Universitat
de les Illes Balears

**MASTER'S THESIS**

# EXPLORATORY STUDIES LAIA@UIB: AI MODELS FOR THE DETECTION OF FRAUDULENT PAYMENTS


**Melih Kurtaran**

**Master's Degree in Intelligent Systems (MUSI)**

**Specialisations: Artificial Intelligence, Computer Vision**

**Centre for Postgraduate Studies**


**Academic Year 2022-2023**

# Exploratory Studies LAIA@UIB: AI Models for the Detection of Fraudulent Payments

Melih Kurtaran
**Tutors:** Antoni Jaume Capó, Isaac Lera Castro
Trabajo de fin de Máster Universitario en Sistemas Inteligentes (MUSI)
Universitat de les Illes Balears
07122 Palma, Illes Balears, Espanya
melih.kurtaran1@estudiant.uib.cat

ABSTRACT

This thesis investigates the use of AI models for detecting fraudulent payments in electronic payment systems. The main challenges in the development of such models are the lack of labeled data, the need to balance minimizing false positives while maximizing true positives, and the complexity of financial transactions. This study aims to explore the performance of different machine learning and deep learning algorithms, such as logistic regression, neural networks, XGBoost, and Random Forest, in detecting fraudulent payments, and to develop techniques to address data scarcity and imbalance. The research involves experimentation with two datasets, one real and the other artificially generated, both exhibiting a high degree of imbalance. The study findings can enhance the development of trustworthy and effective AI models for the detection of fraudulent payments, contributing to enhancing security measures within financial systems.

***Index Terms*—Credit Card Fraud Detection, Machine Learning, Neural network, Data Imbalance, Re-sampling Methods**

## I. INTRODUCTION

The detection and prevention of fraudulent activities in financial transactions have become increasingly important due to the widespread use of electronic payment systems. Traditional fraud detection techniques often rely on rule-based systems, which may not be effective in identifying complex fraudulent activities. To address these challenges, the use of AI models, such as machine learning and deep learning algorithms, has gained attention as a promising approach for the detection of fraudulent payments.

However, the development and implementation of AI models for the detection of fraudulent payments present several challenges. One of the main challenges is the lack of labeled data, as fraudulent activities are typically rare events. Additionally, the models need to balance between minimizing false positives, which can lead to genuine transactions being flagged as fraudulent, and maximizing true positives, which identify actual fraudulent activities. Moreover, the complexity of financial transactions, as well as the rapidly evolving tactics of fraudsters, further complicate the development of accurate and reliable AI models for fraud detection.

To address these challenges, this thesis aims to investigate the effectiveness of AI models for the detection of fraudulent payments and to develop techniques for improving their performance. Specifically, we explore the performance of different machine learning and deep learning algorithms in detecting fraudulent payments, such as logistic regression, neural networks, XGBoost and Random Forest. We also examine techniques for addressing the challenges of data scarcity and imbalance, such as the resampling methods evaluated on a real-life highly imbalanced online credit card payments dataset by de la Bourdonnaye and Daniel [1]. Additionally, we review related work in the field, such as the survey of credit card fraud detection using machine learning by Lucas and Jurgovsky [2].

As a result of this comprehensive study, our primary objective is to significantly enhance the development of trustworthy and effective AI models specifically designed for the detection of fraudulent payments. By addressing the inherent challenges in fraud detection, we aim to mitigate the risks associated with fraudulent activities, benefiting not only financial institutions but also merchants and consumers alike.

The implementation of robust AI models for fraud detection offers numerous advantages to the entire ecosystem of electronic payment systems. Financial institutions can experience reduced financial losses resulting from fraudulent transactions, thereby improving their profitability and overall stability. With enhanced fraud detection capabilities, merchants can protect their businesses from fraudulent activities, ensuring a secure and trustworthy environment for their customers. By identifying and preventing fraudulent payments, consumers are safeguarded from potential financial losses and can have greater confidence in the security of electronic transactions.

Moreover, our research has broader implications for the overall security measures within financial systems. By advancing the state-of-the-art in fraud detection using AI models, we contribute to bolstering the integrity of electronic payment transactions. This is particularly critical in today's digital landscape, where the widespread use of electronic payment systems necessitates robust security measures. Additionally, our study explores the complex nature of financial transactions and the rapidly evolving tactics employed by fraudsters. By leveraging various machine learning and deep learning algorithms, including logistic regression, neural networks, XGBoost, and Random Forest, we push the boundaries of fraud detection capabilities. This research sheds light on the performance of these algorithms and provides valuable insights

into their effectiveness in detecting fraudulent payments.

In conclusion, our study has far-reaching implications for the financial industry and electronic payment systems as a whole. By enhancing the development of trustworthy and effective AI models for detecting fraudulent payments, we aim to mitigate risks, protect financial institutions, empower merchants, reassure consumers, and enhance security measures within financial systems. Ultimately, our research aims to foster a more secure and reliable environment for electronic payment transactions, safeguarding the integrity of the entire financial ecosystem.

## II. Analysis of State of Art

A comparative analysis of various machine learning models, including Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, and Naive Bayes, has been conducted to address the topic of fraud payment detection [3]. The efficacy of these models has been evaluated using key performance metrics, including accuracy, precision, and specificity. The findings suggest that Random Forest and Logistic Regression models outperformed their counterparts in terms of all the evaluated metrics, thereby demonstrating their superiority in identifying fraudulent payment transactions.

Another comparative study was conducted to assess k-nearest neighbor (KNN), random forest and support vector machines (SVM), and deep learning methods such as autoencoders, convolutional neural networks (CNN), restricted boltzmann machine (RBM) and deep belief networks (DBN) on three distinct real-world datasets [4]. The researchers arrived at the conclusion that neural network models exhibit suboptimal performance due to their inherent reliance on larger datasets, which were not available in the present study. On the other hand, support vector machines demonstrated superior performance across the evaluated datasets. Among the neural network models, convolutional neural networks model emerged as the most effective approach, in terms of overall performance. Authors prefers using Matthews correlation coefficient (MCC) and area under the curve (AUC) metrics to compare the models.

The paper Credit Card Fraud Detection under Extreme Imbalanced Data focuses on various sampling methods to deal with imbalance problem [5]. The techniques include Random Undersampling (RUS), which randomly selects a subset from the original dataset and eliminates instances until the dataset is balanced with a disadvantage of losing valuable information. Another technique is Tomeklinks, which eliminates majority class samples where Tomeklinks are available, helping to remove overlapping in the dataset. Cluster Centroids Undersampling is another method discussed, which uses the K-Means clustering algorithm to group the dataset into clusters and calculates the mean feature vector of a random set of K instances. Authors also discuss a combination of oversampling and undersampling to handle the limitations of both techniques. Two popular methods based on this approach are SmoteTomek and Smoteen. SmoteTomek first oversamples the imbalanced dataset using the Smote technique, and then Tomeklinks are identified and removed from the oversampled dataset. Smoteen uses the Edited Nearest Neighbour (ENN) to remove all dataset instances that differ from their neighborhood, and then the Smote technique is applied to balance the dataset by creating synthetic data points.

Given the inherent class imbalance in fraud detection datasets, it is widely recognized that Accuracy is not an appropriate metric for assessing the performance of classification models. Therefore, alternative evaluation measures that are better suited for imbalanced datasets are required. Class Balanced Accuracy (CBA) has emerged as a commonly cited approach for mitigating the impact of class imbalance on model performance metrics [6]. However, some authors argue that Matthews correlation coefficient (MCC) is a more appropriate metric for handling imbalanced datasets [7]. In the present study, we employ both CBA and MCC as evaluation metrics to enable a comprehensive comparison of their respective performance when applied to fraud detection datasets.

## III. The Datasets

This study involves the experimentation of two datasets, one of which is a *real dataset* while the other is a *generated dataset*. Both datasets exhibit a high degree of imbalance, as over 99% of the data points correspond to legitimate payment transactions.

### A. Real Dataset

The dataset comprises credit card transactions made by European cardholders in September 2013, and has been obtained from the public repository, Kaggle [8]. The dataset encompasses transactions that were conducted over two days, out of which 492 instances are indicative of fraudulent activities, while the total number of transactions amounts to 284,807. Owing to the highly imbalanced nature of the dataset, the proportion of the positive class (fraudulent transactions) stands at a mere 0.172% of the entire dataset. The input variables in the dataset are solely numerical and have been obtained through Principal Component Analysis (PCA). Unfortunately, the original features and additional information about the data has not been shared due to confidentiality issues. Consequently, there arose a need for an additional dataset that incorporates the necessary features to facilitate the implementation of explanatory AI techniques. The dataset has 28 features V1 through V28 correspond to the principal components obtained via PCA, while the features `Time` and `Amount` are the only ones that have not been transformed via PCA. The `Time` feature provides information about the time elapsed between each transaction and the initial transaction in the dataset, while the `Amount` feature is indicative of the transaction amount. The `Class` feature serves as the response variable and takes a value of 1 in case of fraudulent transactions and 0 for legitimate transactions.

### B. Generated Dataset

For the generation of dataset, we have used Sparkov Data Generation Tool [9] since it is widely adopted by researchers

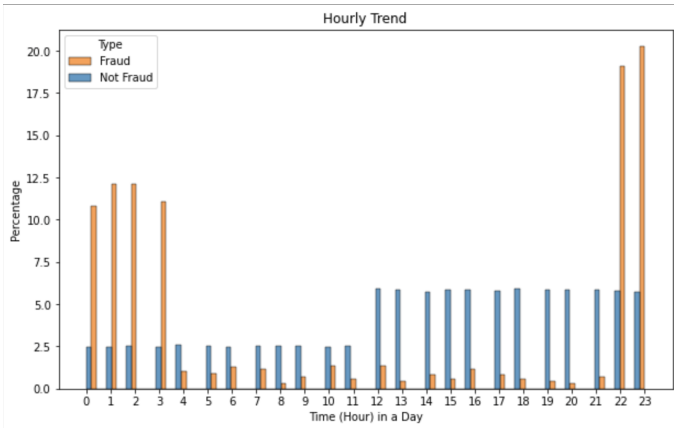


Figure 1. Percentage of frauds in hours

in the field. The tool offers a variety of customer profiles based on gender, residency in urban or rural areas and age range. We have used to generate data for 12 distinct customer profiles for the year 2022. All the generated datasets that are segregated by profiles can be readily accessed [10]. In the present study, customer profiles belonging to adults aged between 25 to 50 and residing in urban areas have been selected for experimentation purposes. The corresponding dataset consists of 60,301 transactions conducted by females and 51,663 transactions conducted by males, with only 0.776% of the transactions being indicative of fraudulent activities, thereby rendering the dataset highly imbalanced, similar to the real-world dataset scenario. The generator employs the Python library called Faker [11] for data generation.

Upon analyzing the generated dataset, it has been observed that a significant proportion of fraudulent transactions occur during late night hours. This finding is substantiated by Figure 1, which illustrates the distribution of fraudulent transactions across different hours of the day. The bar chart represents percentages rather than actual numbers. It indicates that more than 80% of fraudulent payments occur during the night, while genuine payments are more evenly distributed throughout the day, with higher frequencies during daytime hours, as expected. Furthermore, it has been observed that transactions with amounts lower than 10 dollars are mostly non-fraudulent, while fraudulent transactions tend to involve higher amounts.

## IV. Experiment

### A. Implementation

The first implementation utilizes the *real dataset* to conduct a comparative analysis of four distinct machine learning approaches with six models in total. The models in question comprise of a logistic regression model with undersampling, a neural network model with undersampling, a logistic regression model with Synthetic Minority Oversampling Technique (SMOTE) oversampling [12], a neural network model with SMOTE oversampling, an Extreme Gradient Boosting (XGBoost) [13] model without any sampling and a random forest model with SMOTE oversampling.

*1) Logistic Regression Model:* Logistic regression is a widely used statistical modeling technique for binary classification tasks. We have used two logistic regression models were evaluated: one with undersampling and one with Synthetic Minority Oversampling Technique (SMOTE) oversampling.

*1.a) With Undersampling*

Undersampling is a technique that involves the removal of instances from the majority class to balance the class distribution. The logistic regression model with undersampling was trained on a subset of the original dataset, where instances from the majority class were randomly removed until the class distribution was balanced.

*1.b) With SMOTE Oversampling*

SMOTE is a technique that generates synthetic instances of the minority class to balance the class distribution. The logistic regression model with SMOTE oversampling was trained on a modified dataset, where synthetic instances of the minority class were generated until the class distribution was balanced.

*2) Neural Network Model:* Neural networks are a class of machine learning models that are inspired by the structure and function of the human brain. We have used two neural network models: one with undersampling and one with SMOTE oversampling.

*2.a) With Undersampling*

The neural network model with undersampling was trained on a modified dataset that involved randomly removing instances from the majority class until the class distribution was balanced.

*2.b) With SMOTE Oversampling*

The neural network model with SMOTE oversampling was trained on a modified dataset, where synthetic instances of the minority class were generated until the class distribution was balanced.

*3) Extreme Gradient Boosting (XGBoost) Model:* XGBoost is a popular gradient boosting framework for classification and regression tasks. We have used an XGBoost model without any sampling was evaluated as a baseline for comparison against the other approaches and models. The XGBoost model was trained on the original dataset without any modifications to the class distribution.

*4) Random Forest Model:* Random Forest is an ensemble learning method that combines multiple decision trees to make robust predictions. In the context of fraud detection, Random Forest is a valuable tool due to its ability to handle complex, high-dimensional datasets and capture non-linear relationships between features. By aggregating the predictions of individual trees, Random Forest can effectively identify patterns and anomalies indicative of fraudulent activities. Moreover, the model's inherent feature importance estimation allows for the identification of the most influential variables in fraud detection, aiding in the interpretation and understanding of the underlying fraud indicators. Additionally, Random Forest's resilience to overfitting, provided by the combination of multiple trees and random subsampling, enhances its generalization capabilities, enabling it to effectively detect fraudulent patterns in unseen data. Thus, Random Forest holds promise as a reliable and efficient technique for enhancing fraud detection systems in various domains.

In the second implementation, we have used the *generated dataset* and have trained a random forest model [14] that utilizes SMOTE oversampling. The principal objective of this phase is to evaluate the significance of features in the decision-making process of a machine learning algorithm. The Random Forest model with SMOTE oversampling, has been tested using real-world dataset as well to assess its comparative performance against other five alternative models.

The selection of the six models in the implementation was driven by the need to evaluate different machine learning techniques and sampling methods to address the problem of imbalanced datasets. Specifically, the logistic regression and neural network models with undersampling were chosen as they represent common techniques for handling imbalanced data. The logistic regression and neural network models with SMOTE oversampling were selected as they represent popular methods for addressing class imbalance by generating synthetic examples of minority class instances. The Extreme Gradient Boosting (XGBoost) model was chosen as it has demonstrated superior performance in various classification tasks and can handle imbalanced data without the need for oversampling techniques. By comparing the performance of these six models, we seek to provide insights into the most effective machine learning approach for handling imbalanced data in the context of detecting fraudulent payments.

### B. Metrics

Given the heavy imbalance in both datasets, accuracy is not a suitable performance metric as it can lead to overfitting of the majority class by the models. To mitigate this issue, Matthews correlation coefficient (MCC) [15] and Class Balanced Accuracy (CBA) are preferred. CBA offers a comprehensive evaluation of binary classifiers as it considers both the accuracy and completeness of the predictions. MCC is valuable in that it is insensitive to imbalanced class distributions, unlike accuracy or precision, and applicable in cases where class importance varies since it accounts for all values in the confusion matrix. MCC is calculated by taking the covariance between the predicted and true binary labels, normalized by their respective standard deviations, as shown in the following formula:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{1}$$

where TP, TN, FP, and FN represent the number of true positives, true negatives, false positives, and false negatives, respectively. MCC ranges from -1 to 1, with a score of 1 indicating perfect classification, 0 indicating random classification, and -1 indicating perfect inverse classification. The advantage of MCC is that it is insensitive to imbalanced class distributions and applicable in cases where class importance varies since it accounts for all values in the confusion matrix.

Moreover, Class Balanced Accuracy (CBA) is another performance metric that is useful for evaluating binary classifiers on imbalanced datasets. CBA takes into account both the accuracy and completeness of the predictions and can be calculated using the following formula:

$$CBA = \frac{TP}{TP + \alpha(FN + FP)} \tag{2}$$

If the confusion matrix is available, CBA can alternatively be computed as the product of the sensitivity, which is the true positive rate of the minority class, and the specificity, which is the true negative rate of the majority class. This approach is also used in our implementation and can be expressed using the following formula:

$$CBA = \frac{TP_{min}}{TP_{min} + FN_{min}} \times \frac{TN_{maj}}{TN_{maj} + FP_{maj}} \tag{3}$$

where $TP_{min}$ is the number of true positives for the minority class, $FN_{min}$ is the number of false negatives for the minority class, $TN_{maj}$ is the number of true negatives for the majority class, and $FP_{maj}$ is the number of false positives for the majority class. The parameter $\alpha$ is not needed in this formulation, as the trade-off between precision and recall is implicitly balanced by the sensitivity and specificity terms.

CBA ranges from 0 to 1, with a score of 1 indicating perfect classification. CBA is a comprehensive evaluation metric that provides a balanced assessment of the classifier's performance, especially when the cost of false positives and false negatives is not equal.

Another important performance metric that is commonly used in classification tasks is the F1 score. The F1 score is a harmonic mean of the precision and recall metrics, which are calculated using the values from the confusion matrix. Precision measures the proportion of positive predictions that are actually positive, while recall measures the proportion of actual positives that are correctly predicted by the model.

The F1 score is calculated as follows:

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{4}$$

The F1 score ranges from 0 to 1, with a score of 1 indicating perfect precision and recall. Like MCC and CBA, the F1 score is also useful in evaluating classifiers on imbalanced datasets, as it considers both the false positives and false negatives.

### C. Results

The table I shows the values obtained by each ML models.

A *True Positive* refers to a fraudulent transaction that has been correctly classified as fraud, while a *False Positive* indicates a fraudulent transaction that has been erroneously flagged as genuine. False positives are dangerous and we want them to be as low as possible. *True Negative* corresponds to a genuine transaction that has been correctly identified as such, while a *False Negative* denotes a genuine transaction that has been incorrectly labeled as fraudulent. Figure 2 shows all six confusion matrices belonging the models.

It can be observed that all of the models have high accuracy values with the lowest accuracy value being 0.9826. However, it is important to note that a high accuracy value does not necessarily indicate the effectiveness of a model, particularly when dealing with imbalanced datasets. In our case, since the dataset is heavily imbalanced, accuracy is not an appropriate

Table I
MODELS COMPARISON FOR REAL DATASET

| Model | True Pos | False Pos | True Neg | False Neg | Accuracy | F1 score | MCC | CBA | Precision | Recall |
|---|---|---|---|---|---|---|---|---|---|---|
| LR undersampling | 86 | 12 | 55884 | 979 | 0.983 | 0.14 | 0.26 | 0.86 | 0.88 | 0.08 |
| LR oversampling | 85 | 13 | 56214 | 649 | 0.988 | 0.2 | 0.31 | 0.85 | 0.87 | 0.12 |
| NN undersampling | 85 | 17 | 56482 | 360 | 0.993 | 0.3 | 0.38 | 0.82 | 0.83 | 0.19 |
| NN oversampling | 71 | 28 | 56852 | 8 | 0.999 | 0.79 | 0.80 | 0.71 | 0.72 | 0.9 |
| XGBoost | 75 | 23 | 56839 | 24 | 0.999 | 0.76 | 0.76 | 0.75 | 0.76 | 0.76 |
| Random Forest | 68 | 30 | 56862 | 1 | 0.999 | 0.81 | 0.82 | 0.69 | 0.69 | 0.98 |

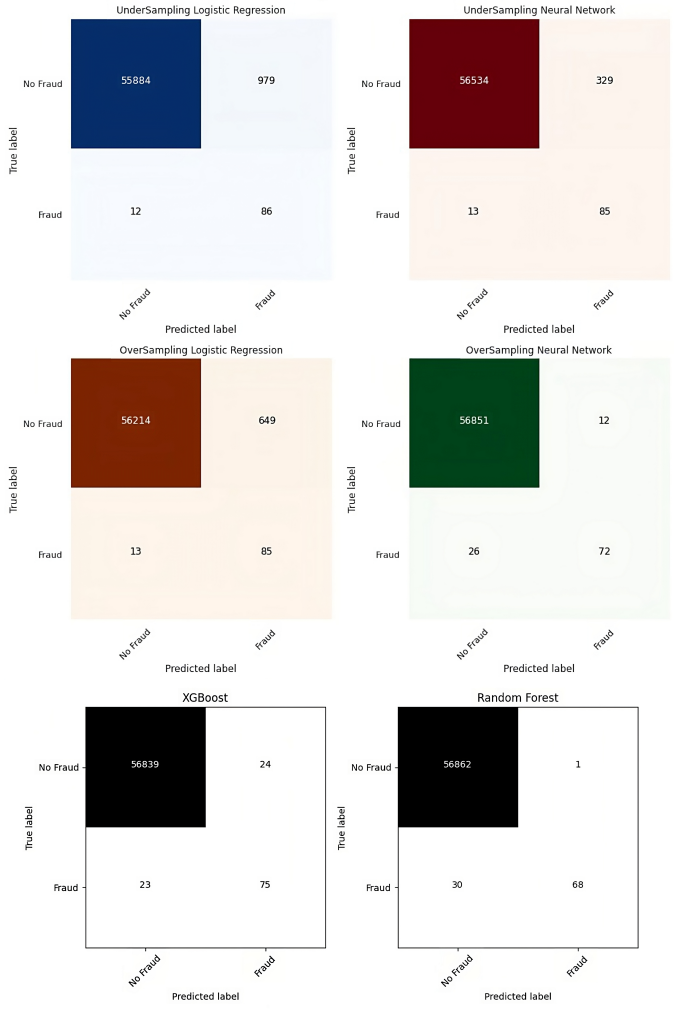


Figure 2. Confusion Matrices

metric to evaluate the performance of the models. Therefore, it is crucial to consider other evaluation metrics, such as the F1 score, MCC, and CBA, in addition to accuracy, to obtain a more comprehensive understanding of the performances.

The logistic regression (LR) model shows varying performance when using different sampling techniques. When using undersampling, the model achieved a low MCC score of 0.26 and a CBA score of 0.86. However, when using oversampling, the model achieved a higher MCC score of 0.31 but a slightly lower CBA score of 0.85. This suggests that oversampling may be more effective in improving the model's performance in this case.

In contrast, the neural network model outperforms the LR model in all cases. When using undersampling, the NN (Neural Network) model achieved an MCC score of 0.38 and a CBA score of 0.82, which is significantly better than the LR undersampling model. Similarly, when using oversampling, the NN model achieved a much higher MCC score of 0.80 and a slightly lower CBA score of 0.71, which suggests that oversampling is also effective for this model. In addition, when comparing the precision and recall scores of the models, it is evident that the neural network models, both with undersampling and oversampling, outperform the logistic regression models. The NN models achieve higher precision scores, indicating a higher accuracy in classifying fraudulent transactions, while also exhibiting significantly improved recall scores, implying a better ability to capture a larger proportion of actual fraudulent transactions. Among the NN models, the oversampling approach yields the highest recall score, indicating its effectiveness in identifying a greater number of true positives.

The XGBoost model shows competitive performance with the NN model, achieving a high MCC score of 0.76 and a CBA score of 0.75. Although the NN oversampling model outperformed the XGBoost model in terms of MCC, the XGBoost model achieved a higher CBA score. Moreover, the XGBoost model shows balanced precision and recall scores, suggesting consistent performance in accurately predicting and capturing fraudulent transactions.

Finally, The Random Forest model achieved 68 True Positive predictions, correctly identifying fraudulent cases, and had 30 False Positive predictions, indicating fraud cases incorrectly classified as non-fraud. The model demonstrated a favorable F1 score of 0.81, indicating a balance between precision and recall. The Matthews Correlation Coefficient value of 0.82 suggests a strong correlation between predicted and actual fraud cases and CBA achieved a value of 0.69, indicating the effectiveness of the model in terms of minimizing costs associated with false positives and false negatives. The precision and recall values for the Random Forest model were 0.69 and 0.98, respectively, highlighting its ability to accurately identify fraudulent instances while minimizing false negatives. Precision and recall represent the ability to correctly identify positive instances and the proportion of correctly identified positive instances, respectively. Random Forest model achieved a relatively high precision value and a very high recall value.

Overall, the results of this study indicate that the random forest model with oversampling is likely the most effective

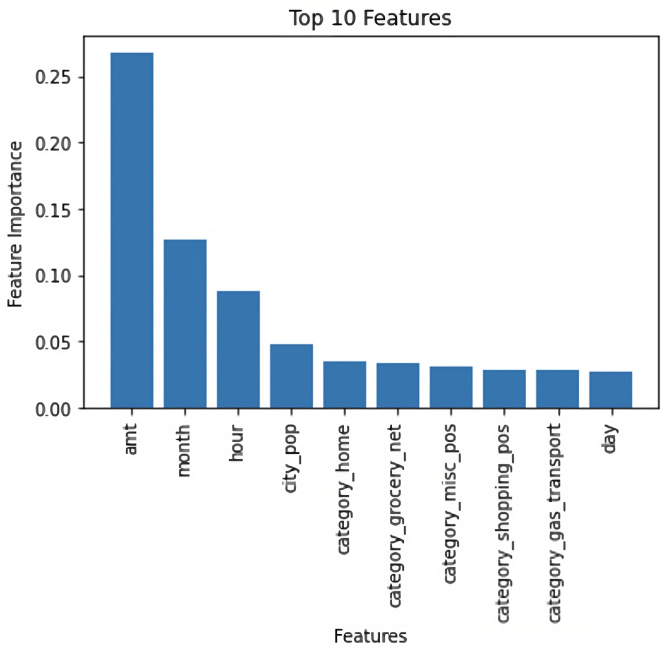


Figure 3. Feature Importances for generated dataset

approach for detecting fraudulent payments, as it achieved the highest MCC score score. However, the XGBoost model also shows promising performance and may be a viable alternative depending on the specific requirements and constraints of the problem. Notably, the XGBoost model does not require any sampling methods and still achieves very good results.

### *D. Explainability*

As previously noted, conducting XAI research on the actual dataset is unfeasible due to confidentiality concerns, as the features have undergone transformation via Principal Component Analysis (PCA). Consequently, the analysis of explainability is carried out on a generated dataset.

A random forest model was developed to identify the most important features in a generated dataset. The dataset was oversampled using SMOTE to mitigate class imbalance. The model achieved a MCC value of 0.8 and CBA value of 0.73. The most ten important features, along with their corresponding values, are displayed in the Figure 3.

The analysis reveals that the most significant feature for the model's decision-making process is the transaction amount, with a feature importance value of 0.27. The second and third most important features are the month and hour of the transaction, with feature importance values of 0.13 and 0.09, respectively. The fourth most significant feature is the city population of the customer, with a feature importance value of 0.05. The fifth most important feature is whether the payment falls under the category of home products, with a feature importance value of 0.035. The next 4 most important features are also the category of payment with a similar importance values around 0.03. Finally, the tenth most important feature is the week day of the payment with the value 0.027. These findings highlight the critical role that transaction amount, transaction time, and customer demographics play in determining the model's decision-making process.

## V. Conclusion

In conclusion, our study demonstrates the potential of AI models for the detection of fraudulent transactions in electronic payment systems. We have worked on two distinct implementations, the first of which involved the use of a real-world dataset to compare the performance of various machine learning algorithms under different types of data imbalance handling techniques. In order to assess and compare the efficacy of different models, a diverse range of metrics, from F1 Score and Matthews Correlation Coefficient are utilized. The second implementation involved the utilization of a generated dataset to gain insights into the decision-making process of a machine learning model. The analysis provided valuable insights into how various features influence the final decision of the model, highlighting their respective levels of importance.

Since the real dataset is publicly available on Kaggle, many models has been implemented for the problem. Most of the people who work with the this data, prefered using SMOTE oversampling and recommended. Our study also proves that SMOTE oversampling has a good impact on the performance of models. On the other hand, XGBoost's capability to have correct predictions without the need of any sampling methods proves that it is a strong model to be used in fraudulent payment detection domain. Most of the authors who share their notebook in Kaggle had a result with lower accuracy that proves the importance of the techniques that is used in this research.

Although the models utilizing undersampling techniques demonstrate much faster performance, the models employing SMOTE oversampling exhibit superior performance. Hence, the models with SMOTE oversampling are considered more favorable due to their better overall performance. Despite the Random Forest model being relatively slow as well, it manages to achieve favorable results in comparison.

For the future work, the present implementation may be further expanded to encompass a wider array of models and employ more sophisticated techniques. While the focus of the current study pertains exclusively to the explainability of the Random Forest model, prospective investigations could extend this domain to encompass diverse machine learning models. By incorporating a diverse range of algorithms, the research can achieve a more comprehensive and in-depth examination of the interpretability landscape in the context of fraud detection. Such endeavors hold the potential to yield valuable insights and advancements in the field of explainable artificial intelligence, thereby contributing to the broader domain of data-driven decision-making and enhancing the overall effectiveness of fraud detection systems.

## References

[1] F. de la Bourdonnaye and F. Daniel. Evaluating resampling methods on a real-life highly imbalanced online credit card payments dataset. *arXiv preprint arXiv:2206.13152*, 2022.

[2] Y. Lucas and J. Jurgovsky. Credit card fraud detection using machine learning: A survey. *arXiv preprint arXiv:2010.06479*, 2020.

[3] S. Shirgave, C. Awati, R. More, and S. Patil. A review on credit card fraud detection using machine learning. *International Journal of Scientific & technology research*, 8(10):1217–1220, 2019.

[4] Pradheepan Raghavan and Neamat Gayar. Fraud detection using machine learning and deep learning. pages 334–339, 12 2019.

[5] Amit Singh, Ranjeet Kumar Ranjan, and Abhishek Tiwari. Credit card fraud detection under extreme imbalanced data: A comparative study of data-level algorithms. *Journal of Experimental & Theoretical Artificial Intelligence*, 34(4):571–598, 2022.

[6] Amalia Luque, Alejandro Carrasco, Alejandro Martín, and Ana de las Heras. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91:216–231, 2019.

[7] S. Boughorbel, Fethi Jarray, and Mohammed El-Anbari. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PLOS ONE*, 12:e0177678, 06 2017.

[8] Machine Learning Group ULB. Kaggle dataset: Credit card fraud detection. *https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud*, Mar 2018.

[9] Brandon Harris. Sparkov data generation: Generate fake credit card transaction data, including fraudulent transactions. *https://github.com/namebrandon/Sparkov_Data_Generation*, 2022.

[10] Melih Kurtaran. Fraud detection datasets: Generated datasets. *https://github.com/melihkurtaran/Fraud_Detection/tree/main/datasets /generated_datasets*, 2023.

[11] Python faker's documentation. *https://faker.readthedocs.io/en/master/*.

[12] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[13] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[14] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[15] Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman. The matthews correlation coefficient (mcc) is more informative than cohen's kappa and brier score in binary classification assessment. *IEEE Access*, 9:78368–78381, 2021.