Student Name : Anjali Verma  
Student Number: 501223571

**Credit Card Fraud Detection**

**Supervisor Name:** Tamer Abdou  
**Date of Submission:** October 20,2023

Table of Contents

1.Revised Abstract [3](#_Toc452457337)

**2.Literature Review …………………………………………………………………………………….5**

**2.a Introduction………………………………………………………………………………………….5**

**2.b Related Studies…………………………………………………………………………………….5**

**3. Methodology ………………………………………………………………………………………….7**

**4. Descriptive Statistics ………………………………………………………………………………8**

**5.References …………………………………………………………………………………………12**

**Dataset**  
For this project, I will use the Credit Card Fraud Detection dataset (<https://www.kaggle.com/datasets/rupakroy/online-payments-fraud-detection-dataset> )from the Kaggle website to conduct the research to fulfill the “Big Data Analytics Project (CIND820)” requirements. The dataset used in the project is titled as “Online Payments Fraud Detection Dataset”.

This dataset has 6362619 rows with main features as:

Step: Represents a unit of time where 1 step equals 1 hour.

type: Type of online transaction.

amount: The amount of the transaction.

nameOrig: Customer starting the transaction.

oldbalanceOrg: Balance before the transaction.

newbalanceOrig: Balance after the transaction.

nameDest: Recipient of the transaction.

oldbalanceDest: Initial balance of the recipient before the transaction.

newbalanceDest: The new balance of the recipient after the transaction.

isFraud: Indicates if the transaction is fraudulent or not.

**Revised Abstract**

With the rapid growth of credit card usage, credit card fraud has become a serious issue costing billions of dollars each year. Advanced fraud detection systems will be needed to identify fraudulent transactions. This research aims to develop an automated credit card fraud detection system using machine learning techniques.

Multiple supervised machine learning algorithms will be implemented, including logistic regression, Decision Tree, random forest. The models will be trained on the dataset to classify transactions as either fraudulent or valid. The best performing model will be integrated into a fraud detection application. This application will enable real-time transactions to be screened by the trained model for fraud prediction. This implementation will successfully demonstrate an accurate machine learning-based fraud detection system. The application of such a solution can significantly reduce credit card fraud losses in real-world scenarios. This work will provide a idea for developing advanced fraud prevention technologies using the latest data science techniques.

Theme

Classification - The goal of the project is to train machine learning models to classify credit card transactions as either fraudulent or valid. This is fundamentally a binary classification problem, where the models have to categorize each transaction into one of two classes.

**Research Questions**

* Which machine learning algorithms (logistic regression, SVM, random forest, etc.) are most effective at detecting different types of fraudulent credit card transactions?
* What are the most important transaction features that allow machine learning models to accurately distinguish between fraudulent and valid credit card transactions?
* Which transaction type exhibits higher incidence of fraudulent transactions ?

List of Libraries will be required

* Pandas - For data loading, manipulation and analysis. Pandas provides easy ways to explore and preprocess the transaction dataset.
* NumPy - For numerical and mathematical operations. NumPy enables efficient ops on model inputs and outputs.
* Scikit-learn - For implementing machine learning models like random forests, SVM, logistic regression etc. Scikit-learn has consistent APIs for model training, evaluation and tuning.
* Matplotlib - For visualizing data, model results and performance metrics through plots and graphs. Useful for exploratory analysis.
* Seaborn - To create more advanced statistical graphics and visualizations for data analysis and insights.

**Tool**

The software tool that will be used is Python.

Github link : <https://github.com/A100Verma/CIND-820-Capstone-Project>

**Literature Review**

Credit card fraud poses a major threat to financial institutions and causes billions in losses annually. With the growth of e-commerce and online payments, the volume and complexity of fraud has grown exponentially. In this project, I collected and analyzed a number of research papers published. Authors of these research papers applied various approaches to their research problems. They have applied different machine learning algorithms and compared the accuracy of the models to identify the most effective one.

**Related studies**

[1] J Afriyie a,∗ , K Tawiah , W Adoma Pels study the performance of three different machine learning models: logistic regression, random forest, and decision trees to classify, predict, and detect fraudulent credit card transactions. They compared these models’ performance and show that random forest produces a maximum accuracy of 96% (with an area under the curve value of 98.9%) in predicting and detecting fraudulent credit card transactions.

[2]Another comparative studyby G. Niveditha, K. Abarna, and G. v. Akshaya investigates different classification algorithms for highly skewed dataset namely logistic regression, random forest, decision trees, and naïve Bayes. According to the results, the random forest classifier has the best performance with an accuracy of 96.77%, precision of 100%, recall of 91.11%, and F1 score of 95.43%.

[3]According to Jain et al. (2020), the two transfers where frauds are most prevalent are Cash Out and the ones where money is transferred to a merchant before being transferred to users or occasionally, unknowingly, to fraudsters. The first transfer involves money being transferred from one user to another, a fraudster, or a customer. The second transfer is where frauds are most prevalent. In their assessment of various machine learning algorithms for the identification of frauds when using credit cards.

[4]Another study by Varun Kumar K S, Mr.Vijaya Kumar V G, Mr. Vijay Shankar A, and Ms. Pratibha K have used time and amount features to detect and decide whether the transaction is fraudulent or not.

[5] In another study by Kilickaya, Ozlem to handle the imbalanced data, SMOTE (The Synthetic Minority Oversampling Technique) technique was used to overcome this problem.

[6]John O. Awoyemi, Mr. Adebayo O. Adetunmbi, and Mr. Samuel A. Oluwadare have used Nave Bayes and K-Nearest Neighbourhood and logistic regression algorithms were developed, and the implementation of these algorithms has been done in Python. In order to solve the problem of data unbalancing, they have used oversampling and undersampling techniques, so the imbalance dataset will be converted into two datasets. The algorithms performances have been evaluated on the basis of various metrics.

Hence after reading the papers, I observe that solving this problem has many approaches, and I will also apply three different machine learning models: logistic regression, random forest, and decision trees to classify, predict, and detect fraudulent credit card transactions. and every model applied would lead to prediction with different accuracies from those evaluation metrics I’ll predict the best classifier to detect fraudulent transactions.

**Methodology approach**

Data Acquistion

Exploratory Data Analysis

Data Processing

Supervised Machine Learning Model Deployment

Result Analysis

Dataset:  
**Dataset link :** <https://www.kaggle.com/datasets/rupakroy/online-payments-fraud-detection-dataset>

The online payments fraud detection dataset provides a valuable resource to train machine learning models for identifying fraudulent transactions. The dataset has 11 features - step, type, amount, nameOrig, oldbalanceOrg, newbalanceOrig, nameDest, oldbalanceDest, newbalanceDest, and isFraud and isFlagged Fraud.

'step' represents the time of transaction in hourly intervals. 'type' denotes the transaction type - CASH\_OUT, PAYMENT or TRANSFER. 'amount' indicates the transaction amount.

'nameOrig' and 'nameDest' contain anonymized alphanumeric identifiers for the sender and recipient respectively. The old and new balance variables capture the account balances before and after the transaction.

The 'isFraud' variable is the target with 1 indicating a fraud transaction and 0 denoting a legitimate transaction. This reveals a highly imbalanced class distribution with just 0.172% fraudulent transactions.

The dataset provides rich features for developing supervised learning models to detect anomalous patterns in the fraudulent minority class. The high class imbalance makes it challenging to train accurate models. The anonymized customer names also limit advanced identity analytics.

**Descriptive Statistics**

**Table**

|  |  |  |
| --- | --- | --- |
| **Features** | **Description** | **Type** |
| Step | Maps a unit of time in the real world. In this case 1 step is 1 hour of time. | Numerical |
| type | CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER | Categorical |
| amount | Amount of the transaction in local currency. | Numerical |
| nameOrig | customer who started the transaction | Nominal |
| oldbalanceOrg | initial balance before the transaction | Numerical |
| newbalanceOrig | new balance after the transaction | Numerical |
| nameDest | customer who is the recipient of the transaction | Nominal |
| oldbalanceDest | Initial balance recipient before the transaction. | Numerical |
| newbalanceDest | New balance recipient after the transaction | Numerical |
| isFraud | This is the transactions made by the fraudulent agents,is fraud or not fraud | Categorical(Nominal) |
| isFlaggedFraud | The business model aims to control massive transfers from one account to another and flags illegal attempts. | Categorical(Nominal**)** |

The online payments fraud detection dataset contains 6,362,619 transactions with 11 variables. The target variable 'isFraud' indicates whether a transaction is fraudulent or not, with 1 denoting fraud and 0 denoting a valid transaction.

Looking at the transaction types, 35% are CASH\_OUT, 34% are PAYMENT while 31% are other transaction types.

The dataset exhibits class imbalance with the fraud cases (isFraud = 1) accounting for just 0.1% of the transactions. This skew is common in fraud detection datasets since illegitimate activities occur relatively infrequently compared to regular day-to-day transactions.

The dataset provides anonymized information about customers through identifiers like nameOrig and nameDest. Temporal information is captured through the 'step' variable which represents hourly time units. The old and new balance variables provide insight into account balances before and after transactions.

I also figure out that fraudulent transactions primarily consist of **'transfer' and 'cash out'** transaction types.

A graph of a bar chart

Description automatically generated with medium confidence

**Dataset Analysis**

The target feature is isFraud which is a binary feature with 0 (not fraud) and 1 (is fraud). There are 6354407 non-fraudulent transactions (99.9%) and 8213 fraudulent transactions (0.1%). As expected, most transactions are non-fraudulent. The following visualization underlines this significant contrast

A blue circle with a red line

Description automatically generated

**Histograms**

After plotting the histograms ,for all the features ,the distribution is left skewed.

A graph of a bar graph

Description automatically generated with medium confidence

**Correlation**

I checked the pearson correlation between the attributes ,found that only a correlation between Oldbalancedest and newbalancedest due to the fact of there is a change in old balance after the transaction takes place, it would automatically changes the newbalance of corresponding destination account.

A green and brown squares with white text

Description automatically generated

**Outliers**  
I identified the outliers in the dataset using boxplot and found that our dataset has numerous outliers, so outlier removal will be not a good option, it may lead to loss of valuable information.

References

**[1]** Jonathan Kwaku Afriyie,Kassim Tawiah,**,**  Wilhelmina Adoma Pels, Sandra Addai-Henne,(2023) Decision Analytics Journal, <https://doi.org/10.1016/j.dajour.2023.100163>

**[2]** G. Niveditha, K. Abarna, and G. v. Akshaya, “Credit Card Fraud Detection Using Random Forest Algorithm,” International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 5, no. 2, pp. 301–306, Apr. 2019, doi: 10.32628/CSEIT195261.

**[3]** Jain, V., Agrawal, M. and Kumar, A. (2020). Performance analysis of machine learning algorithms in credit cards fraud detection., DOI: 10.1109/ICRITO48877.2020.9197762

**[4]** Varun Kumar K S , Vijaya Kumar V G , Vijayshankar A , Pratibha K,2020, Credit Card Fraud Detection using Machine Learning Algorithms, DOI : 10.17577/IJERTV9IS070649

**[5].** Kilickaya, Ozlem, Credit Card Fraud Detection: Smote Technique to Improve Model Performance. Available at [http://dx.doi.org/10.2139/ssrn.4412674](https://dx.doi.org/10.2139/ssrn.4412674)

# [6] John O. Awoyemi, Mr. Adebayo O. Adetunmbi, and Mr. Samuel A. Oluwadare,2017, Credit card fraud detection using machine learning techniques: A comparative analysis,available at, DOI: 10.1109/ICCNI.2017.8123782.