National College of Ireland

National College of Ireland

Project Submission Sheet

Student Name:	Pranav Dinesh Ahire.		
Student ID:	X23302372		
Programme:	Master of Science in Data Analytics	**Year:**	2024-2025
Module:	**Domain Application CA Project**		
Lecturer:	Vikas Sahini		
Submission Due Date:	24-07-2025		
Project Title:	AI-Based Fraud Detection in Mobile Transaction Using Isolation Forest and Random Forest.		
Word Count:	……2855……		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Pranav Dinesh Ahire

Date: 24-07-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

Domain Application CA Project

AI-Based Fraud Detection in Mobile Transaction Using Isolation Forest and Random Forest

Your Name/Student Number	Course	Date
Pranav Dinesh Ahire	Master of Science in Data Analytics	24-07-2025

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool
NA	NA	NA

Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

[Insert Tool Name]	
[Insert Description of use]	
[Insert Sample prompt]	[Insert Sample response]

Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence:

[Place evidence here]

Additional Evidence:

[Place evidence here]

AI-Based Fraud Detection in Mobile Transaction Using Isolation Forest and Random Forest

Pranav Dinesh Ahire
x23302372
MSCDAD_B- Domain Application CA Project
National College of Ireland

Abstract

Fraudulent transaction detection has been found out to be a very essential problem with the flourishing mobile financial service industry since, there is a lot of data that has to be processed that has minimal cases of fraud. The present paper illustrates a comparison of two machine learning models Isolation Forest which is unsupervised anomaly detection model, and the Random Forest which is a supervised classification model to identify the fraudulent transactions among the mobile money transactions. The search is based on PaySim dataset to discuss the assignment of data preprocessing and visual data to detect the pattern of fraud crimes. The first imbalanced dataset is utilized in Isolation Forest and the second one through imbalanced dataset retrieved by downsampling method is taught in Random Forest. Based on the results, as observed it can be said that Isolation Forest has a high overall accuracy, but a low performance based on the fraud-detection since there is a class imbalance. The Random Forest, however, is significantly more precise, recalling and achieving the F1-score and, therefore, the more cost-effective solution, in the real-life scenario, of fraud detection. The results of the experiment reveal that in the detection of fraud on financial data, class balancing supervised method is significant.

1. Introduction

1.1 Overview of the Topic

Due to the financial system in the world increasingly becoming digitalized, mobile money services have emerged as a key instrument of financial inclusion especially in the underbanked areas like M- Pesa. These platforms have the capacity to accommodate millions of transactions each day through the convenience and speed that they present. But they also provide an expanding exposure area to the fraudulent transactions like improper transfer, imitation and synthetic fraud. The fraud detections scheme is based on the conventional rule-based models that are impotent to match the fraud geniuses in terms of complexity, large volume and fluctuating trends.

The financial institutions are addressing this dilemma through uncovering any suspicious activities in real-time using machine learning (ML) models as well as anomaly detection models. ML models in their normal forms have the capacity to process large amount of transactional data and derive certain patterns that are not identifiable using rule-based systems. However, it is not an easy job to define where the fraud is present as the asymmetry with normal transactions is too high and the selection of algorithms and the method of processing them should be reasonable.

1.2 Research Objectives

The objective of conducting this study is to design and test a machine learning program that can be used to detect frauds workers based on mobile transaction dataset generated by PaySim that models the real-life financial system.

The specific objectives are:

- The purpose of this is to apply the unsupervised model of detecting anomalies known as Isolation Forest to highlight the suspicious transactions in the skewed dataset.
- To even out the class imbalance by doing downsampling and dividing a Random Forest supervised classifier on the even classes.
- To compare the two models, in terms of precision, recall, F1-score and accuracy

based on the isFraud label as our after evaluation.

- To establish the applicability of the individual models to the business and which of the specified techniques will be more suitable to the detection of actual frauds within the financial establishments.

1.3 Research Questions

How well can machine learning models (particularly Isolation Forest, and Random Forest) identify fraudulent transactions in a mobile money system and how will class balancing influence the performance of a model?

2. Background and Scope

They have developed such fears like financial frauds in institutions where there is a huge presence of real time transactions. The relatively old rule-based systems cannot track the emerging form of fraud as in the case of the mobile financial system such as M-Pesa whereby the access point is decentralized, and the transactions are numerous and diverse [2]. Practical tool to test the possibilities to detect fraud can be offered with PaySim mobile money transaction simulator. It has more than 6 million of transactions that are attributed to uneven labels on which a small percentage are real criminal acts. It is this asymmetry that means it is hard to learn anything useful with common classifiers that has not been heavily pre-processed, or model pruned.

Within the framework of this project, the authors consider the question that implies the necessity to apply machine learning model to detect fraud in mobile money systems. Not only it refers to the problem of using the unsupervised method of detecting the abnormalities (Isolation Forest) but also compares it with the advantages of the supervised model (Random Forest), and with the balanced data based on which the latter had been trained [2][3]. The prevention of the fraud systems is as well mentioned in the project whereby there is a remark on how the outcomes of the said models can be applied in real life.

3. Project Goal & Business Value

3.1. Project Goals

- Apply an unsupervised learning (Isolation Forest) algorithm in detecting anomalies in mobile transaction data that have no prior labels.
- Take care of the class imbalance through downsampling and employ a Random Forest classifier when conducting a comparative supervised learning.
- Compare the two models according to such measures as accuracy, precision, recall, and F1-score.
- Make a clear comparison of the model performance to know the best way to proceed to detect the fraud.

3.2. Business Value

- Improve the prevention of financial frauds at an early stage and reduce loss of monetary resource and reputation.
- Improve on the work done by the fraud analysts by reducing the count of false positives and thereby granting them an opportunity to make checks manually.
- Allow real-time fraud detection systems scaling to be explained.
- Present better performance once the data balancing increases the performance of the method of detecting fraud in real life practices.

4. Literature Survey

Ever since digital and mobile financial services have been on the rise, fraud detection has been a major concern by the financial institutions. Since the traditional rule-based systems are rather inefficient as far as the detection of new and adaptive fraudulent behaviours is concerned, machine learning (ML) has proven to be the scalable and smart alternative. Various studies and application in the industry show that the increasing use of artificial intelligence (AI) and ML techniques in identifying anomalies of the financial transactions is becoming inevitable.

The isolation forest suggested by Liu et al. (2008) is one of the basic models of unsupervised anomaly detection. Isolation Forest differs with density-based approach of isolating observations through random selection of a feature followed with a split

value. A measure of an observation is the number of splits needed to isolate the observation (this is an anomaly score) [2]. The most notable strength of this model is its efficiency of working with data of high dimensions and it does not require labelled data to identify rare and outlier events. This qualifies it especially to be used in areas such as fraud detection where fraudulent transactions are few and difficult to be tagged with [4].

Nonetheless, unsupervised models alone might not be very effective to use in these cases when there is some access to a labelled set of data or where the interpretability and reliability are essential. Classifiers Just like Random Forest classifiers, supervised models are well known in the fraud detection documents because of their robustness and high level of accuracy [3]. Thus, these models take advantage of ensemble learning, decreasing overfitting as well as having a method of ranking feature relevance and explaining predictions. The power of Random Forest has been proved in many studies together with such class balancing methods as SMOTE (Synthetic Minority Over-sampling Technique) or under-sampling[3].

PaySim dataset used in this study was created to simulate mimicked mobile money transactions based on real financial systems such as M-Pena. It is a feasible method to examine the topic of detecting fraud in a controlled setting. Although synthetic, PaySim maintains realistic transactional behaviour and imbalance between fraudulent and non-fraudulent transactions and as result has become a commonly used benchmark in the research of fraud detection [4]. Multiple ML methods have been tried on PaySim: logistic regression, decision tree and neural networks, and characteristic of PaySim has been mentioned in all of these, which is that very noisy data can be highly skewed, making the detection of some rare partition's problematic [5].

Financially, the banks have been on their toes trying to implement the AI-driven applications to enhance fraud protection systems. One of them is the JPMorgan Chase that uses AI models to identify suspicious behaviour with the use of transactional patterns as analysis in the form of greater accuracy [8]. The Mastercard Decision Intelligence is a risk measurement methodology that can accurately quantify the risk in real-time and pre-emptive the fraudulent transaction [6]. Similarly, HSBC applies voice biometrics to authentication of its users, which seeks to prevent cases of identification fraud by further expanding the applications of AI past the vision of a transactional analysis into biometric protection.

These revelations indicate the existence of meeting ground between research study and actual uses of machine learning in detecting frauds. Both are giving much attention to the fact that the combination of domain knowledge, data preprocessing and model selection is required not only to generate precise systems, but also interpretable and scalable systems [6],[7]. The current project goes further to show these ideals by comparing and testing an unsupervised (Isolation Forest) and a supervised (Random Forest) approach to the problem of detecting fraud in the financial services that is part of the larger debate on the extent to which we can make fraud detection intelligent in the financial services.

5. Ethical Concerns

Although based on a synthetically generated dataset (PaySim), this project simulates the process of detecting fraud in the real world of finance and begs several ethical questions. Privacy of data should also be maintained in the real world, where regulatory act such as the GDPR should be observed. Fairness and bias are essential since the model that was trained using unbalanced data can be biased and discriminate against certain users or forms of transactions.

Also, high false positive rate may negatively affect the customer confidence by locking genuine transactions. Hence, models should trade-off on accuracy and user experience. Grounds like transparency and explainable should be used to justify flagged transactions particularly in regulated business. Finally, human control and surveillance should be integrated into the components of fraud detecting systems to maintain responsible and accounted AI application.

6. Methodology

This project featured the use of two machine learning methods namely Isolation Forest (unsupervised anomaly detection) and Random Forest (supervised classification) to detect the fraudulent transactions in the PaySim data.

6.1. Data Preprocessing

Preliminary exploratory data analysis (EDA) found that fraudulent transactions highly skewed and worked in the TRANSFER and CASH_OUT of transaction types. This imbalance was brought out in a class distribution bar chart. The type of transactions was encoded with a label, and such selected numerical features as the amount, oldbalanceOrg, newbalanceOrig, oldbalanceDest, and newbalanceDest were scaled through standard Scaler.

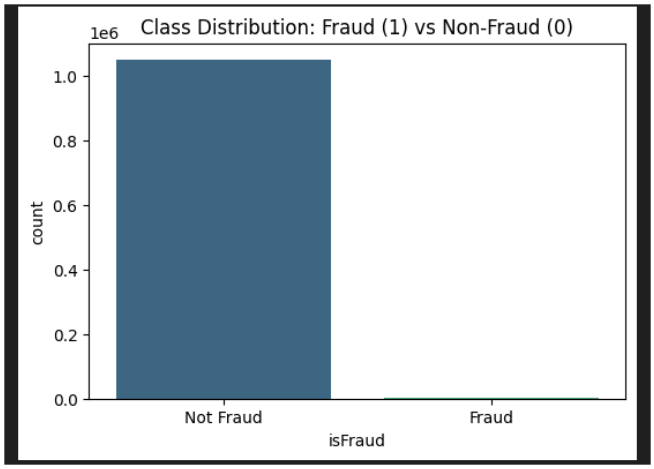

Figure 1. Class Distribution Bar Chart.

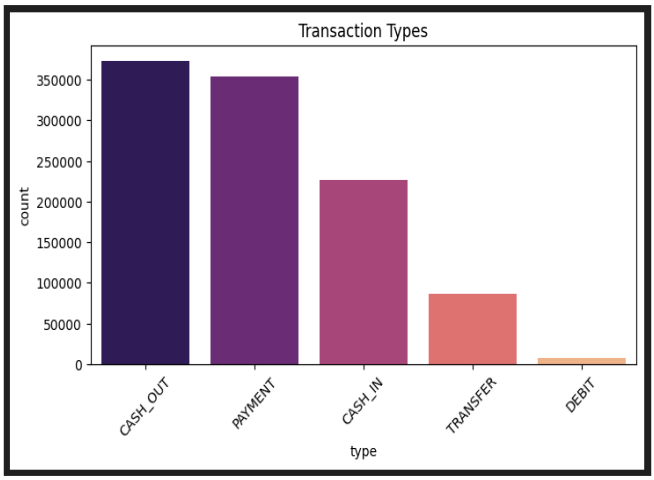

Figure 2. Distribution of Transaction Types

A heatmap to measure multicollinearity amongst variables was created as well. This showed good correlations between origin and destination balances and amounts of transactions, which assisted in the determination of features to be chosen.

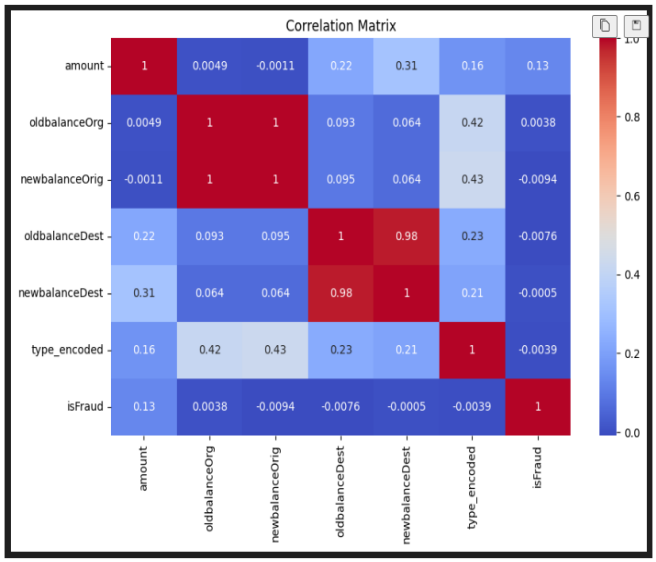

Figure 3. Correlation Matrix of Transaction Features.

6.2. Insolation Forest (Unsupervised)

In anomaly detection, the Isolation Forest model uses all the features (not the target label) in its training. It could walk through each transaction and would score an anomaly and predict fraud using isolation depth. This is even though in this case, the training data was very skewed (~0.13 percent of fraud), the model has failed to distinguish the fraud cases well.

A confusion matrix and a classification report suggested that this model was able to generate just almost 99.79 percent of accuracy, although its other questionable parameters, precision and recall were less than 0.24 percent. This showed the weakness of direct transfer of unsupervised models to an imbalanced data which used financial data.

```
=== Classification Report ===
              precision    recall  f1-score   support

           0     0.9989    0.9990    0.9990   1047433
           1     0.0019    0.0018    0.0018      1142

    accuracy                         0.9979   1048575
   macro avg     0.5004    0.5004    0.5004   1048575
weighted avg     0.9978    0.9979    0.9979   1048575
```

```
=== Confusion Matrix ===
[[1046391    1042]
 [   1140       2]]
```

Figure 4. Classification Report and Confusion Matrix of Isolation Forest (Unbalanced Dataset).

6.3. Class Imbalance Handling

To enhance the performance in detecting frauds, the dataset was balanced through downsampling where a small sample of the records that part of the frauds was not was randomly sampled to allow an equal number as the number of the fraudulence transaction. A post-balancing and a pre-balancing side-by-side plot of the distribution of classes proved the success of this technique.

```
Balanced class distribution:
isFraud
1    1142
0    1142
Name: count, dtype: int64
```

Figure 5. Class Balance After SMOTE Downsampling.

6.4. Random Forest (Supervised)

It was the balanced dataset that was applied to train the Random Forest classifier. The data was split in to training and test (70/30) and the model was tested on ground of standard classification parameters. The accuracy was 97.38 percent, precision was 96.7 percent, recall was 98.32 and it had 97.51 on F1-score. These scores indicated massive increase in the performance of detecting the fraud compared to the unsupervised model.

```
=== Classification Report ===
              precision    recall  f1-score   support

           0     0.9814    0.9634    0.9723       328
           1     0.9670    0.9832    0.9751       358

    accuracy                         0.9738       686
   macro avg     0.9742    0.9733    0.9737       686
weighted avg     0.9739    0.9738    0.9737       686
```

```
=== Confusion Matrix ===
[[316  12]
 [  6 352]]
```

Figure 6. Classification Report and Confusion Matrix of Random Forest (Balanced Dataset).

To make this comparison evident a bar chart was graphed on the grounds of accurateness, precision, recollection, F1-score of the two models. The plot of data proved the truth that the supervised approach with class weighting was more effective to detect the fraud in the current case.

7. Evaluation

The area where the two models were benchmarked took four measures of performance that include, accuracy, precision, recall and F1-score.

Isolation forest, although capable of providing high accuracy, did not detect real frauds because not all transactions were frauds. Quite the opposite, on a balanced set, the Random Forest was also accurate on all the assessment metrics proving it has potential to discover rare cases of fraudulent transactions.

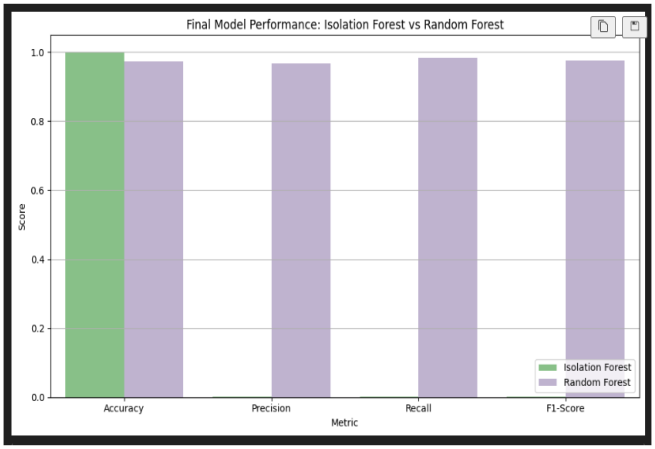

Figure 7. Final Model Performance Comparison — Isolation Forest vs. Random Forest (Based on Accuracy, Precision, Recall, F1-Score).

8. Conclusion and Business Interpretation

8.1. Conclusion

This is a successful attempt to study two machine learning algorithms (Isolation Forest and Random Forest) and apply them in describing fraudulent transactions in cell phone money system based on PaySim dataset. This paper has identified an important issue in trying to detect fraud, the tremendous imbalance in the classes; the fraudulent transactions are only a small percentage of the database.

The Isolation Forest however, even though, being efficient and unsupervised, it had also presented a high accuracy in detecting cases of fraud, due to its extremely low recall and precision. This indicates that unsupervised combinations, in spite of being helpful in locating anomalies in cases with general background, might not work sufficiently in a seriously imbalanced and situation-specific circumstance.

With the distribution of the dataset balanced through downsampling, the Random Forest classifier was far more successful, with an excellent precision (96.7%), recall (98.3%), F1-score (97.5%), equaling and surpassing the performance of all its classifier counterparts. That supervised models are much better at detecting fraudulent patterns, provided representative, balanced data is learned on, is now confirmed.

8.2. Business Interpretations

The results of this project have significant implications in the financial institutions:

- Fraud can be easily and accurately detected at its initial stages using machine learning, which prevents a significant sum of the financial loss and operational costs.
- In the scenario when we know historical fraud labels of course in well balanced manner, supervised models like Random Forest are better.
- This implies that recall via double-key operation will be high hence most of the fraudulent transactions will be identified and precision will be high decreasing the number of false alarms to the legit users.
- This implies that recall via double-key operation will be high hence most of the fraudulent transactions will be identified and precision will be high decreasing the number of false alarms to the legit users.

To conclude, the presented project shows a scalable, interpretable, and efficient pipeline of fraud detection that can be incorporated into mobile money systems in the real world.

9. References

1. Dataset link: https://www.kaggle.com/datasets/ealaxi/paysim1
2. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," *2008 Eighth IEEE International Conference on Data Mining*, Dec. 2008, doi: https://doi.org/10.1109/icdm.2008.1
3. M. S. Kumar, V. Soundarya, S. Kavitha, E. S. Keerthika, and E. Aswini, "Credit Card Fraud Detection Using Random Forest Algorithm," *IEEE Xplore*, Feb. 01, 2019. https://ieeexplore.ieee.org/document/8824930
4. A. Owen, O. Emma, and R. Adams, "Evaluating the Role of Class Imbalance in Fraud Detection Algorithms," *ResearchGate*, Oct. 07, 2024. https://www.researchgate.net/publication/388817710_Evaluating_the_Role_of_Class_Imbalance_in_Fraud_Detection_Algorithms (accessed Jul. 14, 2025).
5. P. Gupta, A. Varshney, M. R. Khan, R. Ahmed, M. Shuaib, and S. Alam, "Unbalanced Credit Card Fraud Detection Data: A Machine Learning-Oriented Comparative Study of Balancing Techniques," *Procedia Computer Science*, vol. 218, pp. 2575–2584, 2023, doi: https://doi.org/10.1016/j.procs.2023.01.231.
6. Team DigitalDefynd, "5 Ways MasterCard is Using AI [Case Study][2024]," *DigitalDefynd*, Dec. 10, 2024. https://digitaldefynd.com/IQ/ways-mastercard-use-ai/
7. Team DigitalDefynd, "5 Ways MasterCard is Using AI [Case Study][2024]," *DigitalDefynd*, Dec. 10, 2024. https://digitaldefynd.com/IQ/ways-mastercard-use-ai/
8. J.P. Morgan, "AI Boosting Payments Efficiency & Cutting Fraud | J.P. Morgan," *www.jpmorgan.com*, 2023. https://www.jpmorgan.com/insights/payments/payments-optimization/ai-payments-efficiency-fraud-reduction