**Credit Card Fraud Analysis via KNN & Decision Tree**

Anthony Rispoli

Computer Science Department

Hood College, Frederick, Maryland 21701, USA

asr16@hood.edu

**ABSTRACT**

This paper focuses on the analysis of credit card transactions to determine if the transaction was fraudulent. There has been an increase in the number of fraudulent credit card transactions, causing many people financial distress. Data for this paper was collected from Kaggle. Both a Decision Tree (DT) model and a K Nearest Neighbor (KNN) model were made to detect these fraudulent transactions. Both techniques proved viable with near-perfect accuracy.

**KEYWORDS**

Machine Learning • Decision Tree • KNN • Credit Card Fraud • Data Analysis

**OVERVIEW**

Credit card fraud is a major concern for both consumers and financial institutions. In recent years, advances in technology have enabled criminals to commit fraud more easily and with greater sophistication. As a result, it is crucial for banks and other financial institutions to develop effective methods for detecting and preventing credit card fraud. This paper explores the use of two popular machine learning algorithms, KNN and Decision Tree, for the analysis of credit card fraud. By comparing the performance of these algorithms, this paper aims to provide insights into their effectiveness for detecting and preventing credit card fraud.

**BACKGROUND**

Credit card fraud is a serious and pervasive problem that affects individuals, businesses, and financial institutions around the world. In 2017, the total amount of credit card fraud losses worldwide was estimated to be over $22 billion, with the United States accounting for the largest share of those losses[3].

The manner in which fraudsters are perpetrating these unauthorized transactions is always changing and evolving to become more difficult to detect. This implies that the methods used to detect this fraudulent activity also needs to evolve and grow with the schemes they are built to stop.

There have already been many studies completed on this topic. Researchers at Universiti Sains Malaysia created a hybrid machine learning architecture that combine popular machine learning techniques in an attempt to correctly identify fraudulent charges in a data set similar to the one this paper focuses on[1].

Another analysis technique used focuses on the Statistical Process Control (SPC) model, which is the current model used in society for real-life credit card fraud analysis[2].

Researchers at Dalian University of Technology completed a comparative study that focused on models using popular machine learning techniques such as a Boosting Algorithm, Random Forest classification, and K-neighbors[3].

There is a plethora of models that have been created that utilize popular machine learning techniques, such as the ones mentioned above. There are many differences found among these models and they are all phenomenal utilities in combatting this fraud. The models created in this paper aim to further the ability to identify these fraudulent transactions quickly and correctly.

**LIBRARIES**

The Python libraries used in this paper and their purpose can be seen below:

* SKLearn- Decision tree and KNN machine learning libraries utilized to create models and to split data into training set and testing set
* Pandas – used in reading CSV file
* Matplotlib – used to statistically visualize data in data analysis
* Seaborn- used to statistically visualize data in data analysis

**DATA**

The dataset used in this paper was taken from online database Kaggle[4]. This dataset contained 100000 credit card transactions with the following attributes:

* distancefromhome - the distance from home and where the transaction happened.
* distancefromlast\_transaction - the distance from where the last transaction happened.
* ratiotomedianpurchaseprice - Ratio of purchased price transaction to median purchase price.
* repeat\_retailer - Did the transaction happen from same retailer.
* used\_chip - Is the transaction through chip (credit card).
* usedpinnumber - Did the transaction happen by using PIN number.
* online\_order - Is the transaction an online order.
* fraud - Is the transaction fraudulent.

This dataset contained no missing values and there were no duplicate entries.

A heatmap was created in Figure 1 to visualize the correlation between the different attributes and the fraudulent attribute in the dataset.

Chart, treemap chart

Description automatically generated

Figure 1. Correlation heat map of dataset attributes.

It is apparent in Figure 1 that the attributes that have the highest correlation to whether or not a transaction is fraudulent are online\_order, distance\_from\_home, used\_pin\_number, and most significantly ratio\_to\_median\_price. This data was further analyzed to determine the distribution of fraudulent transactions and real transactions featured in this dataset. A pie chart of this distribution can be seen in Figure 2.

Chart, pie chart

Description automatically generated

Figure 2. Distribution of real and fraudulent transactions.

About 91% of the transactions featured in this data set were authorized transactions while only about 9% of transactions were fraudulent. This rate of fraudulence is much higher than the observed rate of credit card fraud in the real world, which make up about .2% of all credit card transactions, but that is not of the most concern and this data is still applicable in training a model to detect fraudulent transactions.

60% of the dataset was used to independently train each model and 40% was saved for testing the accuracy of each model.

**KNN**

KNN clustering involves creating separate clusters of the data of size k based on how closely they neighbor other data points when plotted. Determining an optimal k value is crucial when creating this type of model. A plot of expected accuracy scored over varying k values was made and can be seen in Figure 3.

Chart, line chart

Description automatically generated

Figure 3. Plot of accuracy v K values

It is clear that as the K value increases, the accuracy of the model decreases. Because of this, K was set to 1 when training the KNN model used in this paper. After completing the training of the model and testing it with extra 40% of the data saved, the model yielded accurate results 98.505% of the time.

**DECISION TREE**

The same 60% training data was fit to a decision tree model. After fitting the model and testing it with the same 40% used in the KNN model, the decision tree surprisingly yielded more accurate results than the KNN model with 99.9955%.

**CONCLUSION**

Credit card fraud schemes have become more sophisticated and more popular as technology has developed. This evolving issue has called for many equally sophisticated techniques to detect and stop these fraudulent transactions in real time. Even though a plethora of these models already exist, it is important to keep adjusting and innovating upon these models to keep ahead of the evolving fraud schemes. This paper highlights two popular machine learning algorithms that could be utilized to address this ongoing issue that society has been facing.

**CITATIONS**

[1] Esraa Faisal Malik, Khai Wah Khaw, Bahari Belaton, Wai Peng Wong, and XinYing Chew. 2022. Credit Card Fraud Detection Using a New Hybrid Machine Learning Architecture. *Mathematics (2227-7390)* 10, 9 (May 2022), 1480–1480. DOI:https://doi.org/10.3390/math10091480

[2] Phuong Hanh Tran, Kim Phuc Tran, Truong Thu Huong, Cédric Heuchenne, Phuong HienTran, and Thi Minh Huong Le. 2018. Real Time Data-Driven Approaches for Credit Card Fraud Detection. In *Proceedings of the 2018 International Conference on E-Business and Applications* (ICEBA 2018), Association for Computing Machinery, New York, NY, USA, 6–9. DOI:https://doi.org/10.1145/3194188.3194196

[3] Hanbing Zou. 2021. Analysis of Best Sampling Strategy in Credit Card Fraud Detection Using Machine Learning. In *2021 6th International Conference on Intelligent Information Technology*, ACM, Ho Chi Minh Viet Nam, 40–44. DOI:https://doi.org/10.1145/3460179.3460186

[4] Credit Card Fraud. Retrieved December 12, 2022 from https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud