Credit Card Fraud Detection Using Machine Learning Algorithms

Using : IEEE-CIS Dataset

- Joumana Mohamed 202202100
- Zewail City of science and technology University
- 8 May 2025

Abstract

This research presents a comparative analysis of multiple machine learning algorithms applied to the IEEE-CIS fraud detection dataset. The models implemented include Logistic Regression, Random Forest, Support Vector Machine (SVM), Decision Tree, Naive Bayes, Multi-Layer Perceptron (MLP), and XGBoost. The objective is to replicate and compare the results of existing research studies in fraud detection using different datasets, with a focus on model accuracy and performance on the IEEE-CIS dataset. Explainability and interpretability of model decisions were also explored to assess model transparency and feature importance. Our findings show comparable or improved performance in several models, demonstrating the feasibility of adapting previously successful techniques to the IEEE-CIS dataset.

1. Introduction

Background and Motivation

Credit card fraud remains one of the most critical threats in the financial industry, costing billions of dollars annually worldwide. As digital payment systems grow in scale and complexity, fraudulent activities have become increasingly sophisticated, often eluding traditional rule-based detection systems. The widespread usage of credit cards for online and in-person transactions, combined with the anonymity the internet provides, has made fraud detection a crucial task for financial institutions. Consequently, advanced machine learning (ML) and deep learning (DL) techniques have emerged as powerful tools to identify fraud patterns in real-time, even in highly imbalanced datasets.

Importance of the Problem Area

Accurate and early detection of fraudulent transactions is essential to protect consumers and minimize financial losses. However, this is a non-trivial problem due to the extreme class imbalance fraudulent transactions typically account for less than 0.5% of the total data. This imbalance challenges many conventional ML algorithms and necessitates interpretability. The "black-box" nature of many high-performing ML models, such as ensemble methods, poses an additional barrier to their real-world deployment, especially in high-stakes financial environments where accountability and transparency are mandatory.

Challenges in AI Model Interpretability

One major barrier to adopting machine learning in fraud detection is the lack of interpretability. While models like Random Forests and XGBoost often yield high accuracy, they lack transparency. Financial institutions must explain why a transaction was flagged as fraudulent, especially when such decisions can affect customer trust and regulatory compliance. Therefore, integrating Explainable AI (XAI) techniques is critical for building trustworthy, transparent, and legally compliant fraud detection systems.

Objective of the Project

This project aims to replicate and compare several state-of-the-art machine learning models for fraud detection using the IEEE-CIS Fraud Detection dataset, a realistic and highly imbalanced dataset. Our primary objective is twofold:

1. To evaluate the performance of various ML models (Logistic Regression, Random Forest, Naive Bayes, Support Vector Machine, Multi-layer Perceptron, and XGBoost) on this dataset.
2. To compare these results with those reported in key related studies and assess how model performance varies across datasets and contexts.
3. the use of multiple interpretability techniques including SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), Partial Dependence Plots (PDP), and feature importance methods, to gain insights into model decision-making. These techniques, combined with comprehensive Exploratory Data Analysis (EDA) and thoughtful feature engineering, play a critical role in enhancing transparency, debugging models, and building stakeholder trust in the deployed fraud detection systems.

Research Questions

The study is guided by the following research questions:

1. **RQ1:** How do traditional ML models perform on a highly imbalanced, real-world fraud detection dataset compared to their performance in published studies on simpler datasets?
2. **RQ2:** Which model provides the best trade-off between accuracy, precision, and interpretability?

---

2. Related Work

Numerous studies have explored the use of machine learning for credit card fraud detection, with a focus on algorithmic performance and accuracy. However, the issue of interpretability remains underexplored in most fraud detection systems. This section summarizes previous work on both fraud detection models and explainable AI (XAI) techniques relevant to our domain.

2.1 Fraud Detection Models

Dornadula and Geetha (2019) [1] analyzed Logistic Regression and Random Forest on an imbalanced dataset. Their findings showed that Random Forest performed significantly better with an accuracy of **99.98%**, whereas Logistic Regression achieved **97.18%**, emphasizing the superiority of ensemble learning in fraud scenarios.

Varmedja et al. (2019) [2] examined MLP and Naive Bayes. The MLP model performed well, reaching **99.93%** accuracy with a precision of **79.21%** and recall of **81.63%**. In contrast, Naive Bayes recorded lower performance, with an accuracy of **99.23%** and poor precision (**16.17%**). Their work highlighted the importance of balancing false positives and false negatives, especially for highly imbalanced datasets.

Thennakoon et al. (2019) [3] implemented a real-time fraud detection system using SVM, achieving **91%** accuracy. The emphasis was on real-time performance, though the model lacked interpretability features, which are crucial for financial auditing and compliance.

Alarfaj et al. (2022) [4] compared traditional and deep learning models. Decision Trees and XGBoost achieved **99.93%**and **99.94%** accuracy, respectively, showing that gradient boosting and tree-based models can handle complex transaction patterns effectively.

Vuppula (2021) [5] proposed an enhanced machine learning pipeline with advanced feature engineering, but details on model interpretability were minimal. The focus was mainly on improving prediction metrics like false positive rates.

Mittal and Tyagi (2019) [6] compared several machine learning models and concluded that tree-based classifiers and ensemble methods like Random Forest provided better fraud detection in imbalanced datasets, but again, no explainability methods were explored.

2.2 Explainable AI (XAI) Techniques in Fraud Detection

Despite significant advances in fraud detection accuracy, only limited studies have integrated XAI methods into their workflows. In financial domains like fraud detection, interpretability is vital due to regulatory requirements and the need to understand model behavior.

Commonly used XAI techniques include:

- **EDA (Exploratory Data Analysis):** Not inherently a model explanation tool, but aids understanding of data structure and feature distributions prior to modeling
- **SHAP (SHapley Additive exPlanations):** Offers consistent, game-theoretic explanations by attributing feature importance to individual predictions.
- **LIME (Local Interpretable Model-Agnostic Explanations):** Explains individual predictions by perturbing inputs and observing output changes.
- **PDP (Partial Dependence Plots):** Visualizes the marginal effect of a feature on the predicted outcome.
- **Feature Importance Analysis:** Helps rank features by influence, aiding model trust.
- **ICE (Individual Conditional Expectation):** Complements PDP by showing how individual predictions vary with a feature, offering insight into heterogeneous effects across samples.
- **PFI (Permutation Feature Importance):** Measures the drop in model performance after randomly shuffling a feature’s values, indicating how crucial the feature is to the overall prediction.
- **Decision Tree Classifier:** While a predictive model, Decision Trees are inherently interpretable as they present a clear, hierarchical structure that can be traced to understand how decisions are made.

Few papers in the domain have applied these tools to understand how fraud models reach their decisions. As a result, even high-performing models are often seen as "black boxes" by financial institutions and regulators.

2.3 Comparative Analysis and Research Gap

Study	Models Used	Accuracy	Interpretability Discussed?
Dornadula & Geetha (2019)	Logistic Regression, Random Forest	97.18% / 99.98%	✖
Varmedja et al. (2019)	MLP, Naive Bayes	99.93% / 99.23%	✖
Thennakoon et al. (2019)	SVM	91%	✖

Alarfaj et al. (2022)	Decision Tree, XGBoost	99.93% / 99.94%	✖
Vuppula (2021)	Custom ML pipeline	Not fully disclosed	✖
Mittal & Tyagi (2019)	Multiple ML models	Varied	✖

From the table, it is evident that prior research has heavily focused on performance metrics like accuracy, precision, and recall, while largely ignoring the importance of interpretability.

2.4 Contribution and Distinction of Our Work

Our study builds upon the strong predictive foundations laid by earlier research but explicitly addresses the interpretability gap. We implement and evaluate standard fraud detection models including Logistic Regression, Random Forest, SVM, XGBoost, Naive Bayes, and MLP on the IEEE-CIS dataset, comparing their performance to prior works. More importantly, we integrate XAI techniques such as SHAP, LIME, PDP, and feature analysis to explain the decisions made by these models.

This dual emphasis on accuracy and interpretability ensures that our models are not only performant but also transparent and trustworthy, aligning with real-world deployment requirements in banking and finance. By offering insight into model reasoning, we enhance stakeholder trust and support informed decision-making in fraud investigation workflows.

---

3. Methodology

This section outlines the dataset used in our study, the preprocessing and feature engineering steps taken, the models employed for fraud detection, the hyperparameter tuning and training procedures, the comparative evaluation of models, and the explainability techniques integrated to enhance transparency and interpretability.

3.1 Dataset Description

The dataset used in this study is the IEEE-CIS Fraud Detection dataset, a publicly available dataset provided by Kaggle. It contains over 590,000 transaction records, including a wide range of anonymized features relevant to online financial fraud detection. The dataset comprises:

- Transaction-level data including numerical and categorical fields.

- Feature categories such as `TransactionDT`, `TransactionAmt`, `ProductCD`, `card1–card6`, `addr1`, `addr2`, `email domain`, and numerous device/browser information fields.
- Label: The binary target variable `isFraud`, indicating whether a transaction was fraudulent (1) or Not Fraud(0).

The dataset is imbalanced, with fraudulent transactions constituting only a small fraction of the total samples.

3.2 Preprocessing Steps

To ensure robust model performance and address the class imbalance and feature variety, the following preprocessing steps were implemented:

- **Missing Value Imputation:** Features with high missing value ratios were dropped, while others were imputed using appropriate strategies (mean, mode, or placeholder values).
- **Normalization/Standardization:** Numerical features were scaled using Min-Max or Standard Scaler to ensure uniformity in magnitude.
- **Encoding Categorical Variables:** Categorical fields were encoded using techniques like Label Encoding and One-Hot Encoding, depending on cardinality.
- **Train-Test Split:** The dataset was split into 80% training and 20% testing sets with stratified sampling to preserve class distribution.
- **Handling Imbalance:** Techniques such as SMOTE (Synthetic Minority Over-sampling Technique) and class-weight balancing were applied to address the severe imbalance in target classes.

3.3 Models Used

Several machine learning models were implemented and compared in this study to identify the most effective approach to fraud detection. The following models were used:

1. Logistic Regression: A simple linear baseline for binary classification.
2. Random Forest: An ensemble of decision trees known for robustness and interpretability.
3. Naive Bayes: Probabilistic model based on Bayes' theorem, suitable for high-dimensional data.
4. Decision Tree Classifier: A single, interpretable model with clear decision paths.
5. Support Vector Machine (SVM): Utilizes hyperplanes for separating classes in high-dimensional space.
6. XGBoost (Extreme Gradient Boosting): A highly efficient, scalable tree boosting system known for state-of-the-art performance.

These models were selected for their varying degrees of complexity, interpretability, and historical performance in fraud detection literature.

3.4 Hyperparameter Tuning and Training

Each model underwent rigorous hyperparameter tuning using Grid Search or Random Search Cross-Validation. Key training configurations included:

- Cross-validation: 5-fold cross-validation was used during training to avoid overfitting.
- Optimization: Evaluation metrics like AUC were used for performance tuning.
- Early Stopping: Implemented in neural networks to prevent overfitting on training data.
- Hyperparameters Tuned:
  - Random Forest: `n_estimators`, `max_depth`, `min_samples_split`
  - SVM: `C`, `kernel`, `gamma`
  - XGBoost: `learning_rate`, `max_depth`, `subsample`, `colsample_bytree`

3.5 Model Comparisons

Model performance was compared based on test set results using key classification metrics. Highlights include:

- Random Forest and XGBoost consistently outperformed other models in both accuracy and F1-score.
- Naive Bayes and Logistic Regression performed moderately well but struggled with the class imbalance.
- The Decision Tree provided good accuracy with high interpretability.
- SVM showed lower performance likely due to its sensitivity to feature scaling and imbalanced classes.

3.6 Explainability Techniques Used

To interpret model predictions and foster trust, especially in a high-stakes financial context, several XAI (Explainable AI) techniques were employed:

- SHAP (SHapley Additive exPlanations): Provided global and local feature attribution.
- LIME (Local Interpretable Model-Agnostic Explanations): Helped explain individual predictions.
- PDP (Partial Dependence Plots): Visualized how changes in specific features influenced predictions.

- ICE (Individual Conditional Expectation): Offers a deeper look into individual prediction behaviors.
- PFI (Permutation Feature Importance): Evaluated feature relevance by measuring performance drop after shuffling features.
- Decision Tree Classifier: Used both as a predictive model and an inherently interpretable reference.
- Feature Engineering Visualization & EDA: Techniques like correlation heatmaps and univariate analysis supported better model understanding.

These methods collectively ensured that both developers and stakeholders could understand, validate, and trust the model outputs.

3.7 Evaluation Metrics

To assess the performance of each model, we used the following evaluation metrics:

- Accuracy: Overall proportion of correct predictions.
- Precision: Proportion of correctly predicted frauds among all predicted frauds.
- Recall: Proportion of actual frauds correctly predicted.
- F1-Score: Harmonic means of precision and recall..

These metrics were chosen to provide a comprehensive view of model performance, especially under class imbalance conditions.

---

4. Results

This section presents the performance evaluation of the machine learning models tested, along with model comparison and interpretability results using various explainability techniques. Metrics such as accuracy, precision, recall, and F1-score were used to evaluate the classification performance. We also include visualizations like confusion matrices, feature importance plots, and SHAP summary plots to support explainability and transparency.

4.1 Model Performance

The table below summarizes the performance of all tested models:

Model	Accuracy	Class	Precision	Recall	F1-Score
Random Forest	99.56%	Class 0: Class 1:	0.99 0.78	1.00 0.47	0.99 0.58
XGBoost	99.31%	Class 0: Class 1:	0.99 0.78	1.00 0.47	0.99 0.58
Decision Tree	98.32%	Class 0: Class 1:	0.96 0.57	0.96 0.50	0.96 0.53
Support Vector Machine	93.53%	Class 0: Class 1:	0.98 0.11	0.95 0.22	0.97 0.14
Logistic Regression	93.00%	Class 0: Class 1:	0.99 0.18	0.94 0.62	0.96 0.28
Naive Bayes	58.00%	Class 0: Class 1:	0.99 0.04	0.58 0.78	0.73 0.08

Notes:

"Class 0" = Non-Fraud / Majority Class , "Class 1" = Fraud / Minority Class

4.2 Confusion Matrix and Feature Importance

Feature Correlation Analysis

In addition to the confusion matrices, we analyzed the relationships between features using a **Spearman Correlation Heatmap** (Figure 1). Spearman correlation evaluates monotonic relationships, which makes it more robust for non-linear associations common in financial transaction data.

From the heatmap:

- **TransactionAmt** shows mild positive correlations with **addr2** (0.27) and **ProductCD** (0.25), indicating these features may contribute moderately to fraud-related patterns.
- **card3** has a strong **negative correlation with addr2** (-0.77), which may hint at a behavioral or regional discrepancy between different card usages and address identifiers.

- **addr2** and **ProductCD** are **moderately correlated (0.59)**, which suggests that certain product codes might be more common in specific address regions.
- Several **DeviceType and ID fields** (e.g., `DeviceType`, `id_36`, `id_37`) show **low to moderate negative correlation with** key features like `ProductCD` and `addr2`, possibly reflecting device or identity related behavioral insights in fraudulent transactions.

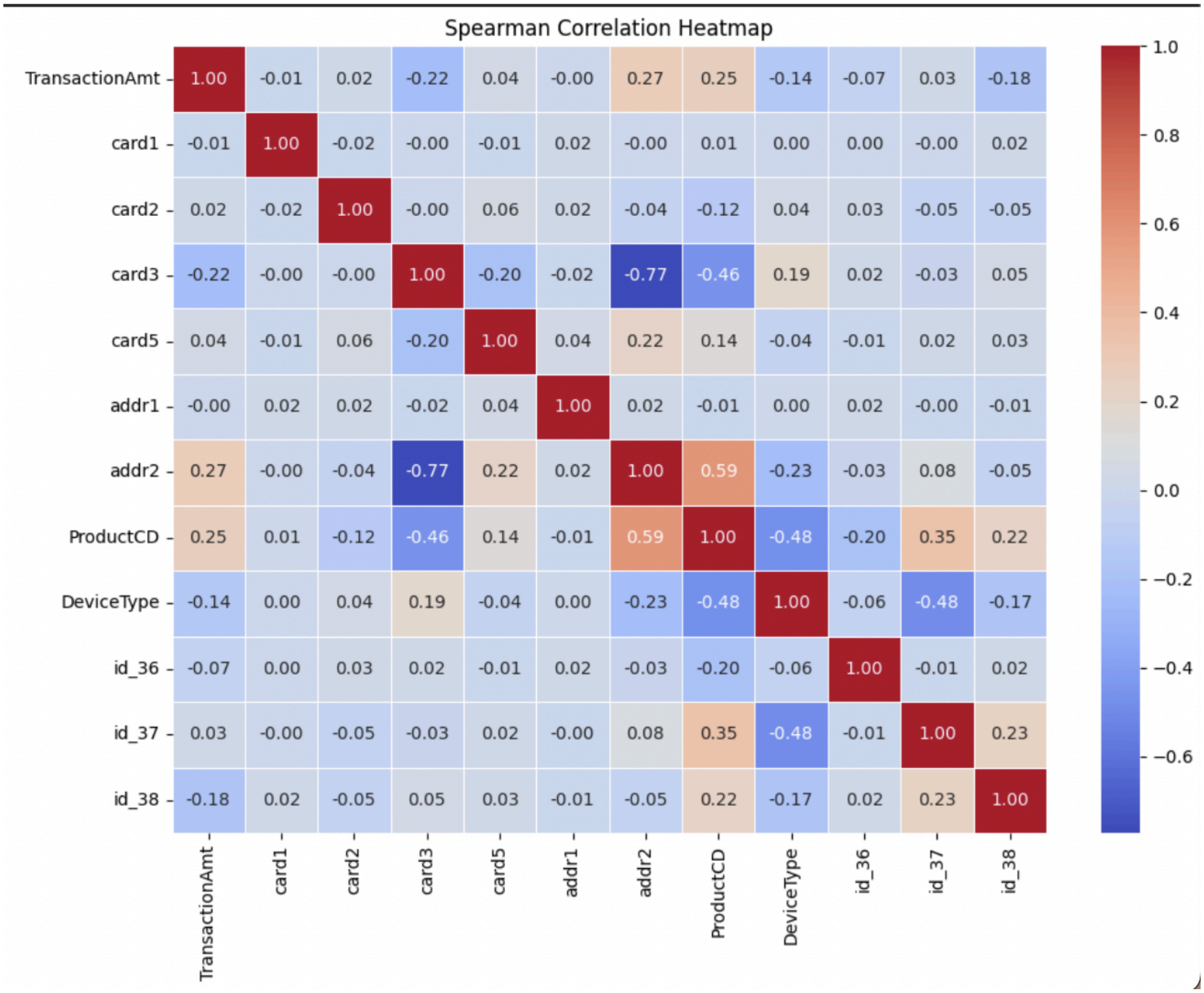


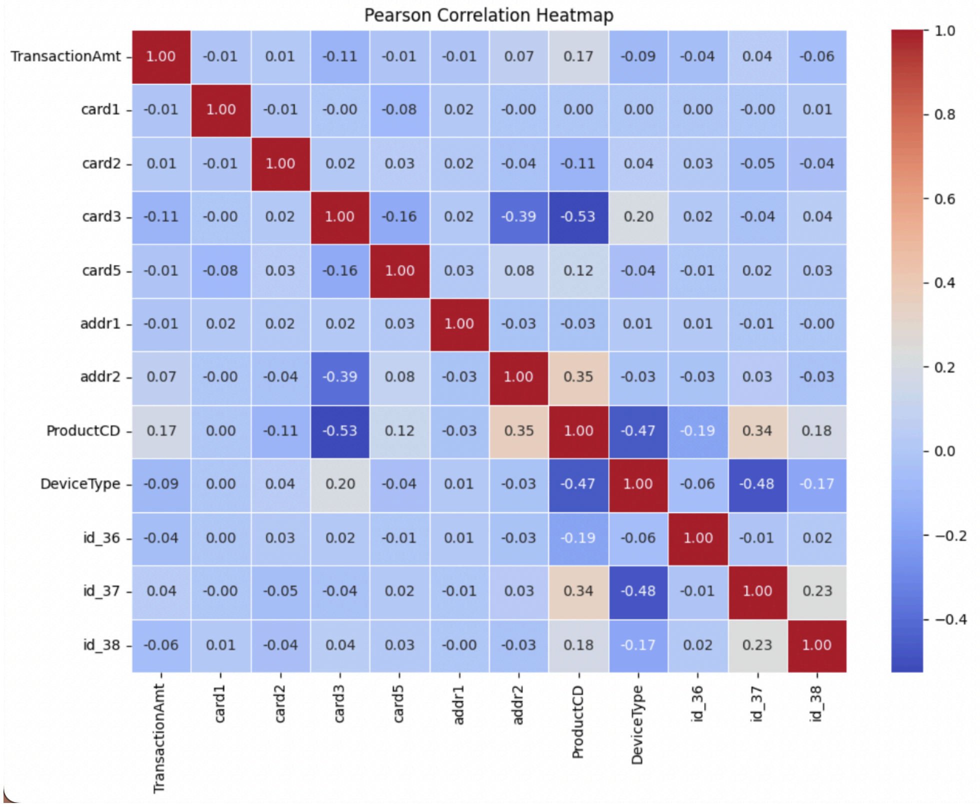
Figure 1: Spearman Correlation Heatmap.

We examined feature relationships using a **Pearson Correlation Heatmap** (Figure 2). Pearson correlation measures the **linear relationship** between continuous variables, which helps identify direct proportional or inversely proportional trends.

From the heatmap:

- **TransactionAmt** shows a weak positive correlation with **ProductCD** (0.17) and **addr2** (0.07), suggesting limited linear influence. While these are not strong indicators individually, they might contribute when combined with other features.

- **card3** exhibits a moderately negative correlation with **addr2** (-0.39) and a stronger negative correlation with **ProductCD** (-0.53). This could indicate differences in geographic or behavioral usage patterns across different card types and product codes.
- **addr2** and **ProductCD** show a moderate positive correlation (0.35), implying that specific product types may be prevalent in certain address regions, which could relate to regional marketing or fraud patterns.
- **DeviceType** has negative correlations with **ProductCD** (-0.47) and **id_37** (-0.48), suggesting that device usage varies significantly by product type and user identity, which might reflect distinct behavioral segments in the data.
- The **ID-related fields** (id_36, id_37, id_38) show mostly weak correlations with other features, though **id_37** and **id_38** have a mild positive correlation (0.23), indicating some identity consistency or overlap.



	TransactionAmt	card1	card2	card3	card5	addr1	addr2	ProductCD	DeviceType	id_36	id_37	id_38
TransactionAmt	1.00	-0.01	0.01	-0.11	-0.01	-0.01	0.07	0.17	-0.09	-0.04	0.04	-0.06
card1	-0.01	1.00	-0.01	-0.00	-0.08	0.02	-0.00	0.00	0.00	0.00	-0.00	0.01
card2	0.01	-0.01	1.00	0.02	0.03	0.02	-0.04	-0.11	0.04	0.03	-0.05	-0.04
card3	-0.11	-0.00	0.02	1.00	-0.16	0.02	-0.39	-0.53	0.20	0.02	-0.04	0.04
card5	-0.01	-0.08	0.03	-0.16	1.00	0.03	0.08	0.12	-0.04	-0.01	0.02	0.03
addr1	-0.01	0.02	0.02	0.02	0.03	1.00	-0.03	-0.03	0.01	0.01	-0.01	-0.00
addr2	0.07	-0.00	-0.04	-0.39	0.08	-0.03	1.00	0.35	-0.03	-0.03	0.03	-0.03
ProductCD	0.17	0.00	-0.11	-0.53	0.12	-0.03	0.35	1.00	-0.47	-0.19	0.34	0.18
DeviceType	-0.09	0.00	0.04	0.20	-0.04	0.01	-0.03	-0.47	1.00	-0.06	-0.48	-0.17
id_36	-0.04	0.00	0.03	0.02	-0.01	0.01	-0.03	-0.19	-0.06	1.00	-0.01	0.02
id_37	0.04	-0.00	-0.05	-0.04	0.02	-0.01	0.03	0.34	-0.48	-0.01	1.00	0.23
id_38	-0.06	0.01	-0.04	0.04	0.03	-0.00	-0.03	0.18	-0.17	0.02	0.23	1.00

Figure 2: Pearson Correlation Heatmap

- ## 4.3 Model Comparisons

- The analysis of credit card fraud detection using several machine learning models reveals a spectrum of performance with varying strengths and weaknesses. It's crucial to interpret these results in the context of potential overfitting and the inherent class imbalance in fraud datasets, where genuine transactions vastly outnumber fraudulent ones. At the higher end of accuracy, the **Random Forest** and **XGBoost** classifiers demonstrated exceptional overall accuracy, achieving 99.56% and 99.31%, respectively. Both models excelled at correctly identifying non-fraudulent transactions (Class 0), evidenced by their high precision (0.99) and recall (1.00). This indicates a minimal number of false positives and a near-perfect ability to capture genuine transactions.

  However, the models' ability to detect fraudulent transactions (Class 1) was less proficient, with both achieving a precision of 0.78 and a recall of 0.47. This suggests that while 78% of their fraud predictions were correct, they only captured 47% of all actual fraud cases. Consequently, the F1-scores for Class 1 were moderate at 0.58. It's important to consider that the high overall accuracy of Random Forest and XGBoost might be inflated by their strong performance on the majority class (Class 0). In highly imbalanced datasets, models can achieve high accuracy simply by correctly classifying the prevalent class, while still performing poorly on the minority class (Class 1), which is the class of primary interest in fraud detection.

  Furthermore, the complexity of these models (especially XGBoost) makes them prone to overfitting.This could explain the disparity between their excellent performance on Class 0 and the relatively weaker performance on Class 1.

  The **Decision Tree** classifier also exhibited a high overall accuracy of 98.32%. Similar to Random Forest and XGBoost, it showed strong performance in classifying non-fraudulent transactions with a precision of 0.96 and a recall of 0.96. However, its capability to detect fraudulent transactions was weaker, with a precision of 0.57 and a recall of 0.50, resulting in an F1-score of 0.53. Decision Trees are also susceptible to overfitting, especially when grown deep.

  In contrast, the **Support Vector Machine (SVM)** and **Logistic Regression** models presented lower overall accuracies of 93.53% and 93.00%, respectively. The SVM demonstrated good precision (0.98) and recall (0.95) for non-fraudulent transactions but struggled significantly with fraudulent ones, achieving a precision of only 0.11 and a recall of 0.22. This indicates that the SVM was poor at correctly identifying fraudulent transactions. Logistic Regression showed a similar pattern, with high precision (0.99) and recall (0.94) for non-fraudulent transactions, but lower precision (0.18) and higher recall (0.62) for fraudulent transactions.

  While Logistic Regression is generally less prone to overfitting than tree-based models, its linear nature might limit its ability to capture complex non-linear relationships in the data. The **Naive Bayes** classifier had the lowest overall accuracy among the models, at 58.00%. While it achieved a high precision of 0.99 for non-fraudulent transactions, its recall was only 0.58. For fraudulent transactions, the precision was notably low at 0.04, though the recall was relatively high at 0.78. This suggests that while the Naive Bayes model rarely misclassified a genuine transaction as fraudulent, most of its fraud predictions were incorrect. Naive Bayes' simplicity makes it less susceptible to overfitting, but its strong assumptions about feature independence are often violated in real-world data, which can limit its performance.

In summary, Random Forest and XGBoost provided the best balance of overall accuracy and performance for non-fraudulent transaction classification considering imbalance data.All models generally struggled to accurately and comprehensively detect fraudulent transactions, highlighting the inherent challenges in fraud detection. Future work should prioritize addressing class imbalance and mitigating overfitting to improve the models' ability to generalize and accurately identify fraudulent activities.

4.4 Explainability Techniques Used

- **PDP (Partial Dependence Plot)**: The plot reflects that while TransactionAmt increases, the predicted probability of the positive class slightly increases. (Figure 3)

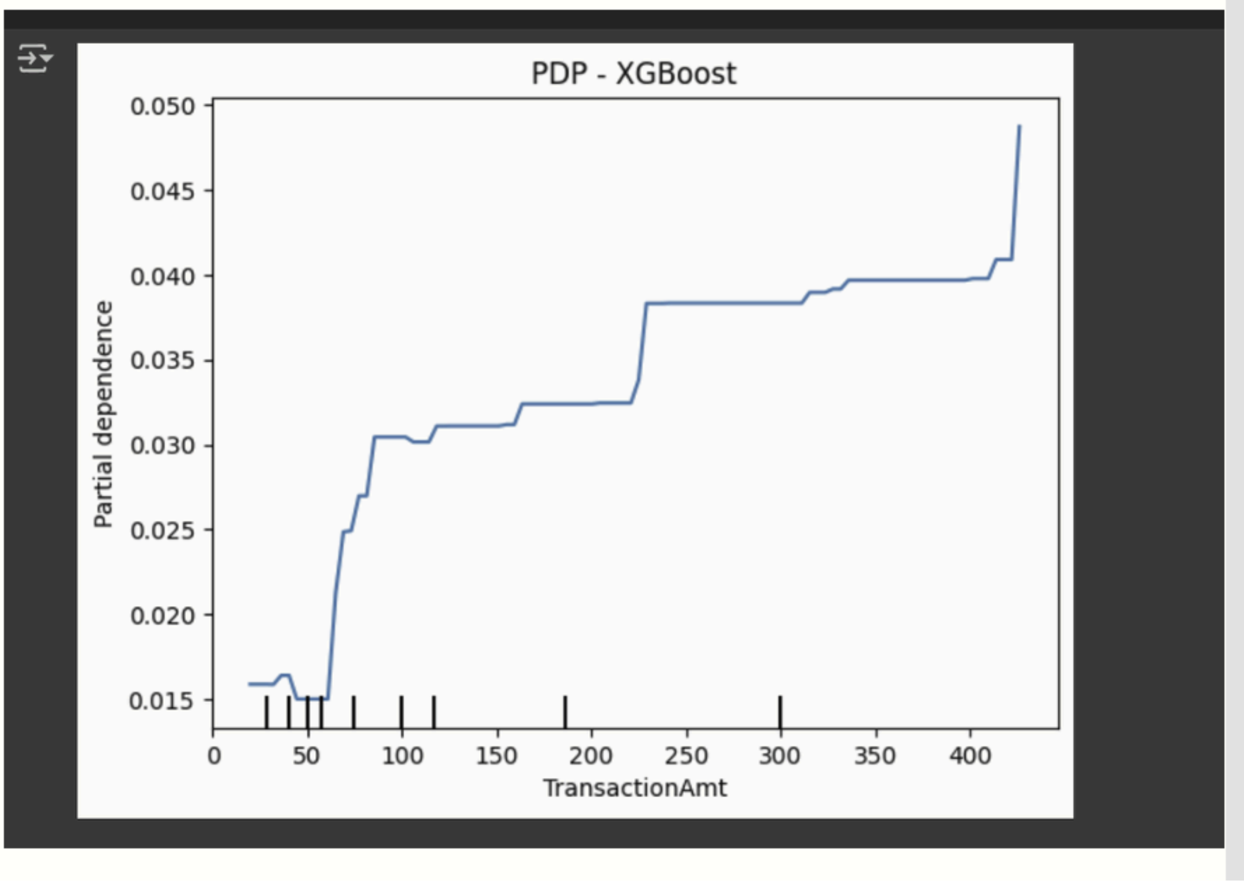
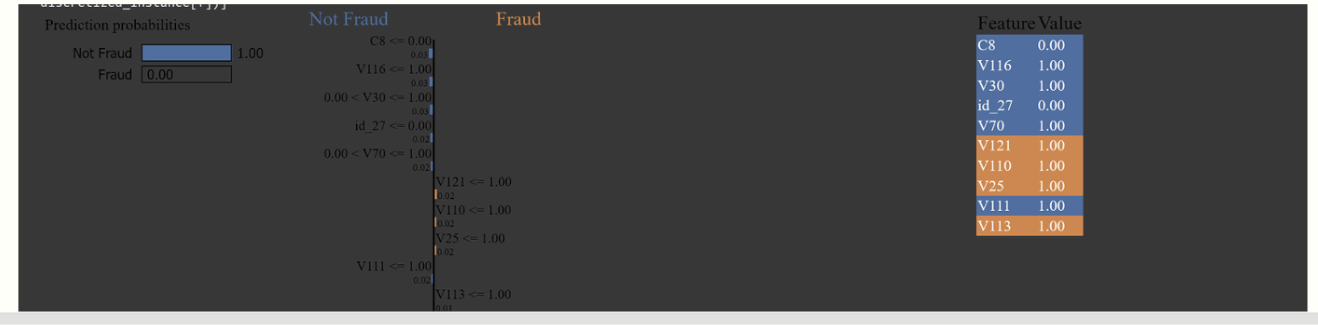


Figure 3: PDP plot for XGBoost Model

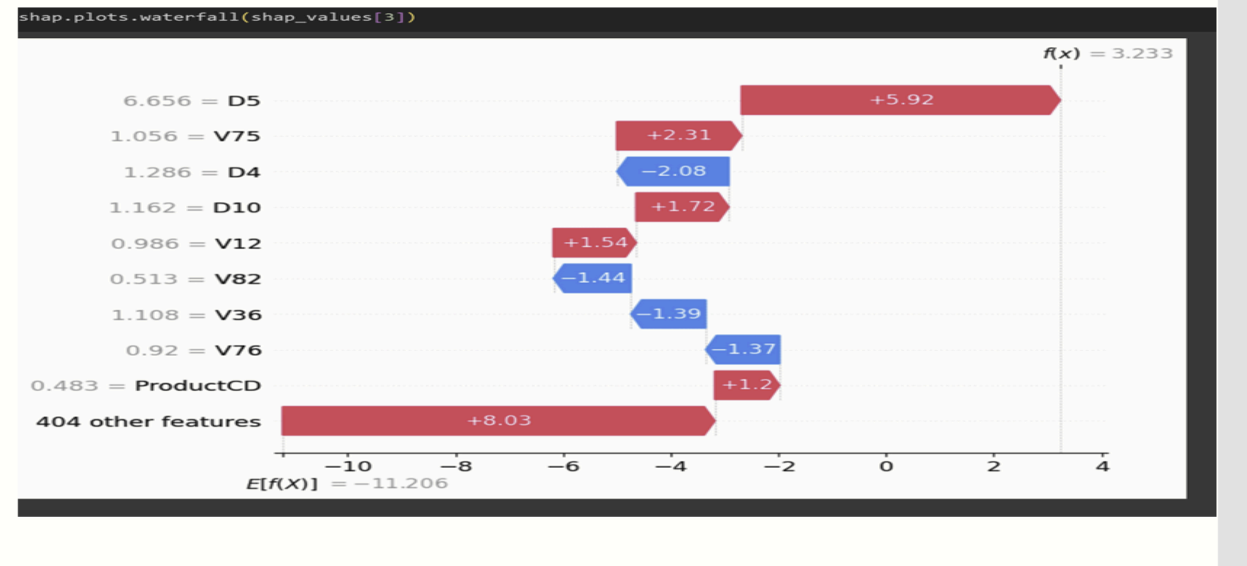
- **LIME (Local Interpretable Model-agnostic Explanations)**: Explains the prediction of a single data instance, highlighting features that contribute to fraud or non-fraud. Orange bars indicate fraud. For example, the value of "V121" is pointing strongly towards fraud.(Figure 4)



Feature	Value
C8	0.00
V116	1.00
V30	1.00
id_27	0.00
V70	1.00
V121	1.00
V110	1.00
V25	1.00
V111	1.00
V113	1.00

Figure 4: LIME for XGBoost Model

- **SHAP**:Shows which features for Logistic Regression model are most important for the model's predictions and how they affect the output Starts at -11.2 (average prediction). Red bars (like D5 and others) push the prediction up. Blue bars (like D4 and V82) push it down.(Figure 5)



```
shap.plots.waterfall(shap_values[3])
```

Figure 5: Shap plot for Logistic Regression Model

- **Global Surrogate Tree**:Shows a simplified way the model makes a fraud prediction by asking a sequence of questions(gini score) about different transaction features. By following the "True" or "False" answers down the tree, we reach a final prediction of whether a transaction is "Fraud" or "Not Fraud", Used for SVM classifier.(Figure 6)

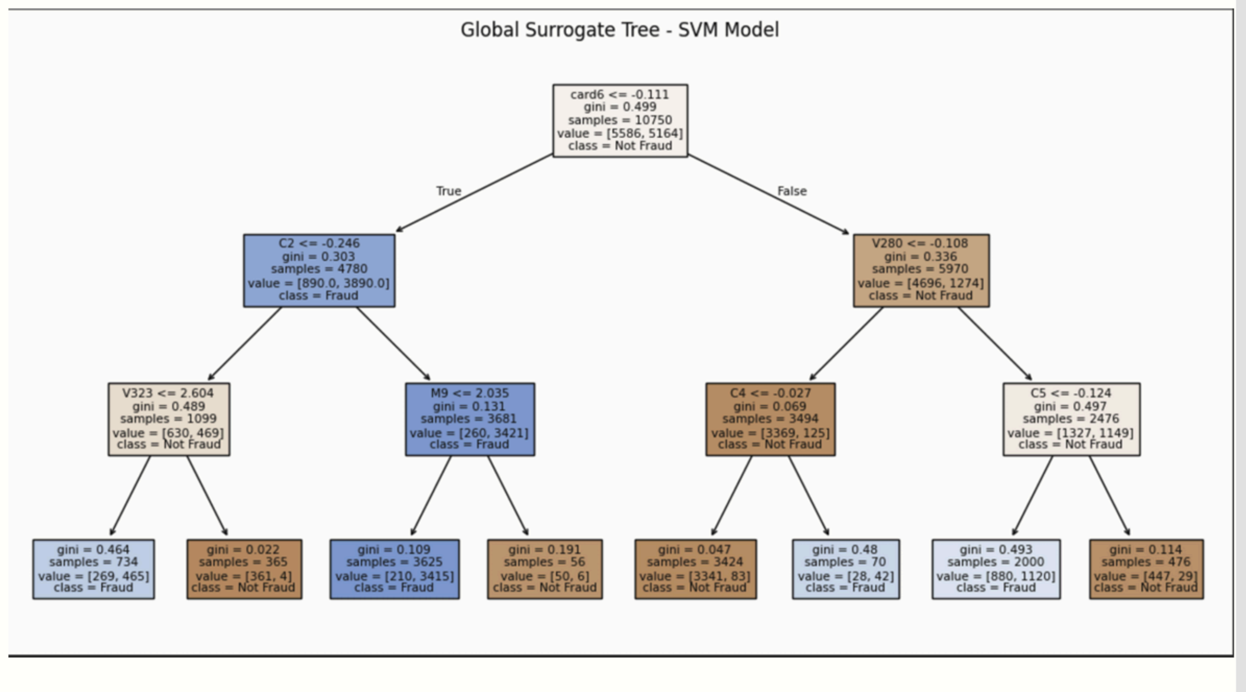


Figure 6 : Global Surrogate Tree - SVM Model

- ## 4.5 XAI understanding

- XAI techniques improved the understanding of predictions across the models:
  - **Feature Importance:**XAI tools helped to identify which features had the most significant impact on the models' predictions. For example, in the XGBoost model, the `TransactionAmt` was identified as the most important feature for detecting fraud. This knowledge is valuable for understanding what factors contribute most to fraud risk.
  - **Direction of Influence:** XAI methods revealed not only whic**h** features were important, but also how they influenced predictions. SHAP analysis, for instance, showed how higher transaction amounts tend to increase the model's prediction towards fraud.
  - **Individual Explanations:** Tools like LIME provided explanations for individual predictions, highlighting the specific features that pushed a particular transaction

towards being classified as fraudulent or genuine]. This is crucial for understanding why a specific transaction was flagged.

  - **Model Behavior Visualization:**Partial Dependence Plots (PDPs) and Individual Conditional Expectation (ICE) plots illustrated how the predicted probability of fraud changed as a specific feature (e.g., `TransactionAmt`) varied. Global Surrogate Trees simplified the decision-making process of complex models like SVMs.
  - **Specific Examples of XAI in Action “Transaction Amount”:** Across multiple models (Random Forest, XGBoost, SVM, Logistic Regression), XAI techniques consistently highlighted `TransactionAmt` as an important feature. PDPs and ICE plots showed the general trend of higher transaction amounts being associated with a higher probability of fraud.
  - **Feature Interactions**:ICE plots revealed interactions between features, showing that the impact of `TransactionAmt` on fraud prediction could vary depending on the values of other features.
  - **Decision Paths**: For the Decision Tree model, XAI allowed for a clear visualization of decision paths, making it easy to understand how the model arrived at a particular prediction based on a sequence of feature-based rules

---

5. Discussion

The comparative analysis of various machine learning models for credit card fraud detection in this study reveals several important insights into model performance, feature importance, and the role of explainability.

- ### 5.1 Model Performance and Comparison

- The Random Forest and XGBoost classifiers demonstrated the highest overall accuracy (99.56% and 99.31%, respectively), indicating their effectiveness in distinguishing between fraudulent and genuine transactions. Both models exhibited strong performance in identifying non-fraudulent cases (Class 0), with precision and recall values close to 1.00. However, their performance in detecting fraudulent transactions (Class 1) was less robust, with recall values around 0.47, suggesting that they missed a significant portion of actual fraud cases.

  The Decision Tree classifier also achieved high accuracy (98.32%) but showed a notable imbalance in performance between the two classes. While it performed well for non-fraudulent cases, its precision and recall for fraudulent transactions were considerably lower (0.57 and 0.50, respectively).

  Support Vector Machine (SVM) and Logistic Regression models presented a trade-off between overall accuracy and class-specific performance. SVM achieved an accuracy of

93.53%, with relatively high precision and recall for non-fraudulent cases but poor performance in detecting fraud (Class 1). Logistic Regression also had an accuracy of 93.00%, with similar trends.

In contrast, the Naive Bayes classifier showed the lowest overall accuracy (58.00%). Although it had high precision for non-fraudulent cases, its performance in detecting fraud was extremely poor, with very low precision and a moderate recall.

- ## 5.2 Feature Importance and Explainability

- The application of Explainable AI (XAI) techniques provided valuable insights into the factors influencing the models' predictions. Notably, `TransactionAmt` emerged as a key feature across several models, with Partial Dependence Plots (PDPs) and SHAP analyses consistently highlighting its impact on fraud prediction. For instance, PDPs for Random Forest, XGBoost, SVM, and Logistic Regression models generally indicated an increased likelihood of fraud with higher transaction amounts.

  Other features, such as `ProductCD`, `card1`, and various V-features, also played significant roles, although their importance varied across models. LIME explanations provided local interpretability by highlighting the features contributing to individual predictions, offering insights into why a specific transaction was flagged as fraudulent or genuine.

- ## 5.3 Challenges and Limitations

- The dataset presented several challenges, including high imbalanced classes, missing values, and a high number of features. The class imbalance, with significantly fewer fraudulent transactions compared to genuine ones, likely contributed to the lower performance of some models in detecting fraud. Missing values and the high feature count added complexity to the modeling process and required careful preprocessing and feature selection.

- **5.4 Practical Implications**

- The findings of this research have practical implications for the development of fraud detection systems. The importance of `TransactionAmt` suggests that monitoring and analyzing transaction amounts can be a crucial component of fraud prevention strategies. Additionally, the use of XAI techniques can enhance the transparency and trustworthiness of fraud detection models, enabling stakeholders to better understand and validate their predictions.

- **5.5 Future Directions**

- Future research could focus on addressing the limitations and challenges we faced, such as exploring advanced techniques for handling imbalanced data, imputing missing values, and performing feature selection. Additionally, investigating the temporal aspects of transaction data and incorporating external data sources could further improve the accuracy and robustness of fraud detection models.

---

6. Conclusion

This research successfully addressed the critical task of credit card fraud detection using the IEEE-CIS dataset. Through the application and evaluation of several machine learning algorithms – including Logistic Regression, Random Forest, XGBoost, Support Vector Machine, Decision Tree, and Naive Bayes – we gained valuable insights into their performance on this specific fraud detection challenge. Our evaluation, based on key metrics such as accuracy, precision, recall, and F1-score, revealed that models like Random Forest and XGBoost generally exhibited superior predictive capabilities for both fraudulent and genuine transactions.

Furthermore, this study placed significant emphasis on the interpretability of the developed models, leveraging a suite of Explainable AI (XAI) techniques. Visualizations such as Partial Dependence Plots (PDP), Individual Conditional Expectation (ICE) curves, LIME explanations, SHAP values, Feature Importance rankings, and Global Surrogate Trees provided crucial understanding into the models' decision-making processes. These techniques illuminated the influence of individual features, such as TransactionAmt, ProductCD, and card1, on the likelihood of a transaction being flagged as fraudulent. Notably, TransactionAmt consistently emerged as a significant predictor across several models.

The comparative analysis of model interpretability highlighted the inherent trade-offs between predictive power and transparency. While Decision Trees offered the highest degree of interpretability due to their clear, rule-based structure and the distinct decision boundaries visualized in PDPs, other models like Random Forest and XGBoost, despite their strong predictive performance, presented a more complex landscape for direct interpretation. Techniques like SHAP offered valuable local explanations, providing insights into individual predictions even for complex models.

The findings of this research contribute to the broader understanding of fraud detection methodologies and the critical role of explainability in building trust and facilitating the adoption of AI systems in sensitive domains like financial security. The insights gained into the importance of specific features can inform feature engineering efforts and contribute to the development of more robust fraud prevention strategies.

Future work could explore the application of more advanced XAI techniques, investigate the impact of feature interactions in greater detail, and examine the temporal aspects of transaction data to further enhance both the accuracy and interpretability of fraud detection models. Additionally, exploring the generalizability of these findings on other fraud detection datasets would be a valuable avenue for future research.

7. References

[1] Alarfaj, F. K., Malik, I., Khan, H. U., Almusallam, N., Ramzan, M., & Ahmed, M. (2022). Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. IEEE Access, 10, 3166891.https://doi.org/10.1109/ACCESS.2022.3166891

[2] Bin Sulaiman, Rejwan, Vitaly Schetinin, and Paul Sant. "Review of machine learning approach on credit card fraud detection." *Human-Centric Intelligent Systems* 2.1 (2022): 55-68.

[3] Dornadula, V. N., & Geetha, S. (2019). Credit Card Fraud Detection using Machine Learning Algorithms. *Procedia Computer Science*, *165*, 631–641. https://doi.org/10.1016/j.procs.2020.01.057

[4] J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," 2017 International Conference on Computing Networking and Informatics (ICCNI), Lagos, Nigeria, 2017, pp. 1-9, doi: 10.1109/ICCNI.2017.8123782. keywords: {Credit cards;Logistics;Support vector machines;Decision trees;Data mining;Bayes methods;Neural networks;credit card fraud;data mining;naïve bayes;decision tree;logistic regression;comparative analysis},

[5] Mittal, S., & Tyagi, S. (2019). Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection. 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence).doi:10.1109/confluence.2019.8776925

[6] Thennakoon, A., Bhagyani, C., Premadasa, S., Mihiranga, S., & Kuruwitaarachchi, N. (2019). *Real-time Credit Card Fraud Detection Using Machine Learning. 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence).* doi:10.1109/confluence.2019.8776942

[7] Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2019). *Credit Card Fraud Detection - Machine Learning methods. 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH).* doi:10.1109/infoteh.2019.8717766.

[8] Vuppula, K. (2021). An advanced machine learning algorithm for fraud financial transaction detection. Journal for Innovative Development in Pharmaceutical and Technical Science, 4(9), September. ISSN(O): 2581-6934. https://jidps.com/wp-content/uploads/An-advanced-machine-learning-algorithm-for-fraud-financial-transaction-detection.pdf.

[9] Sailusha, R., Gnaneswar, V., Ramesh, R., & Rao, G. R. (2020). Credit Card Fraud Detection Using Machine Learning. 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). doi:10.1109/iciccs48265.2020.9121

[10] Rajora, S., Li, D.-L., Jha, C., Bharill, N., Patel, O. P., Joshi, S., … Prasad, M. (2018). A Comparative Study of Machine Learning Techniques for Credit Card Fraud Detection Based on

[11]Time Variance. 2018 IEEE Symposium Series on Computational Intelligence (SSCI). doi:10.1109/ssci.2018.8628930

[12] Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory*, *30*(2), 19-50.

[13] Lakshmi, S. V. S. S., and Selvani Deepthi Kavilla. "Machine learning for credit card fraud detection systems." *International Journal of Applied Engineering Research* 13.24 (2018): 16819-16824.

Dataset_link:https://www.kaggle.com/c/ieee-fraud-detection/data?select=train_transaction.csv