transaction: the distance from last transaction happened.

3. ratio_to_median_purchase_price: Ratio of purchased price transaction to median purchase price.

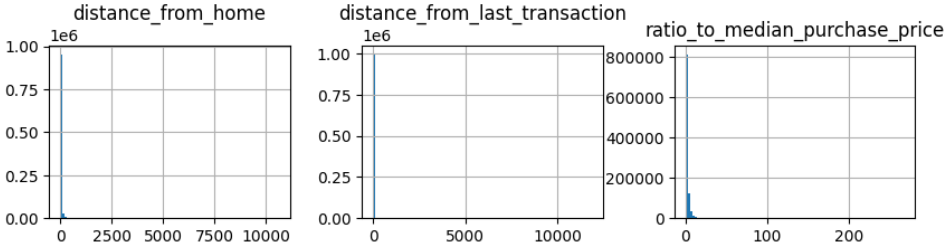
4. repeat_retailer: Is the transaction happened from same retailer.

5. used_chip: Is the transaction through chip (credit card).

6. used_pin_number: Is the transaction happened by using PIN number.

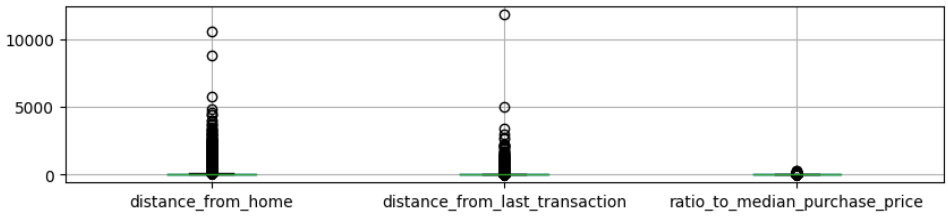
7. online_order: Is the transaction an online order.

8. fraud: Is the transaction fraudulent.



Among these 8 variables, the first three are numerical while the others are binary. For the five binary variables, the proportion of zeros are 0.118464, 0.649601, 0.899392, 0.349448, 0.912597, respectively.

Just as what others said about this dataset, we have 91.26% of Genuine transactions and only 8.74% of fraud transactions! So, accuracy seems not to be a reasonable choice when evaluating the models. Besides, the dataset has a lot of huge outliers and is highly skewed.



Method

The data set will require minimal preprocessing because it is well organized into features and labels and has no missing values. No type conversion is needed because all the data is already in numeric form. The only issues with preprocessing will be addressing outliers and skewed distributions in the data set. For example, fraudulent transactions make up only 8% of the data set, meaning that downsampling may be needed. This is possible due to the large amount of samples in the data set. Additionally, there are some large outliers in the distance columns. A transaction that is far from home could be an indicator of fraud, so removing these outliers could hurt the performance of the model. All the following methods will be done with and without removing the outliers and the performance will be compared.

Previous work has used deep neural networks, decision trees, support vector machines, k nearest neighbors, and random forest classification to classify the data set.

Decision trees had the most success by far, with near perfect classification in 2 different code examples on Kaggle. A decision tree will be trained to provide a benchmark for other methods since it will most likely be the best performing classifier. After this, a naïve bayesian classifier, deep neural network, k nearest neighbors classifier, and logistic regression classifier will be used to classify the data set. The performance of these methods will be compared to the decision tree and other code Examples on Kaggle to see if they can achieve comparable performance to a decision tree in this application.

Outcome

In the project, we will explore methods to deal with imbalanced data, build classifiers and then evaluate them with ROC-AUC, F1-score, etc. Since the data set is imbalanced, we will not be able to use accuracy as a measurement of performance because the model may learn to classify all transactions as non-fraudulent.

We will also try to improve the benchmark already established by others with the concept and tools in the field of neutral network. We hope to achieve similar results to others using decision tree classification and achieve better results than others with the other methods described.

Project Plan

Deadlines:
Progress report due: 4/7
Project presentation: 5/1-5/5
Final report due: 5/6

Gantt Chart:

Task/Week	3/6	3/17	3/20	3/27	4/3	4/10	4/17	4/24	5/1
Revise project ideas and proposal if needed									
Write preprocessing code									
Build classification models									
Improve classification models									
Start final report									
Revise Final Report									
Create Presentation									

References

[1] “2023 Credit Card Fraud Report” Accessed: Mar. 3,2023.

[2] S. Benson Edwin Raj and A. Annie Portia, "Analysis on credit card fraud detection methods," 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET), Tirunelveli, India, 2011, pp. 152-156, doi: 10.1109/ICCCET.2011.5762457.

[3] ‘Credit card fraud statistics” by John Egan. Accessed: Mar. 2, 2023.

[4] Delamaire, L., Abdou, H., & Pointon, J. (2009). Credit card fraud and detection techniques: a review. Banks and Bank systems, 4(2), 57-68.

[5] CARD, R. T. T. B. C. (2023) A Proposed Framework of Dimensionality Reduction Techniques to Boost Credit Card Fraud Classification.

[6] Strelcenia E, Prakoonwit S. Improving Classification Performance in Credit Card Fraud Detection by Using New Data Augmentation. AI. 2023; 4(1):172-198.