JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PREDICTING MOBILE MONEY TRANSACTION FRAUD USING MACHINE LEARNING ALGORITHMS

a project report submitted in partial fulfillment of the requirement for the award of the degree of master of computer applications

Mr. SSP Mkunyah a student, Dr S R Raja Associate professor

Hindustan Institute of Technology & Science (Deemed To Be University) Centre for open and digital education July 2024

Abstract: With the increasing popularity of Mobile Money Transaction Services adopted in developing countries, there has been a corresponding rise in fraud and money laundering cases. Therefore, there should be a modern and effective way for combating fraudulent activities with the new technological innovation techniques. This study paper will come up with effective Machine Learning model for detecting and predicting fraud and any suspicious transaction activities in mobile money.

The study will adopt comprehensive and Case-Based -Reasoning methodologies, The comprehensive methodology will divide literature review into various stages includes mobile money background and mobile money ecosystem ,machine learning theory, research experimental design and implementation of the model which covers data collection , data prepossessing, exploratory data analysis data modelling, machine learning model classification algorithms like Logistic regression, random forests, gradient descent and also model evaluation using the Receiver Operating Characteristic (ROC) AUC (Area Under Curve) and F1-Score . The findings indicate that using supervised machine learning especially logistic regression algorithm can effectively help to detect fraud in mobile money transaction services and make mobile money service be among the tools that improve fair trade in financial sectors on developing countries that adopt the mobile money services. Through discoveries in the experiment demonstrated and illustrated, this paper shows how Machine Learning algorithms can be used to predict mobile money transaction

Introduction

The use of mobile devices have become a rudimentary part of our daily lives. The way we conduct our daily activities have become heavily dependent on mobile devices. One significant aspect of our interactions with mobile devices that cannot be overemphasized is the way we conduct financial transactions most mobile money users are not using smartphones especially in developing countries like Malawi and neighboring countries they use Conventional phones, phones that have basic functions such as calling and texting, but lack the features of smartphones. These phones uses GSM multiple-access technologies not CDMA (Code Division Multiple Access). GSM (Global System for Mobiles) and CDMA (Code Division Multiple Access) are shorthand for two older radio systems (also known as 2G and 3G) used in cell phones .Most of these phones have less security features which rely on Unstructured supplementary service data (USSD) for mobile money transaction services like paying for services banks transactions, and other e-services from our mobile phones with just dialing codes ending with the hash sign (#) to make transactions, it has made life easier for us all, but it comes along with its challenges and shortcomings, the same questions come on our mind each time the services is utilized "How secure it is", fraud has been the major challenges these e-service systems had to battle with since its inception, fraudsters take advantage of innocent users on these platforms to maximize the vulnerabilities in the system for their financial gain. The proposed system could be used by various

mobile network operators (MNO) especially in the field of mobile money services like airtel money, Telecom Networks Malawi (TNM) mpamba in Malawi and many various companies involved in mobile money services in the developing countries in Africa to fight against fraud transactions.

MACHINE LEARNING

Artificial Intelligence (AI) is a field under computer science which emphasizes on the creation of intelligent machines that work and behave like humans. Machine Learning is a sub field of AI where the concept has been defined differently by different school of thoughts.

Machine Learning Algorithms

Technology is pivotal in investigating and detecting fraudulent or laundered mobile money transactions and data mining have proven effective in detecting MMF and money laundering activities in the MMS. More specifically, ML algorithms teach computers to learn human behavior and detect patterns in the data. Supervised ML algorithms like logistic regression, decision tree, gradient descent, and random forest have all been successfully used in detecting financial fraud from labelled data. The following sections are devoted to reviewing the literature on these algorithms

Selection Of Ml Algorithms

The goal is to predict transactions flagged as fraud in mobile money transfers with help of Real-time transactions that simulate a well-known mobile transfer fraud scheme. The baseline model is Logistic regression is used as the baseline model. The study compares logistic regression with ensemble methods such as random forests and gradient descent models.

Key Findings

- The established logistic regression model performed reasonably well compared to other models.
- The random forest classifier showed outstanding performance across all measures.
- The amount of money transferred was a crucial feature for predicting money laundering transactions through mobile money transfers.

The algorithms selected for this project were explicitly chosen for their effectiveness in predicting financial crimes. These algorithms have undergone rigorous testing and have been utilized in financial crime research with notable success. From the proven reliability of logistic regression to the flexibility of random forest, the interpretability of Decision Trees, and the efficiency of gradient descent, these algorithms offer robust and diverse approaches to solving predictive problems in the complex realm of financial crimes.

RESEARCH DESIGN AND EXPERIMENTAL SETTING

Tools and Requirements

In order to come up with well experimental settings following tools and software are required:

- Dataframe(from kaggle)
- High specs computer
- Python
- Python libraries (sklearn,pandas,seaborn, nampy,multplotlib etc)
- Dataset
- Online Python editor (Google colab ,Jupiter notebook)
- Decision tree, random forest and logistic regression algorithms

Data Generation and Stimulation

Data

The data used in the study comes from Kaggle (2016) and it contains data from mobile money transactions. The dataset is synthetic and was generated by the simulator PaySim. PaySim uses data collected from a private dataset to generate a synthetic dataset that mirrors the real transactions. The sample of real transactions comes from one month of financial logs from a mobile money service. PaySim is a part of the research project called Scalable resource-efficient systems for big data analytics. This synthetic dataset was created to enrich the research in fraud detection since there is a scarcity of transaction data available to the public.

```
data = pd.read_csv("/content/ResearchDataSet.csv")

data.head()
```

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.0	0.0	0	0
1	1	PAYMENT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.0	0.0	0	0
2	1	TRANSFER	181.00	C1305486145	181.0	0.00	C553264065	0.0	0.0	1	0
3	1	CASH_OUT	181.00	C840083671	181.0	0.00	C38997010	21182.0	0.0	1	0
4	1	PAYMENT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.0	0.0	0	0

Fig: 0.1 the dataset

The dataset used in this study is a synthetic representation of mobile money transactions, designed to reflect the complexities of real-world financial activities while incorporating fraudulent behaviors for research purposes. It was derived from a simulator named PaySim, which uses aggregated data from actual financial logs of a mobile money service operating in an African country.

Exploratory data analysis (EDA)

This an important step that starts once business hypothesis is ready. This step takes 40-50% of total project time as the model outcome depends on the quality of input data being fed to train the model. Exploratory data analysis involves data attributes identification, data preprocessing and feature engineering. Attributes' identification involves identification of predictor/features variables (inputs) and target/class variable (output), along its data types (string or numeric or datetime) and classification of features into categorical and continuous variables that helps in applying appropriate treatment to be given to the variable by the algorithm while building the model

Summary statistics: Calculations like mean, median, mode, variance, and standard deviation that describe the central tendency and dispersion of the data.

```
data.describe()
```

	amount	oldbalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest	isFraud
count	1.048575e+06	1.048575e+06	1.048575e+06	1.048575e+06	1.048575e+06	1.048575e+06
mean	1.586670e+05	8.740095e+05	8.938089e+05	9.781600e+05	1.114198e+06	1.089097e-03
std	2.649409e+05	2.971751e+06	3.008271e+06	2.296780e+06	2.416593e+06	3.298351e-02
min	1.000000e-01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	1.214907e+04	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
50%	7.634333e+04	1.600200e+04	0.000000e+00	1.263772e+05	2.182604e+05	0.000000e+00
75%	2.137619e+05	1.366420e+05	1.746000e+05	9.159235e+05	1.149808e+06	0.000000e+00
max	1.000000e+07	3.890000e+07	3.890000e+07	4.210000e+07	4.220000e+07	1.000000e+00

Fig :0.2 statistical data description

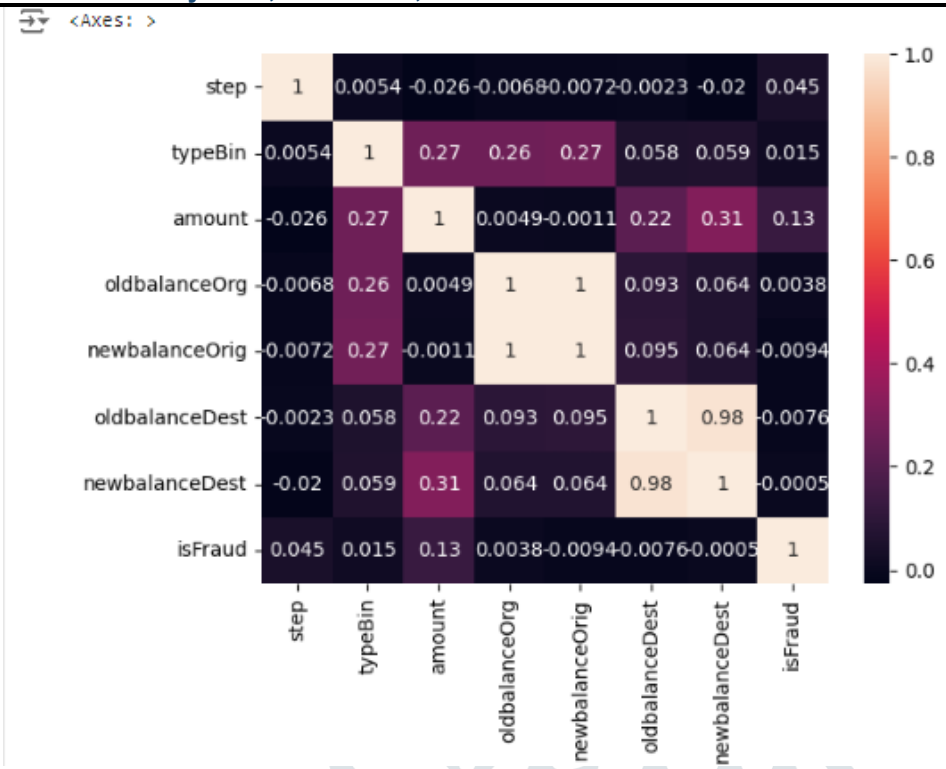

Fig :0.3 heatmap

Building, Training and Test Data

Training a model in machine learning is the process of teaching a machine learning algorithm to make predictions or decisions based on data. Whenever you build machine learning models, you will be training the model on a specific dataset (X and y). Once trained, you want to ensure the trained model is capable of performing well on the unseen test data as well. The train test split is a way of checking if the ML model performs well on data it has not seen. This is applied to supervised learning problems, both classification and regression.

In this case the dataset has gone through data import from data source, data cleaning, feature engineering and Exploratory Data Analysis (ED)

IMPLEMENTATION OF SELECTED MACHINE LEARNING MODELS

The methodology was implemented in python

Logistic regression

Logistic regression will be used as the baseline algorithm to compare with the other models. Logistic regression uses a linear combination of input variables (*x)* to predict an output variable (*y)*. The output variable is usually (0 or 1), representing the two possible outcomes of a binary classification task (e.g., fraud or not fraud). The coefficients of the input variables (b) are estimated using maximum likelihood estimation. The Sigmoid function is a mathematical function that is the foundation of logistic regression and takes an actual number and translates it into a value between 0 and 1. The Sigmoid translation is important for ML learning classification tasks because it allows the algorithm to easily separate data points into different classes. The sigmoid function is denoted in equation 1 below.

Where

F(*x)* is the value bounded between 0 and 1

X is he derivative of the sigmoid function

e is the mathematical constant

$$f(x) = 1/(1 + e^{\wedge} - x) \qquad eq.1$$

The output of the sigmoid function can be interpreted as a probability. For example, if the output of the Sigmoid function is 0.8, this can be interpreted as an 80% chance that the data point belongs to one class and a 20% chance that the data point belongs to the other class. Using the Sigmoid function, a logistic regression model can be trained to predict the class to which a new data point belongs. The logistic regression classifier uses the Sigmoid function to estimate the probability that $y = 1$, given the size of x. Equation 2 below denotes the logistic regression model.

Where

Y = values between 0 and 1,

$$^{0}_{e}\beta + \beta 1 + \mathrm{x}$$

represents the independent features, and *B0* and *B1* will give different estimations of

$$pr(y = x1_{j^x}\ x) = 1 + e\beta\ e\,\beta \quad 0{+}0{+}\,\beta 1 + x \qquad eq.2$$

Decision tree

Another useful ML algorithm for fraud detection is the Decision Tree classifier.Decision Tree algorithms are highly effective in identifying fraud and often employ other methods such as rule-based systems and artificial neural networks.In this project, the data is labeled using the Gini Index. The mathematical formula for the Gini Index is shown in Equation (3):

$$\mathrm{IG}(f) = \sum\nolimits_{-}(k = 1)^{\wedge} mfk(1 - fk)\,,$$

Random forest

The random forest classifier has generated significant interest in fraud detection and money laundering research due to its ability to accurately identify unusual patterns in the data.For instance, fraudsters may create multiple accounts with different email addresses and make small purchases to evade detection. Alternatively, they may attempt to return items they never purchased to receive refunds. By identifying these and other irregular patterns, the random forest algorithm can aid in detecting fraud before it leads to significant losses

Model Evaluation

Model evaluation is the process that uses some metrics which help us to analyze the performance of the model. As we all know that model development is a multi-step process and a check should be kept on how well the model generalizes future predictions. Therefore evaluating a model plays a vital role so that we can judge the performance of our model. The evaluation also helps to analyze a model's key weaknesses. There are many metrics like Accuracy, Precision, Recall, F1 score, Area under Curve, Confusion Matrix, and Mean

Square Error. In this context we adopt Receiver operating characteristics-Area under Curve (ROC -AUC) explained belowin this study two model evaluation techniques were used Receiver operating characteristics-Area under Curve (ROC -AUC) and F1-Score.

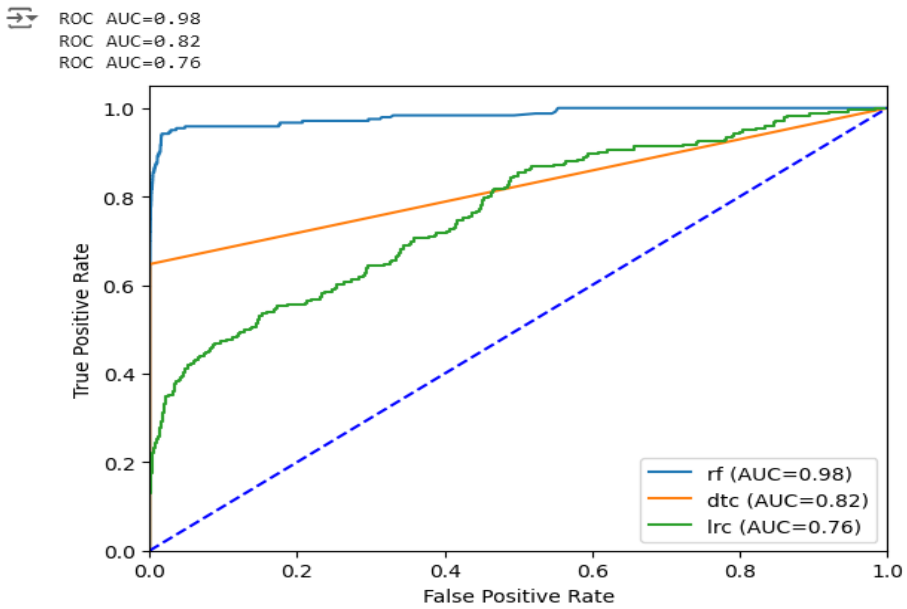

Fig 0.4 : ROC –AUC evaluating three models used in this research paper

AUC-ROC Curve

AUC (Area Under Curve) is an evaluation metric that is used to analyze the classification model at different threshold values. The Receiver Operating Characteristic (ROC) curve is a probabilistic curve used to highlight the model's performance. The curve has two parameters:

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Fig : 0.5 Receiver Operating Characteristic (ROC) curve

TPR: It stands for True positive rate. It basically follows the formula of Recall.

FPR: It stands for False Positive rate. It is defined as the ratio of False positives to the summation of false positives and True negatives. This curve is useful as it helps us to determine the model's capacity to distinguish between different classes. Let us illustrate this with the help of a simple Python.

F1-Score

F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset. This can be a reliable metric only if the dataset is class-balanced; that is, each class of the dataset has the same number of samples.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Fig :0.6 F-1 Score

However, f1-score works better in imbalanced but it is used in this study because our data is synthetically generated by simulator. Here is python code for f1-score

ACKNOWLEDGEMENTS

First and foremost, praises and thanks to God the Almighty, for the blessings throughout my project work to complete the project successfully. I express my gratitude to all the authorities who were instrumental in making this project successful. I am thankful to Head of the Department, CODE for support and encouragement to complete my work. I am highly grateful to mu supervisor, for stimulating guidance, help and useful suggestions throughout the present work. I would like to express my gratitude to the project coordinator for providing valuable comments and hints to improve my project work. I would like to dedicate this project to, my amazingly loving and supportive family who has always been with me, no matter where I am.

FUTURE ENHANCEMENTS

In this study paper, Supervised Machine Learning especially logistic regression algorithm is a valuable algorithm used for fraud detection, other algorithms can also be implemented with python libraries for further investigation and research on mobile money transaction fraud. Case based reasoning has also introduced in this paper because by keeping and retrieving various cases about mobile money transaction fraud could also help to find a proper machine learning way of detecting fraudulent activities. Although we were able to develop a high precision model, we heavily relied on patterns in the entirety of the mobile money transactions dataset. Consequently, neural networks and deep learning algorithms could perform similarly or worse than the machine learning techniques explored. One thing that may be considered to improve the accuracy and precision in future work is leveraging patterns in the information in individual subscriber transactions.

The limitation that must be considered in this study document is that the study is limited to PaySim synthetic dataset structure, future works may request actual MMO transaction data to develop a more realistic fraud detection model. Also, the security layer and the techniques explored may only be efficient in certain types of frauds. Frauds that may not result in account balance inconsistencies, and benign-looking fraudulent transactions may go through our proposed security undetected.

Future research can employ ML and AI to identify behavior patterns that may indicate fraudulent activity, while human intelligence can provide context and insights that an automated system may miss. By combining these two approaches, future research can significantly improve the fight against fraud.

Link to the Code

https://github.com/sabali-sez/predicting-mobile-money-transaction-fraud./blob/main/complete%20project%20code.ipynb

REFERENCE

1. Arsene. (2022). *Synthetic MTN mobile money dataset - rwanda*. Kaggle.com. https://www.kaggle.com/datasets/amuhire/synthetic-mtn-momo-data-rwanda

2. Association of Certified Fraud Examiners. (2023). *Fraud 101: What is fraud?* Www.acfe.com. https://www.acfe.com/fraud-resources/fraud-101-what-is-fraud

3. Bagga, S., Goyal, A., Gupta, N., & Goyal, A. (2020). Credit Card Fraud Detection using Pipeling and Ensemble Learning. *Procedia Computer Science*, *173*, 104–112. https://doi.org/10.1016/j.procs.2020.06.014

4. Bruehl, C. (2023, November 14). *The hardest part: Defining A target for classification*. Medium. https://towardsdatascience.com/the-hardest-part-defining-a-target-for-classification-50c34d37e0b8

5. Chugh, V. (2024, September 10). *AUC and the ROC curve in machine learning*. Datacamp.com; DataCamp. https://www.datacamp.com/tutorial/auc

6. *Classifier*. (2019, May 17). DeepAI. https://deepai.org/machine-learning-glossary-and-terms/classifier

7. Collins Ogombo. (2023). *Kenya's mobile money payments data*. Kaggle.com. https://www.kaggle.com/datasets/collinsogombo/kenyas-mobile-money-payments-data

8. *Correlation-based feature selection in python from scratch – johannes S. fischer*. (n.d.-a). https://johfischer.com/2021/08/06/correlation-based-feature-selection-in-python-from-scratch/

9. *Correlation-based feature selection in python from scratch – johannes S. fischer*. (n.d.-b). https://johfischer.com/2021/08/06/correlation-based-feature-selection-in-python-from-scratch/

10. DataScience-ProF. (2024a, March 20). *A comprehensive guide to feature selection using variance threshold in scikit-learn*. Medium; **AI monks.io**. https://medium.com/aimonks/a-comprehensive-guide-to-feature-selection-using-variance-threshold-in-scikit-learn-0b10146aa71f

11. DataScience-ProF. (2024b, March 20). *A comprehensive guide to feature selection using variance threshold in scikit-learn*. Medium; **AI monks.io**. https://medium.com/aimonks/a-comprehensive-guide-to-feature-selection-using-variance-threshold-in-scikit-learn-0b10146aa71f

12. Gilman, L., & Joyce, M. (n.d.). *Managing the Risk of Fraud in Mobile Money*. https://www.gsma.com/solutions-and-impact/connectivity-for-good/mobile-for-development/wp-content/uploads/2012/10/2012_MMU_Managing-the-risk-of-fraud-in-mobile-money.pdf

13. Jain, A. (2023, August 21). *Types of machine learning*. GeeksforGeeks. https://www.geeksforgeeks.org/types-of-machine-learning/

14. James, L. (2024). *Risk management in mobile money : observed risks and proposed mitigants for mobile money operators*. World Bank. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/155161501149762325/risk-management-in-mobile-money-observed-risks-and-proposed-mitigants-for-mobile-money-operators

15. Kazhing'a, P. (2023). *Mobile money agents summary*. Kaggle.com. https://www.kaggle.com/datasets/prestonkazhinga/mobile-money-agents-summary

16. Lokanan, M. (2023). Predicting mobile money transaction fraud using machine learning algorithms. *Applied AI Letters*, *4*(2). https://doi.org/10.1002/ail2.85

17. *Model deployment: Considerations, benefits & best practices*. (n.d.-a). Www.markovml.com. https://www.markovml.com/blog/model-deployment

18. *Model deployment: Considerations, benefits & best practices*. (n.d.-b). Www.markovml.com. https://www.markovml.com/blog/model-deployment

19. *ROC Curves & AUC: What Are ROC Curves | Built In*. (n.d.). Builtin.com. https://builtin.com/data-science/roc-curves-auc

20. Segan, S. (2022, January 20). *CDMA vs. GSM: What's the Difference?* PCMAG; PCMag. https://www.pcmag.com/news/cdma-vs-gsm-whats-the-difference#The_Technology_Behind_CDMA_and_GSM

21. *What is: Variance threshold*. (2024a). LEARN STATISTICS EASILY. https://statisticseasily.com/glossario/what-is-variance-threshold-in-data-science/

22. *What is: Variance threshold*. (2024b). LEARN STATISTICS EASILY. https://statisticseasily.com/glossario/what-is-variance-threshold-in-data-science/