Capstone Paper

Fraud Detection Tool

Aleksandr Shaghbatyan, Gurgen Hovakimyan

BS in Data Science

American University of Armenia

May 9, 2024

Contents

Abstract

Utilizing machine learning algorithms has proven to be a potent method for managing risk and preventing financial fraud. However, the requirement for coding expertise and deep knowledge of machine learning stands as a barrier for users who need such skills. This study offers an innovative, user-friendly framework that uses machine learning to address this difficulty. The paper studies the performance of three machine learning models, Random Forest, SVM, and XGboost, to classify, detect, and predict fraudulent transactions.

Keywords: Fraud detection, machine learning, risk management tool, data cleaning, data visualization, classification, machine learning algorithms, streamlit, random forest, XGboost, SVM, finance, detection models, dynamic environments.

1 Introduction

1.1 *Risk Management Problem*

Risk management is getting progressively more comprehensive in today's constantly developing online information technologies. Credit card payments in 2018 totaled $44.7 billion in the U.S. alone, according to The 2019 Federal Reserve Payments Study. The speed at which these transactions process is awe-inspiring. Credit cards can settle 5,000 transactions per second.[1] Moreover, Credit card usage has increased from 18% to 23%, from 2016 to 2018.[1] The rapid growth of online banking follows an increasing number of frauds. According to the 2017 Financial Institutions Payments Fraud Mitigation Survey by the Federal Reserve Bank of Minneapolis, Ninety-six percent of the respondents are debit card issuers, and 77% of credit card issuers experienced card fraud losses in 2016. Loss increases are more prevalent on debit and credit cards than on other payment types. Fraud losses increased in 2016 compared to 2015 on debit cards (63% of FIs) and credit cards (41% of FIs).[2]

Governmental institutions suffer from fraudulent transactions as well. The Institute of International Finance and Deloitte LLP White Paper report, Though this global fight against financial crime is critical, the current financial crime risk management framework is not as effective as it should or could be. For example, each year, the amount of money laundered globally is estimated to be 2% to 5% of global GDP, or between 715 billion EUR and 1.87 trillion EUR.[3] The UN estimated US$800 billion to US$2 trillion is laundered every year. But unfortunately, about 90% of this amount remains undetectable today.[4]

In this new landscape, traditional fraud detection approaches such as rule-based engines have largely become ineffective. AI and machine learning solutions using graph computing principles have gained significant interest.[5] Execution of an accurate and effective fraud detecting system is of major importance to all financial card issuing bodies. Several ways are based on approximate reasoning, AI, Data mining, sequence alignment that identifies regions of similarities, inheritable programming, etc., which are highly used in detecting these credit card frauds.[6]

The project aims to study and provide non-technical users with data processing, data visualization, and various autonomation methods for identifying and predicting fraudulent transactions. It is done using Streamlit to create a user-friendly interface and a shareable app. The application has three tabs: Data Cleaning, Data Visualization, and Risk Identification.

1.2 *Research Questions*

This paper aims to create a user-friendly environment for risk managers and other specialists in the fraud detection field, allowing them to improve their work quality and decrease the time for data preprocessing. For this purpose, the following research questions have been defined.

Research Question 1: Will the user-friendly framework let non-technical users analyze their data effectively?

Research Question 2: How do the implemented models behave in terms of fraud detection problems?

To address the above-mentioned research questions, we applied the following methods.

Addressing Research Question 1: The project was implemented using Streamlit to find the answer to the first research question. Users can clean, process, visualize, and use classification machine learning algorithms to analyze their specific data by simply choosing the corresponding tab. The application does not require any coding skills. The results are discussed in the following sections.

Addressing Research Question 2: To address the second research question, three machine learning algorithms — random forest, XGboost, and SVM—were implemented and tested on a specific real dataset. The results are discussed in the next sections.

1.3 Structure Of The Paper

The remainder of this paper is structured as follows:

Section 2, Literature Review This section provides a thorough literature review, delving into papers that address similar and the same problems. It includes detailed information about dealing with outlier removals, models used to solve fraud detection tools, and their results, ensuring the validity of our research.

Section 3, Data Collection This section provides a detailed account of the datasets used, including their characteristics, sources, and additional information. It ensures a clear understanding of the data used in our research.

Section 4, Related Work This section covers the information needed to understand the evaluation of the models that were used to accomplish the fraud detection problem. Different metrics of model evaluation are described and explained.

Section 5, Tabs This section describes the data cleaning, data visualization, and model development tabs that exist in the application. It covers the functionality information and methods that the application uses to operate.

Section 6, Conclusion And Future Work This section describes the conclusion made based on the results and contains information about the future work that can be done to develop the application and improve the results.

2 Literature Review

2.1 Literature Review For Outlier Detection

Outlier detection has a wide range of applications, including data quality monitoring, identifying price arbitrage in finance, detecting cybersecurity attacks, healthcare fraud detection, banknote counterfeit detection, and more.[7] There are several ways to deal with outliers. Box plot plots the Q_1 (25th percentile), Q_2 (50th percentile or median) and Q_3 (75th percentile) of the data along with $Q_1 - 1.5 \times (Q_3 - Q_1$ and $Q_3 + 1.5 \times (Q_3 - Q_1)$. Outliers, if any, are plotted as points above and below the plot.[8] IQR method: The data points that fall below $Q_1 - 1.5 \times IQR$ or above the third quartile $Q_3 + 1.5 \times IQR$ are outliers, where Q_1 and Q_3 are the 25th and 75th percentile of the dataset, respectively. IQR represents the inter-quartile range and is given by $Q_3 - Q_1$.[9]

2.2 Logistic Regression

Logistic regression is a technique used to predict a binary outcome variable. This technique does not demand that explanatory variables follow a normal distribution or are correlated. Using the logistic function, it models the dependent variable and predicts the probability of a target variable.[10]

The nature of these variables is dichotomous. It is represented as an equation that combines the input values linearly using the coefficient values to predict an output. The sigmoid function is used in equation[6] $S(X) = \frac{1}{1+e^x}$

Logistic regression is preferred in these scenarios to build the classifier due to its better efficiency in detecting frauds based on the data isolation provided to binary classes.[6] Advantages of Logistic regression : Logistic regression is easy to implement but more advanced than linear regression because linear regression is not good with widely distributed data. No assumptions were made regarding the distribution of classes in the feature space. It is easier to extend to multiple classes in logistic regression. It works well with the classification of unknown data.[6] Logistic regression was used to create numeric fraud detection tools. Implementing logistic regression resulted in a 92% accuracy score. F1-Score, Recall, and Precision were 0.08, 0.76, and 0.04, respectively.

2.3 Decision Tree

Decision tree is a nonlinear classification technique that divides a sample into increasingly smaller subgroups using a collection of explanatory variables. At each branch of the tree, the process iteratively chooses the explanatory variable that, by a predetermined criterion, has the strongest correlation with the outcome variable.[11] The advantage of the suggested method is that it is easy to implement, understand, and display. However, a disadvantage of this system is the requirement to check each transaction individually. Nevertheless, similarity trees have given proven results.[12] The decision tree algorithm has the benefit of not needing feature scaling, being robust to outliers, and handling missing values automatically. It is quicker to train and is very good at resolving classification and prediction problems. The decision tree uses the Gini index, information gain, and entropy as a metric for classification into two or more nodes.[11] The experiment of implementing a decision tree algorithm for predicting fraud credit card transactions resulted in a 92% accuracy score with 0.09 F1-Score, 0.93 Recall, and Precision of 0.05 estimates.

2.4 Unsupervised Learning Methods

Other than supervised learning approaches like the abovementioned Logistic regression and decision tree algorithms, there are also unsupervised learning methods. Such methods do not require labeled data. This presents a significant advantage, as fraudulent transactions are often rare, and labeling them can be expensive and time-consuming.[13] Moreover, unsupervised learning has one more advantage, which is its adaptability. Unsupervised techniques can adjust to new fraud tactics without constant retraining.[14] However, these techniques also face limitations. The main challenge is the high rate of false positives. In other words, these models may identify legitimate transactions as fraudulent or identify clusters containing both valid and fraudulent activities. Additionally, unsupervised techniques might need help to provide the granular detail required to identify specific fraudulent transactions or the individuals involved.[14]

3 Data Collection

3.1 First Dataset

As the paper aims to solve multiple challenges, two different datasets were used for the following paper. Both of them were taken from the Kaggle open-source web page. The first dataset is called Credit Card Fraud Detection. The dataset contains information about transactions made by credit cards in September 2013 for two days by European cardholders. It includes 284,807 transactions, out of which 492 were labeled as fraudulent. Due to confidentiality, the original features except ’Time’ and ’Amount’ were changed to (V1, V2, ... V28). The ’Class’ feature only takes values of 0 and 1, which labels legitimate or fraudulent transactions accordingly.

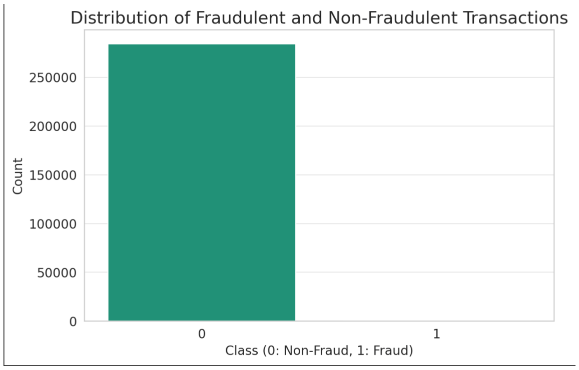

Fig. 1. Histogram of the fraud and non-fraud transactions in the dataset

Fig. 1. shows a significant imbalance between classes non-fraudulent and fraudulent transactions

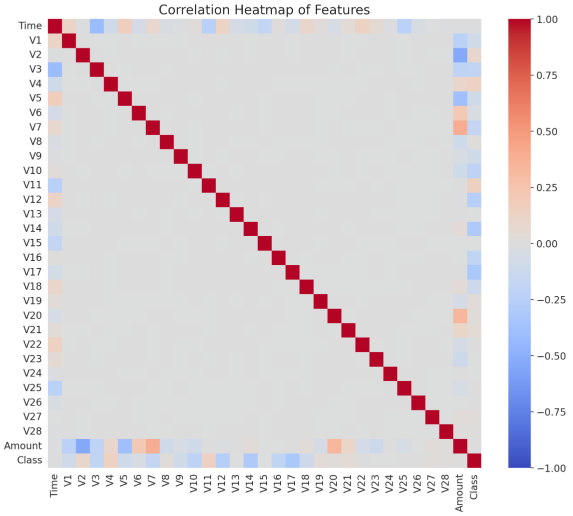

Fig. 2. Heatmap of all features in the dataset

Fig. 2. shows that the features do not show strong correlation with the feature class, which indicates that detection of fraudulent transactions requires combining multiple features

3.2 Second Dataset

As the first dataset was cleaned before publishing, the second dataset has been chosen for the data cleaning part. The dataset is called the Vehicle Dataset. It contains information about used cars for sale. The dataset's features are name, year, selling_price, km_driven, fuel, seller_type, transmission, owner, mileage, engine, max_power, torque, and seats.

4 Related Work

The model evaluation is described by the accuracy score, confusion matrix, and classification matrix, which includes precision, recall, f1-score, and support.

- Accuracy is the ratio of actual results to all occurrences. It shows the probability that the model would correctly anticipate a specific outcome out of all the predictions it has made.
- The confusion matrix is a visualization tool that displays the number of true positive, false positive, true negative, and true positive values the model predicted.
- The precision score shows the percentage of correctly predicted values. It is calculated by dividing true positive values into the sum of false positive and true positive values.
- Recall shows the percentage of the model's ability to catch positive values correctly. The recall score is calculated by dividing the true positive values by the sum of true positive and false negative values.
- F1-score shows the weighted harmonic mean of precision and recall. The formula of the f1 score is $2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$ The best score for f1 is 1, and the worst one is 0
- Support shows the number of actual class instances in the given dataset. This metric diagnoses the model's evaluation process and does not change when the model is switched.[15]

5 Tabs

5.1 First Tab

The first tab of the project provides users with the opportunity to process the data. After uploading a dataset, it finds the missing values in each column and returns the number. After getting this information, the user can fill in these missing values using methods like mean, median, mode, backward fill, or forward fill. Moreover, users can also detect and remove outliers with the Z-score method. Z-score is the number of standard deviations a variable's value is away from the variable' mean. $Z-\text{score} = (\frac{\bar{X}}{\sigma})$. Transforming a variable's values into Z-scores creates a standard normal distribution, where the average value (mean) is zero, and the spread of the data (standard deviation) is one.[8] The cut-off is set to be three, so it will capture around 99.7% of the data points.

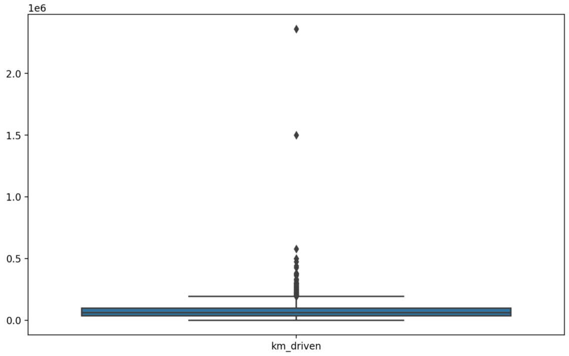

Fig. 3. Box plot before outlier removal

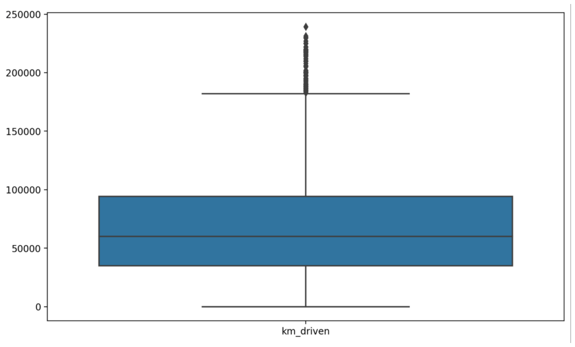

Fig. 4. Box plot after outlier removal

Further, the tab allows users to normalize or standardize data. The min-max method performs the data normalization, which maps the data into a range of 0 to 1. This makes model training less sensitive to the scale of the features, which allows the model to converge to better weights and leads to a more accurate model[16]

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

. Before implementation of a distance-based classifier like K nearest neighbors, the standardization is required to make all variables contribute equally to similarity measure [17]

$$Z = \frac{X - \mu}{\sigma}$$

.

5.2 Second Tab

The second tab of the projects is dedicated to data visualization. After processing data, users can generate different visualizations. Users are prompted to select the type of visualization, such as a histogram, bar chart, line chart, scatter plot, or heatmap. Following the selection of visualization, users can specify the corresponding columns if needed and make personal modifications like color changes or adding hue to the scatter plot. The title of the visualization is added automatically after the choice of columns. The visualizations are generated using Matplotlib and Seaborn Python libraries.

5.3 Third Tab

Users can create personalized fraud detection tools after processing data and getting helpful information from the data. As banks and companies have different standards, it is essential to have a tool that will operate on those standards. After uploading the dataset, it automatically gets encoded for future processing. Users are asked to select features on which the machine learning algorithms will train and the target variable that will be predicted. Further, users are requested to choose the ML algorithm on which they want the data to be operated. Three algorithms are provided for selecting: Random Forest, XGboost, and SVM. After training the algorithm on the data, users can see the model evaluation, which contains information about the accuracy score, classification report, and confusion matrix. After this step, users can upload new transaction datasets that are not labeled for prediction. A new column named predictions will be added to new datasets, which will label the new transactions as fraudulent or legitimate as 1 and 0 accordingly.

5.3.1 Random Forest

Random Forest is a supervised machine learning algorithm used for classification. It constructs a group of decision trees of the training data and matches them with test data. The advantage of the random forest is that it uses many decision trees to improve the predictions instead of one decision tree. As a result, the random forest employs a bagging method to generate a forest of decision trees. Given a dataset (X,Y) with N total observation where X being the predictor variables, and Y the outcome variable, the random forest algorithm first creates K_i random variables $(i = 1, 2, \ldots, N)$ to form a vector and then, it converts each K_i random vector into a decision tree to obtain the dK_i decision tree $dK_1(X), dK_2(X), \ldots, dK_N(X)$ The final classification results are as follows:[11]

$$D(X) = argmax(\sum_{i=1}^{N} dK_i(X)(K_i = \text{Fraud}) +$$

$$\sum_{i=1}^{N} d_i(X)(K_i = \text{Not Fraud}))$$

The model correctly predicted 19946 legit transactions as not fraudulent, two transactions were incorrectly predicted as fraud, 40 fraudulent transactions were correctly detected, and 12 transactions were mispredicted as legit.

Accuracy	Precision	Recall	F1-score	Support
99.93	**0.95**	**0.77**	**0.85**	**52**

Table 1. Classification report of the random forest model

5.3.2 XGboost

Extreme Gradient Boosting or XGBoost is another classification method that can perform various functions such as regression classification and ranking. The most crucial advantage of XGBoost is its scalability.[18] The algorithm works by sequentially adding weak learners to the ensemble, with each new learner focusing on correcting the errors made by the existing ones. It uses a gradient descent optimization technique to minimize a predefined loss function during training.[19] The XGBoost algorithm is known for its high accuracy compared to other algorithms. Nevertheless, the algorithm works faster by using multiple cores to build decision trees faster. It gathers information about all data points at once and then splits them up for processing across multiple cores.

The model correctly predicted 19947 transactions as legit, 1 transaction was incorrectly predicted as fraudulent, 43 fraudulent transactions were detected, and 9 fraudulent transactions were not detected.

Accuracy	Precision	Recall	F1-score	Support
99.95	0.98	0.83	0.9	52

Table 2. Classification report of the XGboost model

5.3.3 SVM

A support vector machine or SVM is another supervised machine learning model for classification or regression problems. They handle linear and nonlinear data by finding the hyperplane that best divides the data into classes. The key idea behind SVMs is to transform the input data into a higher-dimensional feature space. To do this, SVMs use a kernel function. Instead of explicitly calculating the coordinates of the transformed space, the kernel function enables the SVM to implicitly compute the dot products between the transformed feature vectors and avoid handling expensive, unnecessary computations for extreme cases.[20] The support vector machine classifies new data points better when there are clear separations between classes and is memory efficient. However, it only works well with large datasets.

The model correctly predicted 19943 legit transactions, 5 legit transactions were predicted as fraudulent, 19 fraudulent transactions were detected, and 33 fraudulent transactions were incorrectly predicted as legit.

Accuracy	Precision	Recall	F1-score	Support
99.81	**0.79**	**0.37**	**0.5**	**52**

Table 3. Classification report of the SVM model

6 Conclusion And Future Work

With the growth of online transactions worldwide, risk management, especially fraud detection, has become more complicated than ever. Non-technical solutions often show wrong results. Hence, modernization of this field is urgent -the project aimed to study and develop a user-friendly tool for non-technical risk managers and compliance officers. In conclusion, tree models were implemented and tested on the above-mentioned dataset. According to the results, the best performance showed the XGboost algorithm, and the worst result showed the SVM algorithm. In the future, time series analysis tabs can be developed to let users implement complex time series analysis models on specific datasets. Moreover, unsupervised and hybrid models can be explored and tested as well.

REFERENCES

[1] Erica Sandberg. The average number of credit card transactions per day year. *CardRates*, 2020.

[2] A Dorphy and H Hultquist. 2017 financial institution payments fraud mitigation survey. *Federal Reserve Bank of Minneapolis*, 2018.

[3] M Shepard, T Adams, A Portilla, M Ekberg, R Wainwright, K Jackson, T Baumann, C Bostock, A Saleh, and P Saplains Lagoss. The global framework for fighting financial crime enhancing effectiveness & improving outcomes. *Deloitte Report*, 2019.

[4] Rizqi Shafira Chairunnisa, Lana Shabrina, Julia Rahma, and Zaidan Allam. Tracking the money: The case of 1mdb scandal. *Global Focus*, 3(1):48–64, 2023.

[5] Eren Kurshan, Hongda Shen, and Haojie Yu. Financial crime & fraud detection using graph computing: Application considerations & outlook. In *2020 Second International Conference on Transdisciplinary AI (TransAI)*, pages 125–130. IEEE, 2020.

[6] Aditi Aditi, Aman Dubey, Ankit Mathur, and Preeti Garg. Credit card fraud detection using advanced machine learning techniques. In *2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, pages 56–60, 2022.

[7] Sadrach Pierre. A guide to outlier detection in python. *builtin*, 2023.

[8] KSV Muralidhar. Outlier detection methods in machine learning. *Towards Data Science*, 2021.

[9] CHIRAG GOYAL. Outlier detection removal — how to detect remove outliers. *Analytics Vidhya*, 2024.

[10] Ray-I Chang, Liang-Bin Lai, Wen-De Su, Jen-Chieh Wang, and Jen-Shiang Kouh. Intrusion detection by backpropagation neural networks with sample-query and attribute-query. *International Journal of Computational Intelligence Research*, 3(1):6–10, 2007.

[11] Jonathan Kwaku Afriyie, Kassim Tawiah, Wilhemina Adoma Pels, Sandra Addai-Henne, Harriet Achiaa Dwamena, Emmanuel Odame Owiredu, Samuel Amening Ayeh, and John Eshun. A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. *Decision Analytics Journal*, 6:100163, 2023.

[12] Wei Fan, Matthew Miller, Sal Stolfo, Wenke Lee, and Phil Chan. Using artificial anomalies to detect unknown and known network intrusions. *Knowledge and Information Systems*, 6:507–527, 2004.

[13] Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Yacine Kessaci, Frédéric Oblé, and Gianluca Bontempi. Combining unsupervised and supervised learning in credit card fraud detection. *Information sciences*, 557:317–331, 2021.

[14] Oladimeji Kazeem. Fraud detection using machine learning.

[15] Shivam Kohli. Understanding a classification report for your machine learning model. *medium*, 2019.

[16] Educative. Data normalization in python. *Educative*, 2024.

[17] Zakaria Jaadi. When and why to standardize your data. *builtin*, 2023.

[18] Syarifah Diana Permai and Kevin Herdianto. Prediction of health insurance claims using logistic regression and xgboost methods. *Procedia Computer Science*, 227:1012–1019, 2023.

[19] guest blog. Understanding the math behind the xgboost algorithm. *Analytics Vidhya*, 2018.

[20] Fred Tabsharani. What is a support vector machine? — definition from whatis. *WhatIs*, 2023.