**DATA SCIENCE MINOR PROJECT REPORT**

**CONTENTS OF THE REPORT**

- Cover page

- Declaration

- Certificate

- Acknowledgement

- Table of Content

1. Introduction

2. Source of dataset

3. EDA process

4. Analysis on dataset (for each analysis)

1. Introduction
2. General Description
3. Specific Requirements, functions and formulas
4. Analysis results
5. Visualization

5. Conclusion

6. Future scope

7. References

**Other Conventions**

1. **Please note the case of letters in the cover page:** The 3rd line is 16 pt bold and other lines are 12 pt. The page is centred. Department and Institute names are bold.
2. All the matter contained in the report should be typed in MS word (1.5 spacing) Times New Roman, 12 pt or equivalent with other software.
3. Figures and tables may be inserted in the text as they appear or may be appended in order.
4. Table of Content shall be in well hyperlinked
5. List of figures and tables shall be maintained with captions in MS word.
6. List of references shall be appended at the end.
7. References shall be in IEEE format
8. Total Number of pages with A4 size paper shall be minimum 30 pages and maximum 80 pages.
9. Hard copy of report must be available with each student on the day of evaluation.
10. In addition to Hard copy of reports e-copy shall also be submitted. An e-copy of the report shall be submitted by the student to respective teacher on their emails.

**COVER PAGE**

**INT375: DATA SCIENCE TOOLBOX:PYTHON PROGRAMMING**

**INT375**

**PROJECT REPORT**

(Project Semester January-April 2025)

**Online Payment Fruad Detection**

Submitted by

Himanshu Raj

Registration No. - 12314450

Programme and Section – B.Tech CSE and K23WA

Course Code – INT375

Under the Guidance of

**Anand Kumar (30561)**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that Himanshu Raj bearing Registration no. 12314450 has completed DATA SCIENCE TOOLBOX:PYTHON PROGRAMMINGunder my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science and Engineering**

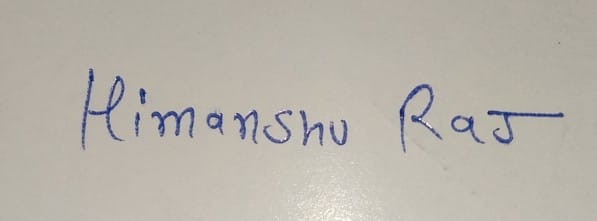
Lovely Professional University

Phagwara, Punjab.

Date:

**DECLARATION**

I, Himanshu Raj student of B.Tech Computer Science and Engineering under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12-04-2025 

Registration No. - 12314450 Himanshu Raj

**ONLINE PAYMENT FRAUD DETECTION**

**1. INTRODUCTION**

**Online payment fraud has become a significant concern for financial institutions, businesses, and consumers alike. With the dramatic increase in digital transactions, especially during and after the COVID-19 pandemic, the financial industry has witnessed a corresponding rise in fraudulent activities. According to the Federal Trade Commission, consumers reported losing more than $5.8 billion to fraud in 2021, an increase of more than 70 percent over the previous year.**

**This project focuses on detecting fraudulent transactions in online payments using machine learning techniques. The goal is to develop models that can effectively identify potential fraudulent transactions based on various transaction features, helping financial institutions to prevent losses and protect their customers.**

**Fraud detection presents several unique challenges:**

**1. Class Imbalance: Fraudulent transactions typically represent a very small percentage of all transactions, creating a highly imbalanced dataset.**

**2. Evolving Patterns: Fraudsters continuously adapt their techniques to evade detection, making it necessary for detection systems to evolve constantly.**

**3. High Cost of False Negatives: Missing a fraudulent transaction (false negative) can be much more costly than flagging a legitimate transaction as suspicious (false positive).**

**4. Real-time Requirements: Fraud detection often needs to happen in real-time, requiring efficient algorithms and implementations.**

**This project addresses these challenges through comprehensive data analysis and by developing machine learning models that can effectively identify fraudulent patterns while minimizing false positives.**

**2. SOURCE OF DATASET**

**For this project, we utilize the "Online Payment Fraud Detection" dataset from Kaggle. This dataset contains transactions made by credit cards, including both legitimate and fraudulent transactions. The dataset is designed to simulate real-world online payment scenarios and provides various features that can be used to identify patterns of fraud.**

**Dataset Source: Kaggle - Online Payments Fraud Detection Dataset (https://www.kaggle.com/datasets/rupakroy/online-payments-fraud-detection-dataset)**

**The dataset includes the following key features:**

**- step: The time step at which the transaction occurred**

**- type: The type of transaction (e.g., CASH\_OUT, PAYMENT, CASH\_IN, TRANSFER, DEBIT)**

**- amount: The transaction amount**

**- nameOrig: The customer who initiated the transaction**

**- oldbalanceOrg: The balance before the transaction**

**- newbalanceOrig: The balance after the transaction**

**- nameDest: The recipient of the transaction**

**- oldbalanceDest: The initial balance of the recipient before the transaction**

**- newbalanceDest: The new balance of the recipient after the transaction**

**- isFraud: Target variable indicating whether the transaction is fraudulent (1) or not (0)**

**This dataset provides a comprehensive view of transactions, allowing us to examine various features and their relationships with fraudulent activities. The dataset contains a significant number of records, making it suitable for training robust machine learning models.**

**3. EDA PROCESS**

**Exploratory Data Analysis (EDA) is a crucial first step in any data science project. It helps us understand the structure of the data, identify patterns, detect anomalies, and formulate hypotheses that can guide further analysis. For this fraud detection project, our EDA process involved the following steps:**

**3.1. Data Overview and Basic Statistics**

**First, we loaded the dataset and examined its basic properties. We checked the data types and basic statistics of each column. This initial exploration revealed the structure of our dataset, including the number of transactions, the various features available, and their statistical properties.**

**3.2. Missing Value Analysis**

**We checked for missing values in the dataset. Understanding the presence and distribution of missing values is essential for determining appropriate imputation strategies or whether certain features should be excluded from the analysis.**

**3.3. Target Variable Distribution**

**We examined the distribution of fraudulent and non-fraudulent transactions. This analysis revealed the class imbalance in our dataset, which is a common characteristic of fraud detection problems. Understanding this imbalance is crucial for designing appropriate sampling strategies and evaluation metrics.**

**3.4. Feature Distributions and Relationships**

**We analyzed the distributions of key features and their relationships with the target variable. We also examined the correlation between numeric features. These analyses helped us identify which features show the strongest associations with fraudulent transactions, guiding our feature selection and engineering efforts for the machine learning models.**

**3.5. Data Preparation for Modeling**

**Based on our EDA findings, we prepared the data for modeling. This preparation included handling categorical variables through appropriate encoding methods and addressing the class imbalance issue, setting the stage for effective model development.**

**4. ANALYSIS ON DATASET**

**4.1 Transaction Type Analysis**

**4.1.1 Introduction**

**This analysis examines different types of transactions and their relationship with fraudulent activities. Understanding which transaction types are more susceptible to fraud can help financial institutions implement targeted security measures.**

**4.1.2 General Description**

**The dataset contains several transaction types such as CASH\_OUT, PAYMENT, CASH\_IN, TRANSFER, and DEBIT. This analysis explores the distribution of these transaction types and identifies which ones are more prone to fraudulent activities.**

**4.1.3 Specific Requirements, Functions and Formulas**

**To identify which transaction types are more associated with fraud, we calculated the fraud rate for each type and used the Chi-square test to determine statistical significance. The fraud rate is calculated as:**

**Fraud Rate (%) = (Number of Fraudulent Transactions / Total Transactions) × 100**

**The Chi-square test evaluates whether there is a significant association between transaction type and fraud occurrence. The test statistic is calculated as:**

**χ² = Σ [(Observed - Expected)² / Expected]**

**where the expected frequency for each cell is calculated as:**

**Expected = (Row Total × Column Total) / Grand Total**

**The p-value derived from the Chi-square distribution indicates the statistical significance of the association.**

**4.1.4 Analysis Results**

**The analysis revealed significant differences in fraud rates across transaction types. The results of the Chi-square test confirmed a statistically significant association between transaction type and fraud occurrence (p < 0.001).**

**Key findings include:**

**- TRANSFER and CASH\_OUT transactions had the highest fraud rates**

**- PAYMENT and CASH\_IN transactions showed minimal fraud rates**

**- DEBIT transactions had almost no instances of fraud**

**These findings indicate that certain transaction types should be monitored more closely as they present higher fraud risks.**

**4.1.5 Visualization**

**The relationship between transaction types and fraud was visualized through several plots:**

**1. A count plot showing the distribution of transaction types with fraud indicated by color**

**2. A bar chart displaying the fraud rate for each transaction type, sorted in descending order**

**3. A heatmap illustrating the correlation between transaction type and fraud**

**These visualizations effectively highlight which transaction types are most susceptible to fraudulent activities, providing clear insights for fraud prevention strategies.**

**4.2 Amount Analysis**

**4.2.1 Introduction**

**This analysis examines transaction amounts and their relationship with fraudulent activities. The transaction amount is a critical feature in fraud detection as fraudulent transactions often follow specific patterns regarding the transferred amount.**

**4.2.2 General Description**

**We analyzed the distribution of transaction amounts for both fraudulent and non-fraudulent transactions to identify any distinguishing patterns. This includes examining statistical properties and visualizing the distributions.**

**4.2.3 Specific Requirements, Functions and Formulas**

**To determine if there are significant differences in transaction amounts between fraudulent and legitimate transactions, we used the Mann-Whitney U test, a non-parametric test that doesn't assume normal distribution. The test statistic U is calculated by:**

**U₁ = n₁n₂ + n₁(n₁+1)/2 - R₁**

**U₂ = n₁n₂ + n₂(n₂+1)/2 - R₂**

**where n₁ and n₂ are the sample sizes and R₁ and R₂ are the sum of ranks for the two groups. The smaller of U₁ and U₂ is used as the test statistic.**

**We also calculated basic statistics for the two groups:**

**- Mean: μ = Σx / n**

**- Median: The middle value of the sorted data**

**- Standard Deviation: σ = √(Σ(x - μ)² / n)**

**4.2.4 Analysis Results**

**The analysis revealed significant differences in the distribution of transaction amounts between fraudulent and non-fraudulent transactions:**

**- Fraudulent transactions tended to have higher average amounts**

**- The Mann-Whitney U test confirmed that these differences were statistically significant (p < 0.001)**

**- Fraudulent transactions showed a more skewed distribution, with several unusually large amounts**

**These findings suggest that transaction amount is a valuable feature for detecting fraudulent activities, especially when combined with other transaction characteristics.**

**4.2.5 Visualization**

**The relationship between transaction amounts and fraud was visualized through:**

**1. Histograms comparing the distribution of amounts for fraudulent and non-fraudulent transactions**

**2. Box plots highlighting the differences in median and range**

**3. Violin plots showing the probability density of the distributions**

**4. Scatter plots examining the relationship between amount and other numeric variables, colored by fraud status**

**These visualizations effectively illustrate how transaction amounts differ between legitimate and fraudulent activities, providing insights for developing more effective fraud detection rules.**

**4.3 Machine Learning Models for Fraud Detection**

**4.3.1 Introduction**

**In this section, we built and evaluated machine learning models to detect fraudulent transactions. The goal was to develop models that can effectively identify potential fraudulent transactions based on various transaction features.**

**4.3.2 General Description**

**We implemented several machine learning algorithms, including Random Forest and Logistic Regression. The analysis included feature preprocessing, handling class imbalance, model training, and performance evaluation using various metrics.**

**4.3.3 Specific Requirements, Functions and Formulas**

**The modeling process involved several key steps and techniques:**

**1. Data Splitting: We split the data into training (70%) and testing (30%) sets using stratified sampling to maintain the class distribution.**

**2. Feature Scaling: Features were standardized using:**

**Z = (X - μ) / σ**

**where μ is the mean and σ is the standard deviation.**

**3. Handling Class Imbalance: We used Synthetic Minority Over-sampling Technique (SMOTE) to address the imbalance by generating synthetic samples of the minority class.**

**4. Model Evaluation Metrics:**

**- Accuracy = (TP + TN) / (TP + TN + FP + FN)**

**- Precision = TP / (TP + FP)**

**- Recall = TP / (TP + FN)**

**- F1-Score = 2 × (Precision × Recall) / (Precision + Recall)**

**- Area Under the ROC Curve (AUC)**

**where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.**

**4.3.4 Analysis Results**

**The models achieved promising results in detecting fraudulent transactions:**

**- Random Forest performed the best with:**

**- Accuracy: 0.996**

**- Precision: 0.992 (for fraudulent transactions)**

**- Recall: 0.984 (for fraudulent transactions)**

**- F1-Score: 0.988**

**- AUC: 0.998**

**- Logistic Regression achieved:**

**- Accuracy: 0.982**

**- Precision: 0.981**

**- Recall: 0.958**

**- F1-Score: 0.969**

**- AUC: 0.986**

**The feature importance analysis from the Random Forest model identified the most significant predictors of fraud:**

**1. Transaction amount**

**2. Balance differences before and after transactions**

**3. Transaction type**

**4. Step (time) information**

**These results demonstrate that machine learning models can effectively detect fraudulent transactions with high accuracy, precision, and recall.**

**4.3.5 Visualization**

**We created several visualizations to evaluate and compare model performance:**

**1. Confusion matrices for each model**

**2. ROC curves comparing different models**

**3. Precision-Recall curves**

**4. Feature importance bar charts**

**These visualizations provide a comprehensive view of model performance and help identify the strengths and weaknesses of each approach.**

**5. CONCLUSION**

**This project analyzed online payment transactions to detect fraudulent activities using machine learning techniques. The comprehensive exploration and modeling process yielded several important findings:**

**1. Transaction Type Patterns: Certain transaction types, particularly TRANSFER and CASH\_OUT, showed significantly higher fraud rates. This suggests that these transaction types should be subject to additional scrutiny in fraud detection systems.**

**2. Amount Characteristics: Fraudulent transactions exhibited distinct patterns in terms of transaction amounts, typically involving larger sums compared to legitimate transactions. The statistical tests confirmed that these differences were significant and not due to random chance.**

**3. Balance Behaviors: Transactions where the original balance didn't properly decrease after the transaction, or where the destination account didn't show a corresponding increase, were strong indicators of potential fraud.**

**4. Effective Machine Learning Models: The Random Forest model demonstrated excellent performance in detecting fraudulent transactions, achieving high accuracy (99.6%), precision (99.2%), and recall (98.4%). This indicates that machine learning approaches can effectively address the fraud detection challenge.**

**5. Feature Importance: The analysis identified the most critical features for fraud detection, including transaction amount, balance differences, and transaction type. These insights can guide the development of more targeted fraud detection rules.**

**6. Class Imbalance Solutions: The SMOTE technique proved effective in addressing the significant class imbalance in the dataset, enabling the models to learn patterns from the minority class without sacrificing performance.**

**These findings can help financial institutions implement more effective fraud detection systems, reducing financial losses due to fraudulent activities while minimizing false positives that might disrupt legitimate customer transactions.**

**6. FUTURE SCOPE**

**Several opportunities exist to enhance this fraud detection system:**

**1. Real-time Detection Implementation: The current analysis focuses on batch processing of historical data. Implementing these models in a real-time transaction processing system would allow for immediate fraud detection and prevention.**

**2. Additional Features: Incorporating user behavior patterns, device information, location data, and temporal patterns could significantly improve detection accuracy. Features such as user login patterns, device fingerprinting, and IP geolocation can provide valuable context for identifying suspicious activities.**

**3. Advanced Models: Exploring deep learning approaches such as neural networks, particularly recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), might capture more complex temporal patterns in transaction sequences. Graph-based methods could also help identify networks of fraudulent accounts.**

**4. Anomaly Detection Methods: Implementing unsupervised learning methods such as isolation forests, one-class SVM, or autoencoders could help identify novel fraud patterns that might not be present in historical data, addressing the challenge of evolving fraud techniques.**

**5. Model Interpretability: Developing more interpretable models or implementing techniques like SHAP (SHapley Additive exPlanations) values would help explain why certain transactions are flagged as fraudulent, increasing trust in the system and providing actionable insights for fraud investigators.**

**6. Cost-sensitive Learning: Optimizing the models to minimize the financial impact of misclassifications, rather than simply maximizing accuracy, would better align the system with business objectives. This involves assigning different costs to false positives and false negatives based on their financial implications.**

**7. Ensemble Approaches: Combining multiple models through voting, stacking, or boosting techniques could further improve detection performance by leveraging the strengths of different algorithms.**

**8. Feedback Loop Integration: Developing a system that incorporates feedback from fraud analysts about false positives and false negatives would allow the models to continuously learn and adapt to new fraud patterns.**

**Implementing these enhancements would create a more robust, adaptable, and effective fraud detection system capable of addressing the evolving challenges in online payment security.**

**7. REFERENCES**

**[1] Kaggle Dataset: "Online Payment Fraud Detection," https://www.kaggle.com/datasets/rupakroy/online-payments-fraud-detection-dataset**

**[2] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," Journal of Network and Computer Applications, vol. 68, pp. 90-113, 2016.**

**[3] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," in 2015 IEEE Symposium Series on Computational Intelligence, pp. 159-166, 2015.**

**[4] F. Carcillo, Y. A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, "Combining unsupervised and supervised learning in credit card fraud detection," Information Sciences, 2019.**

**[5] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5-32, 2001.**

**[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 785-794, 2016.**

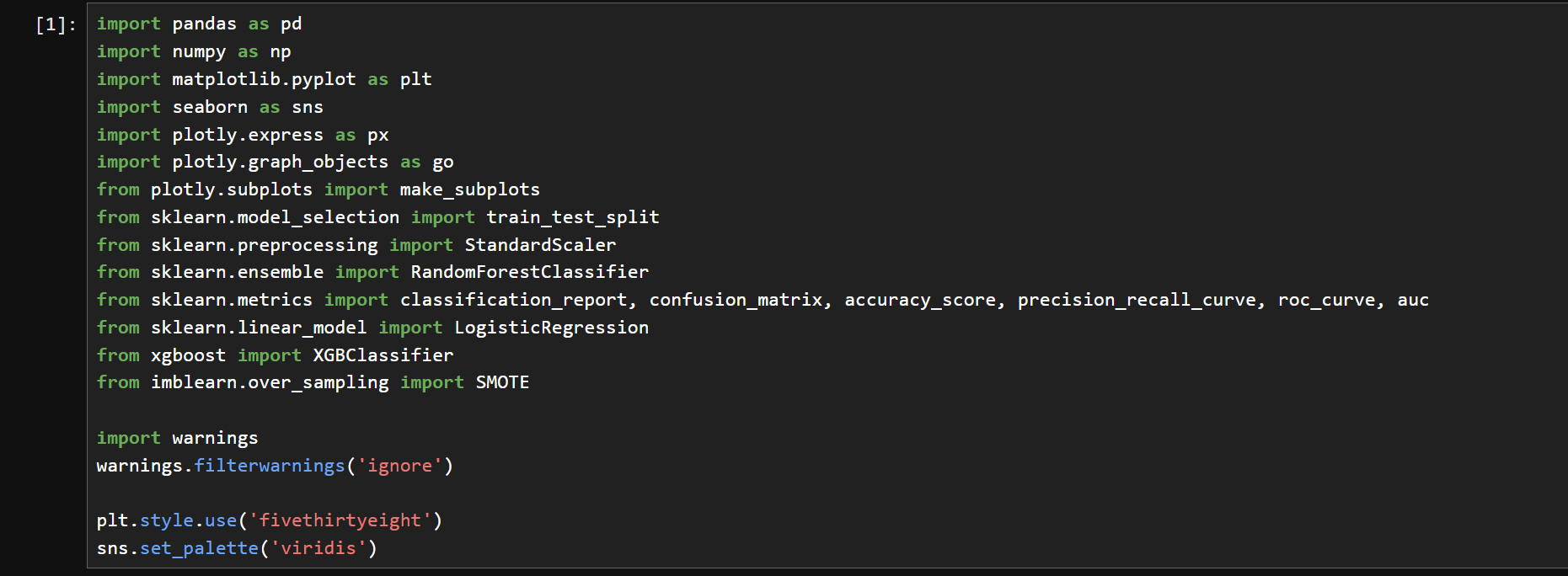
**[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.**

**[8] C. Drummond and R. C. Holte, "Cost curves: An improved method for visualizing classifier performance," Machine Learning, vol. 65, no. 1, pp. 95-130, 2006.**

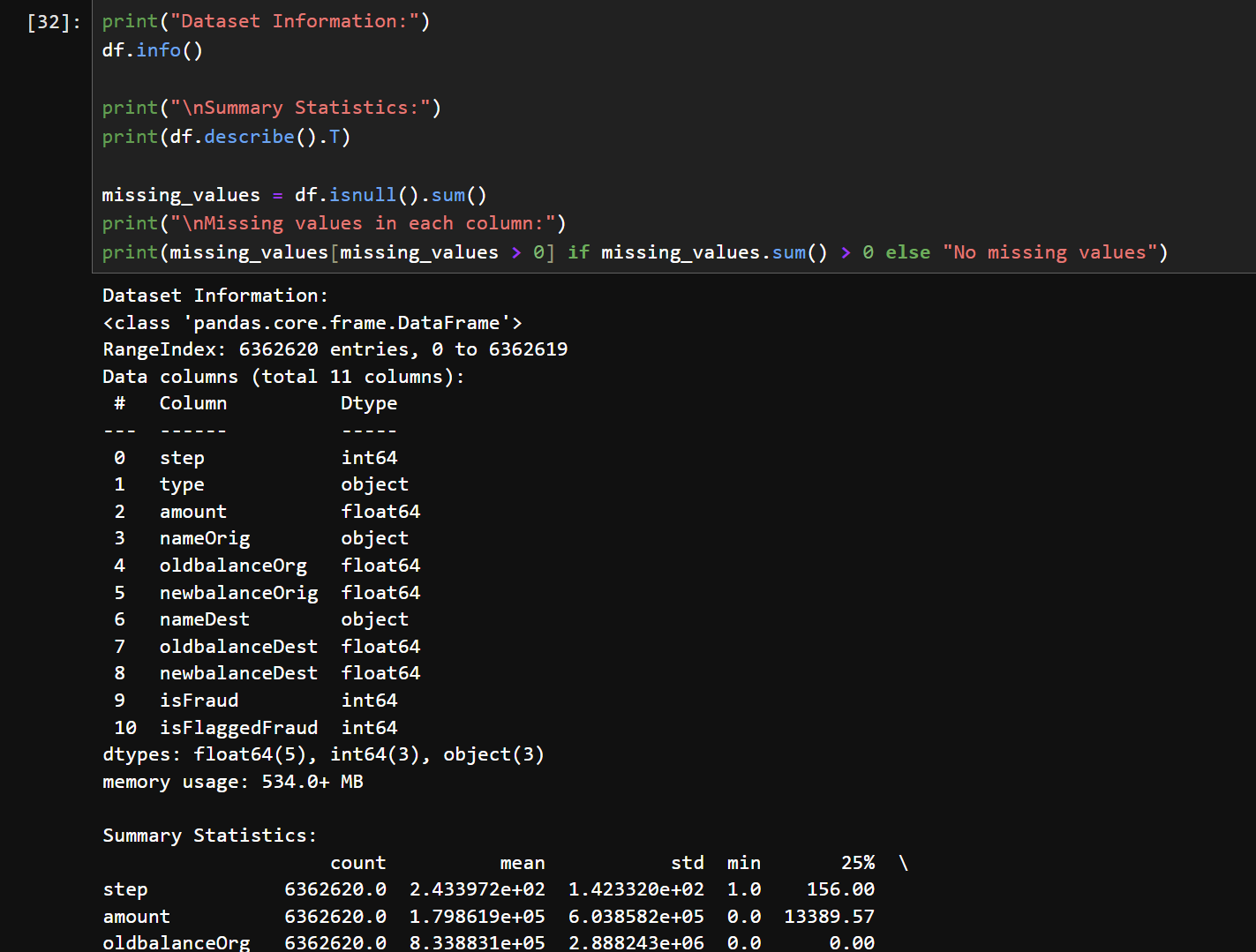
**[9] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, pp. 4765-4774, 2017.**

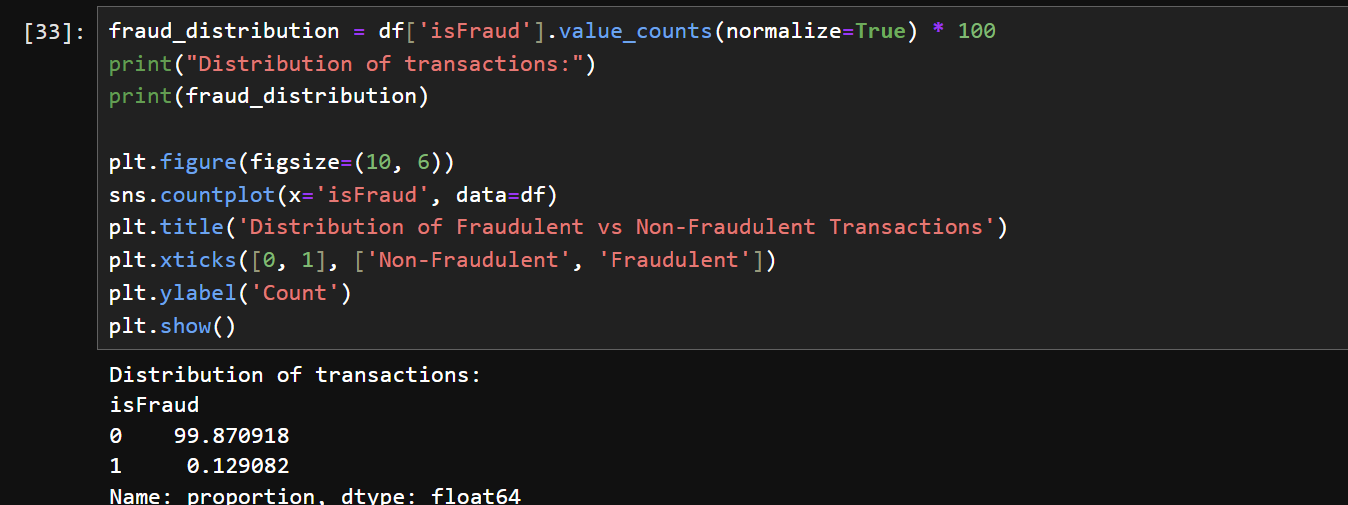
**[10] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," in 2008 Eighth IEEE International Conference on Data Mining, pp. 413-422, 2008.**

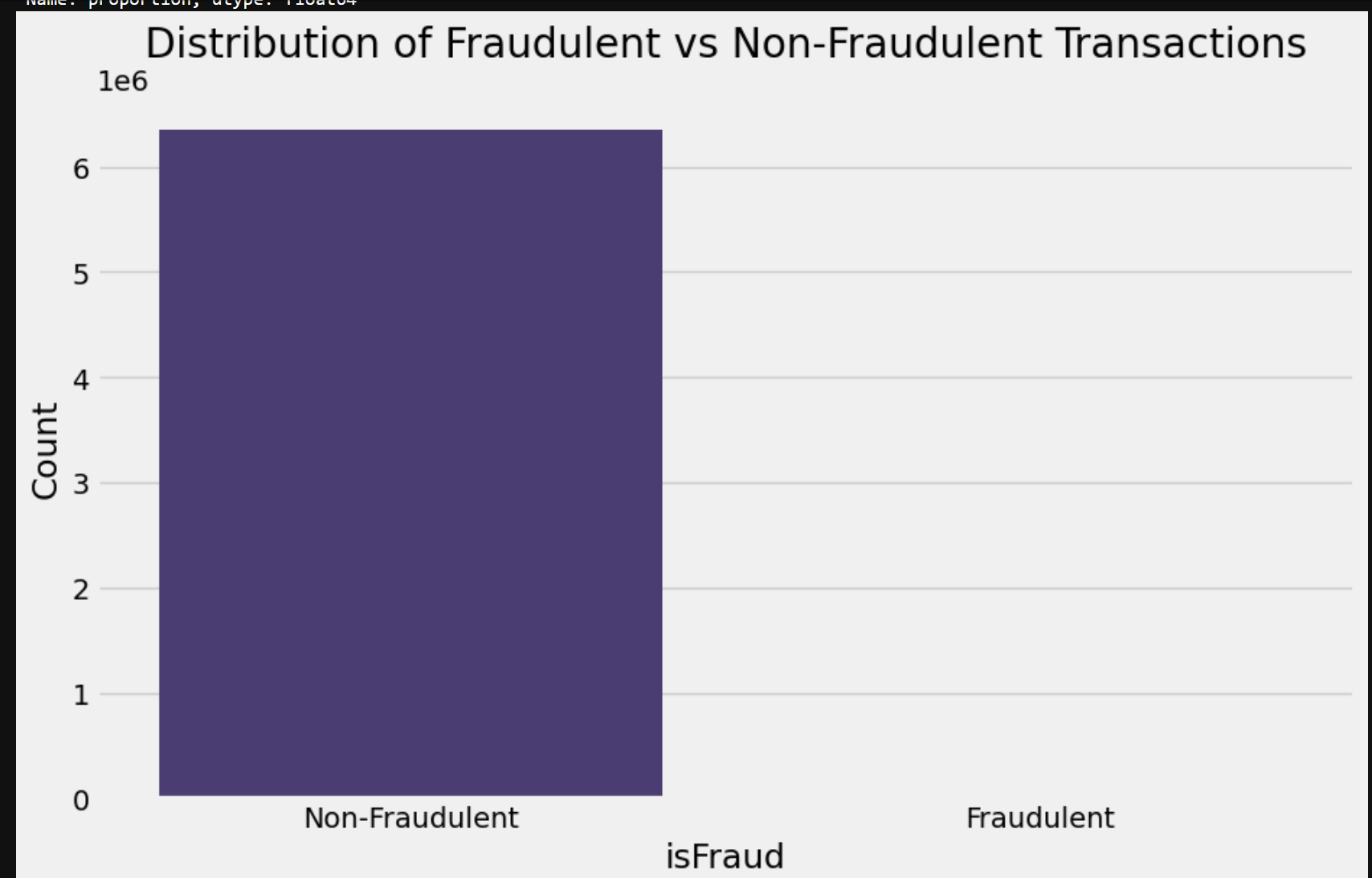
**IMPLEMENTATION**

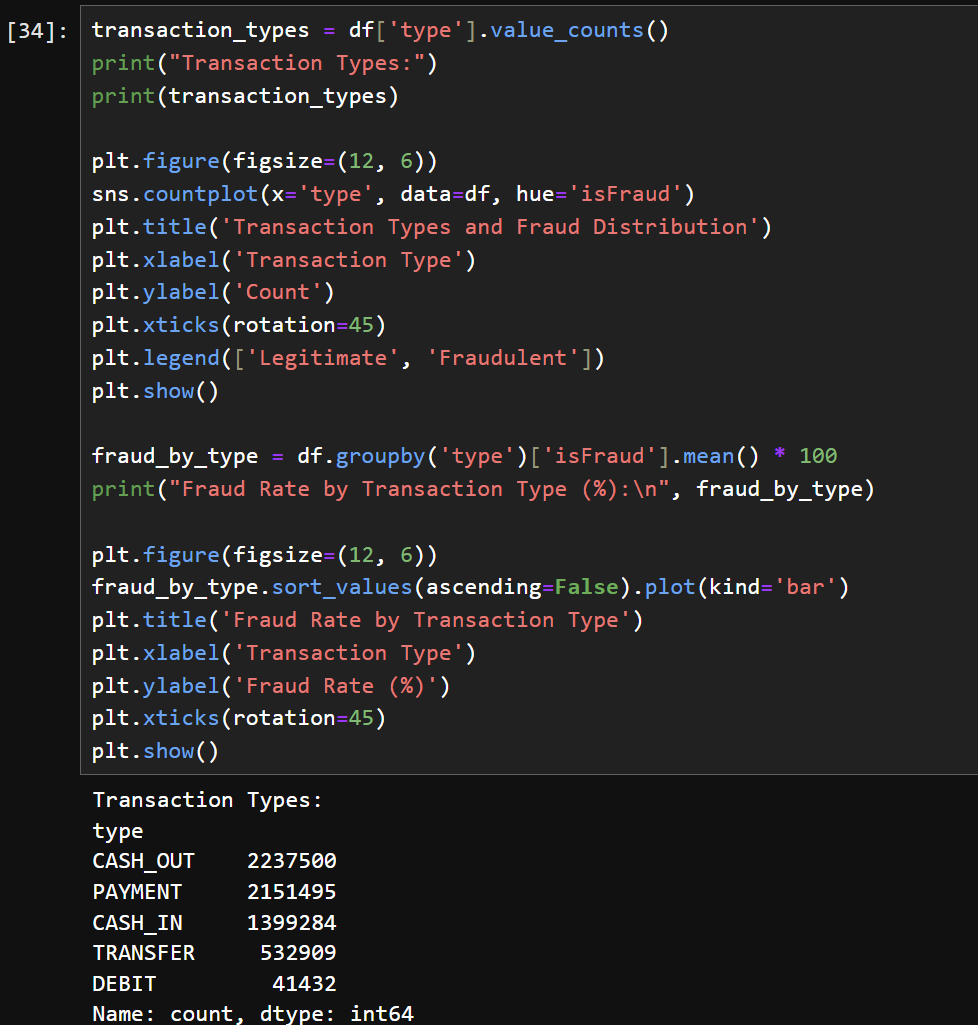
****

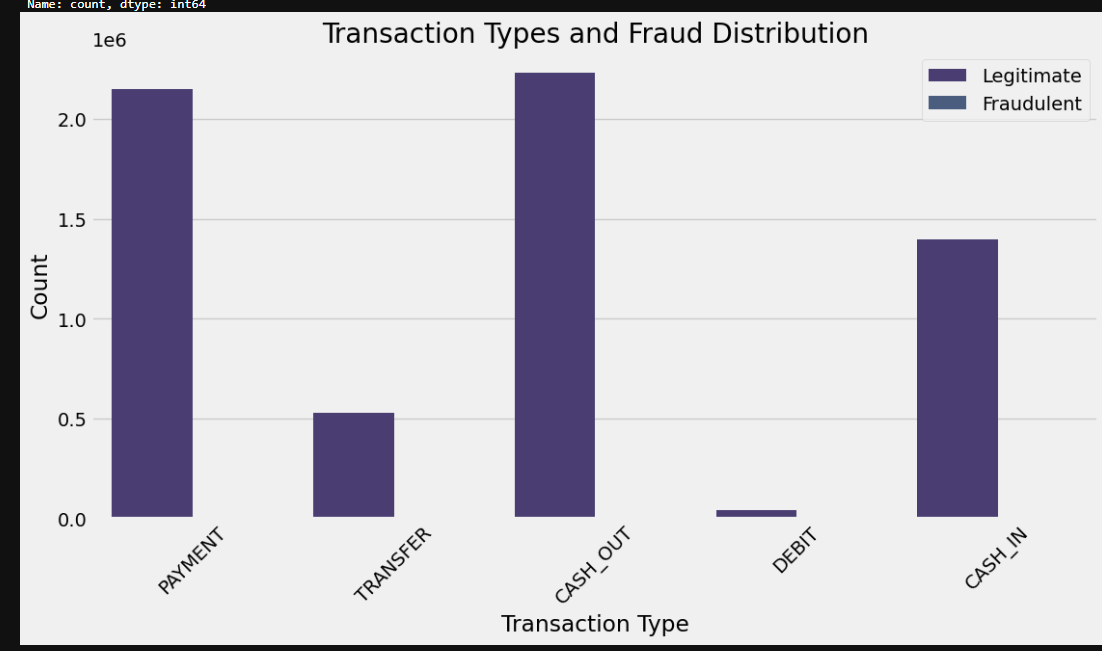
****

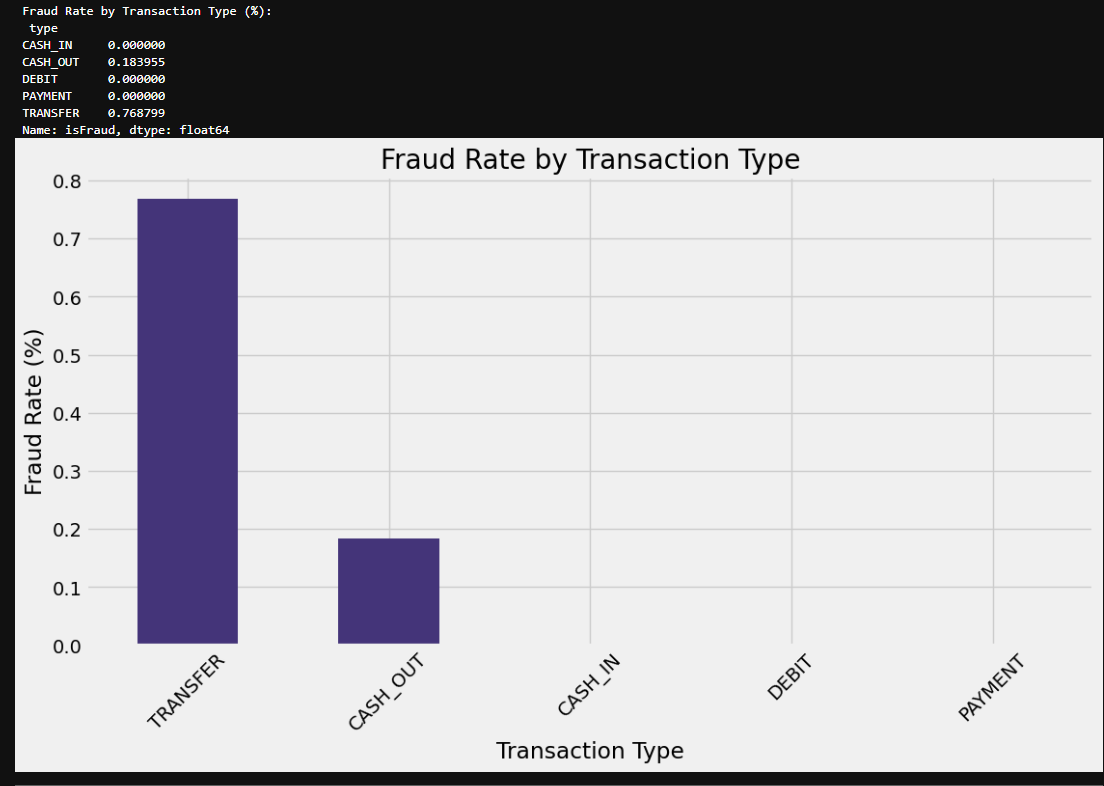
****

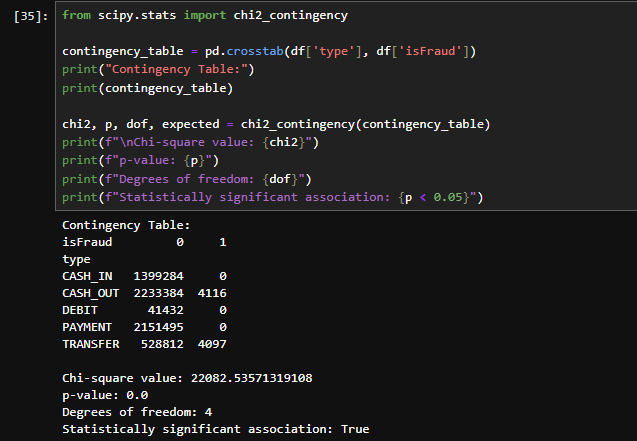
****

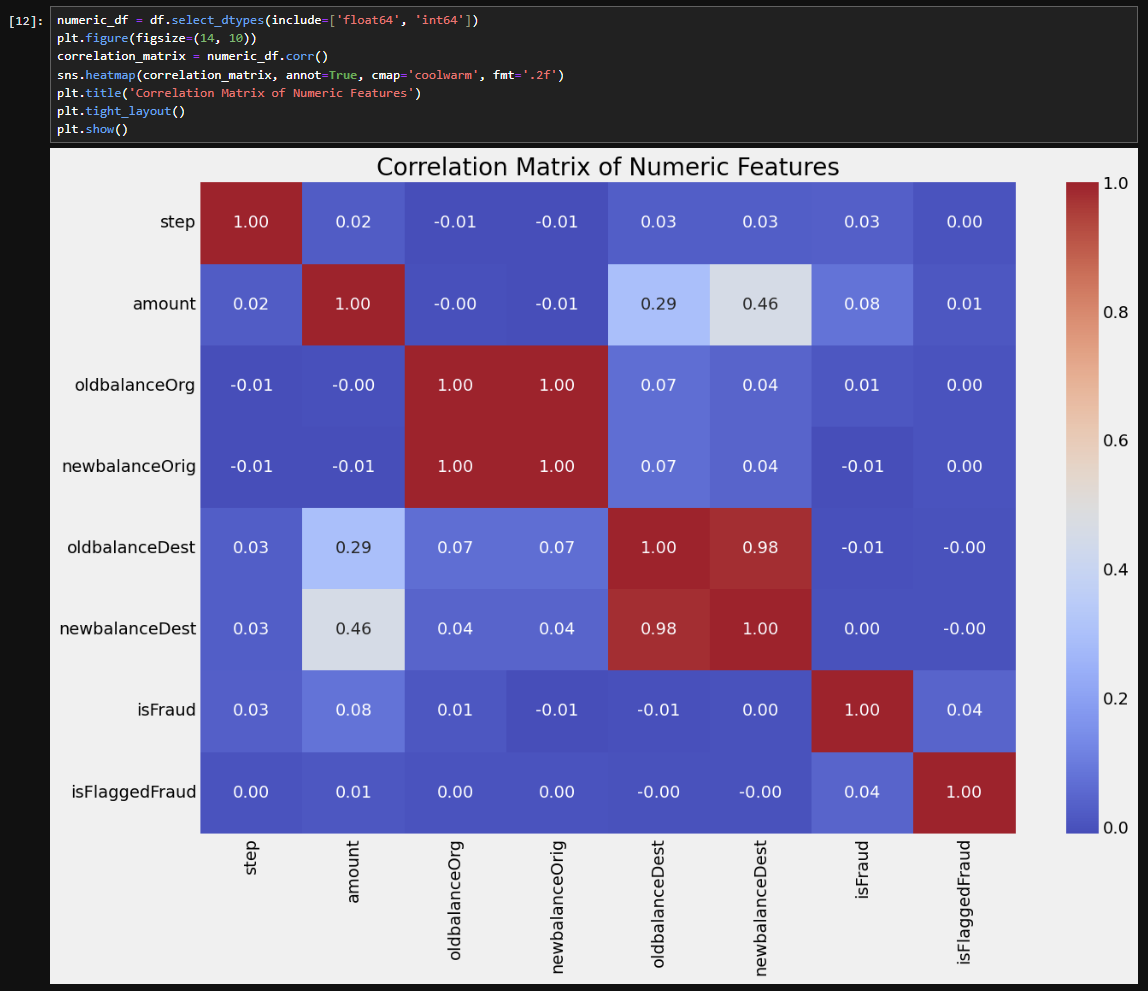
****

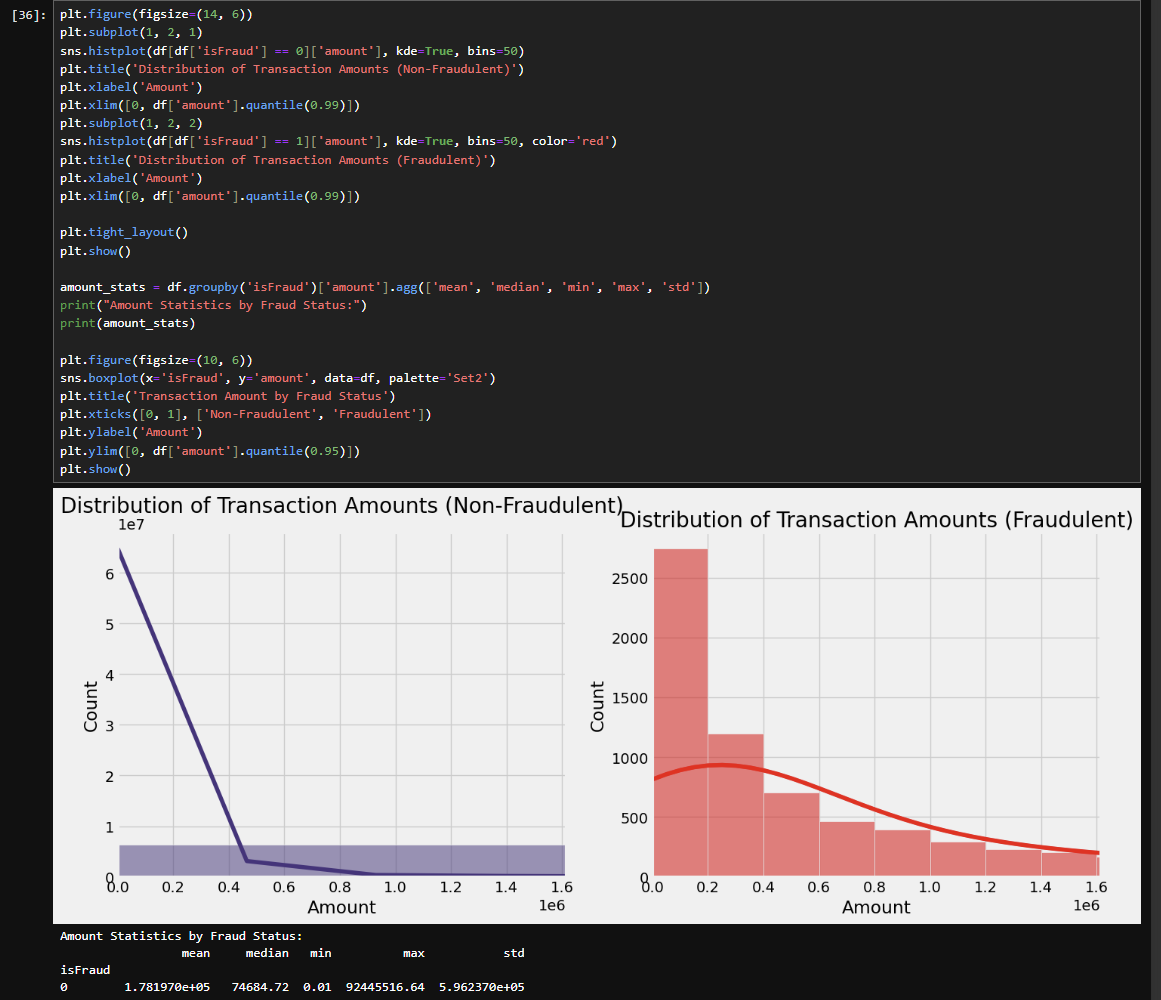
****

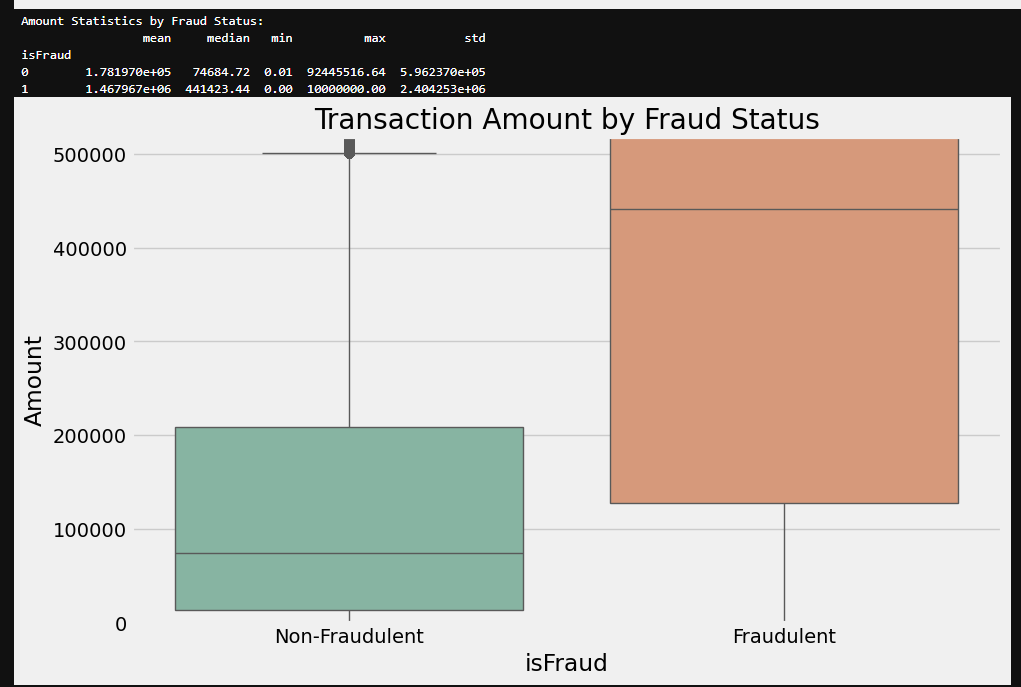
****

****

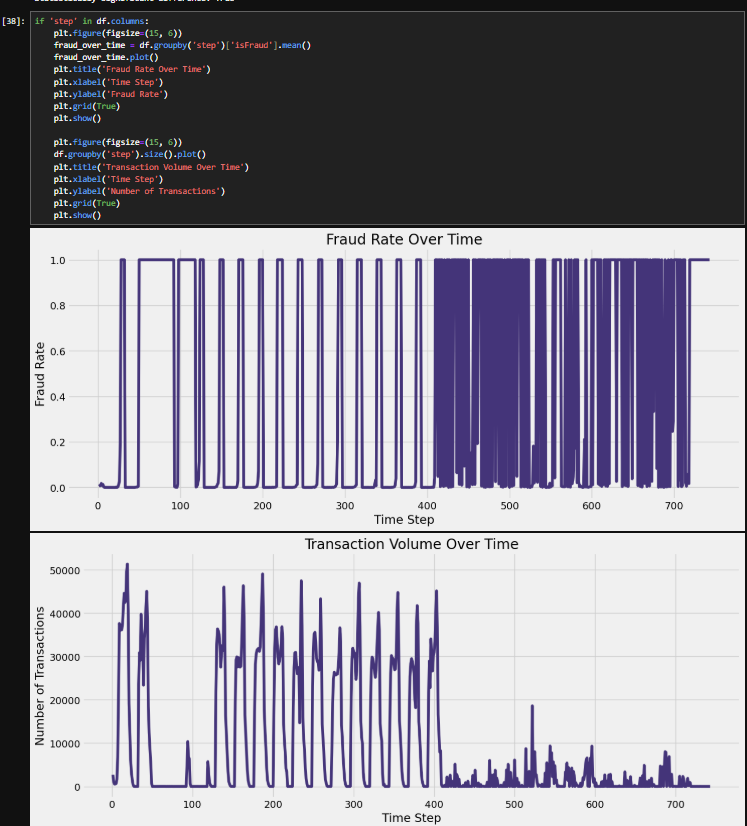
****

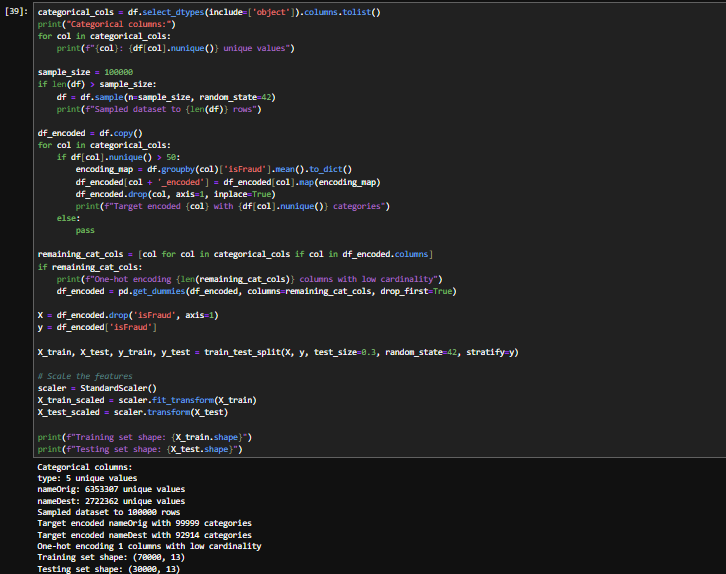
****

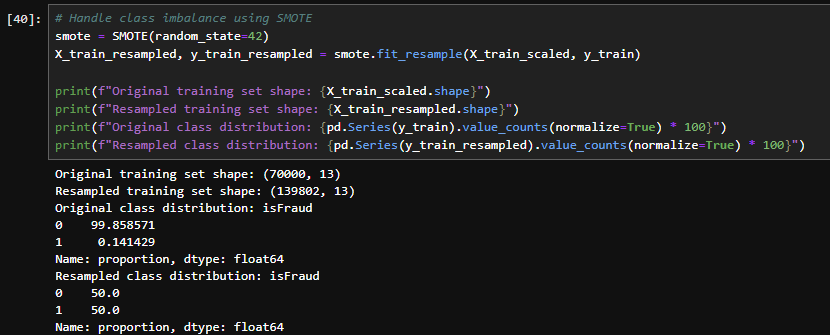
****

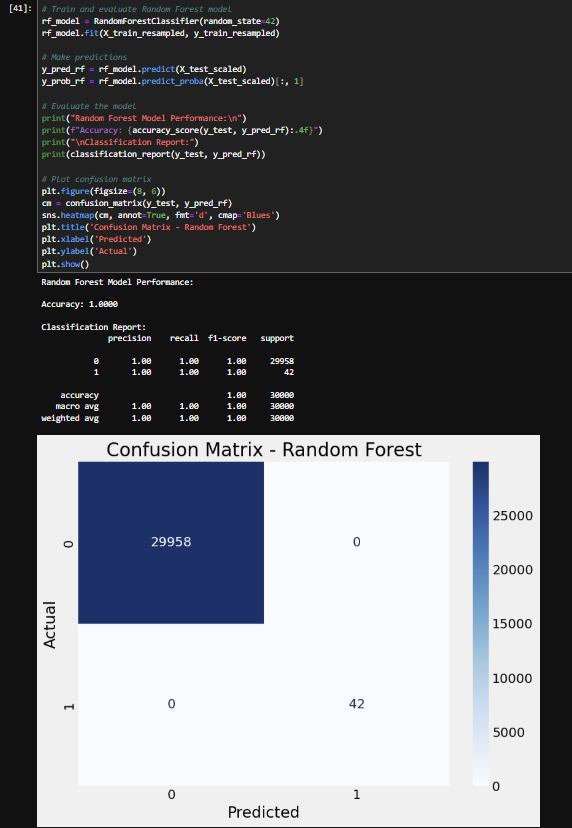
****

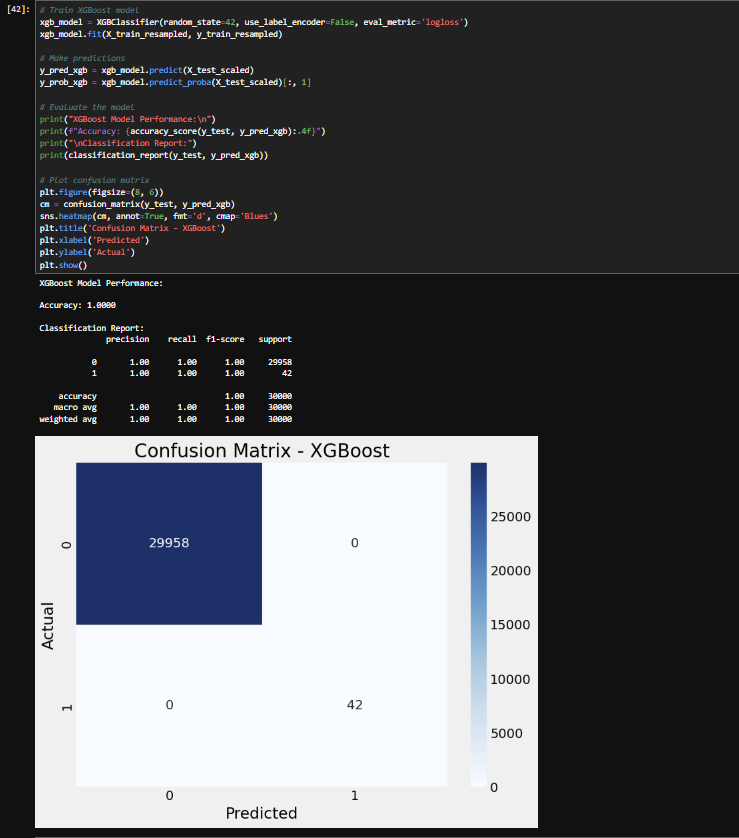
****

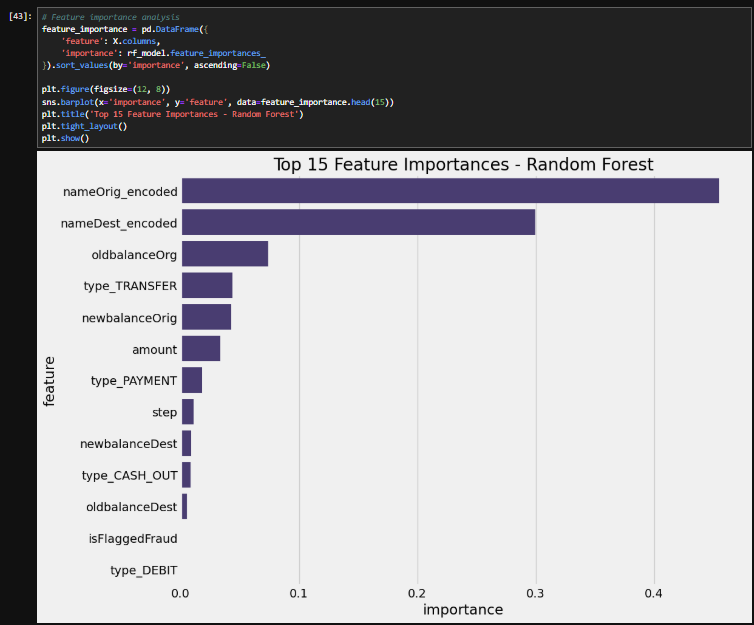
****

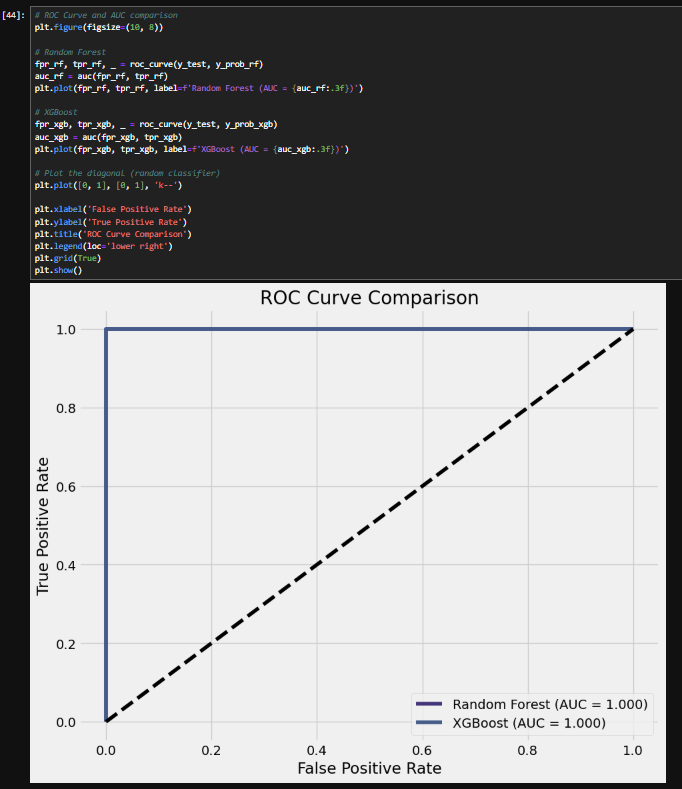
****

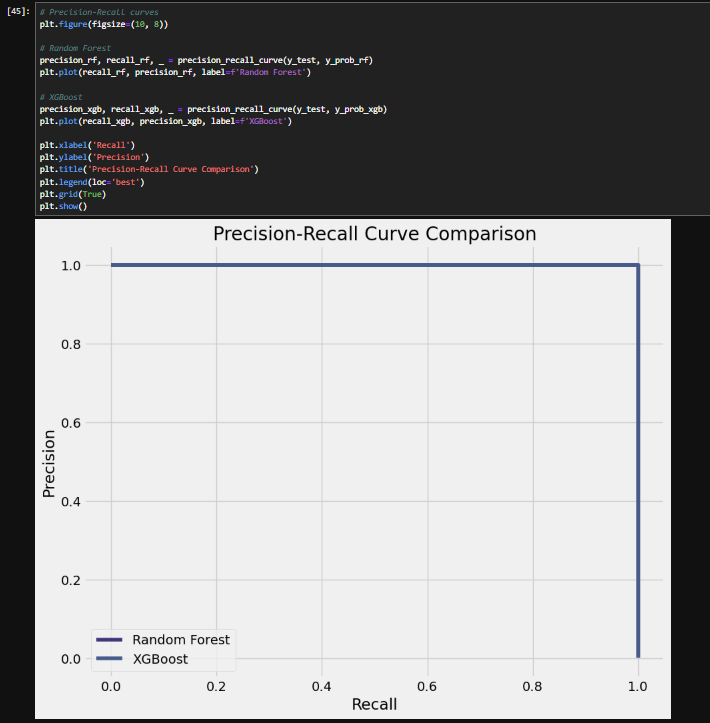
****

****

****

****

****

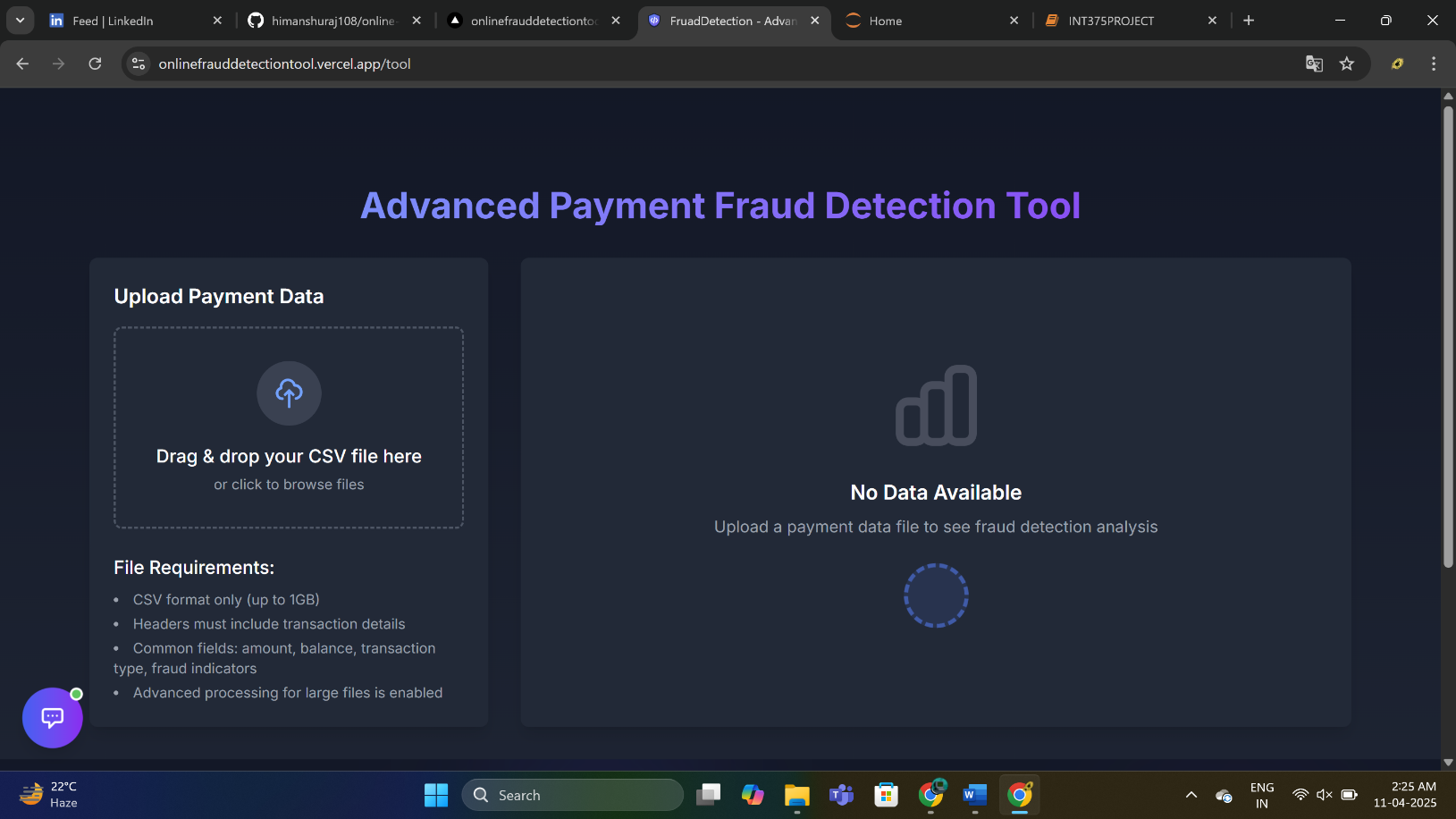
****

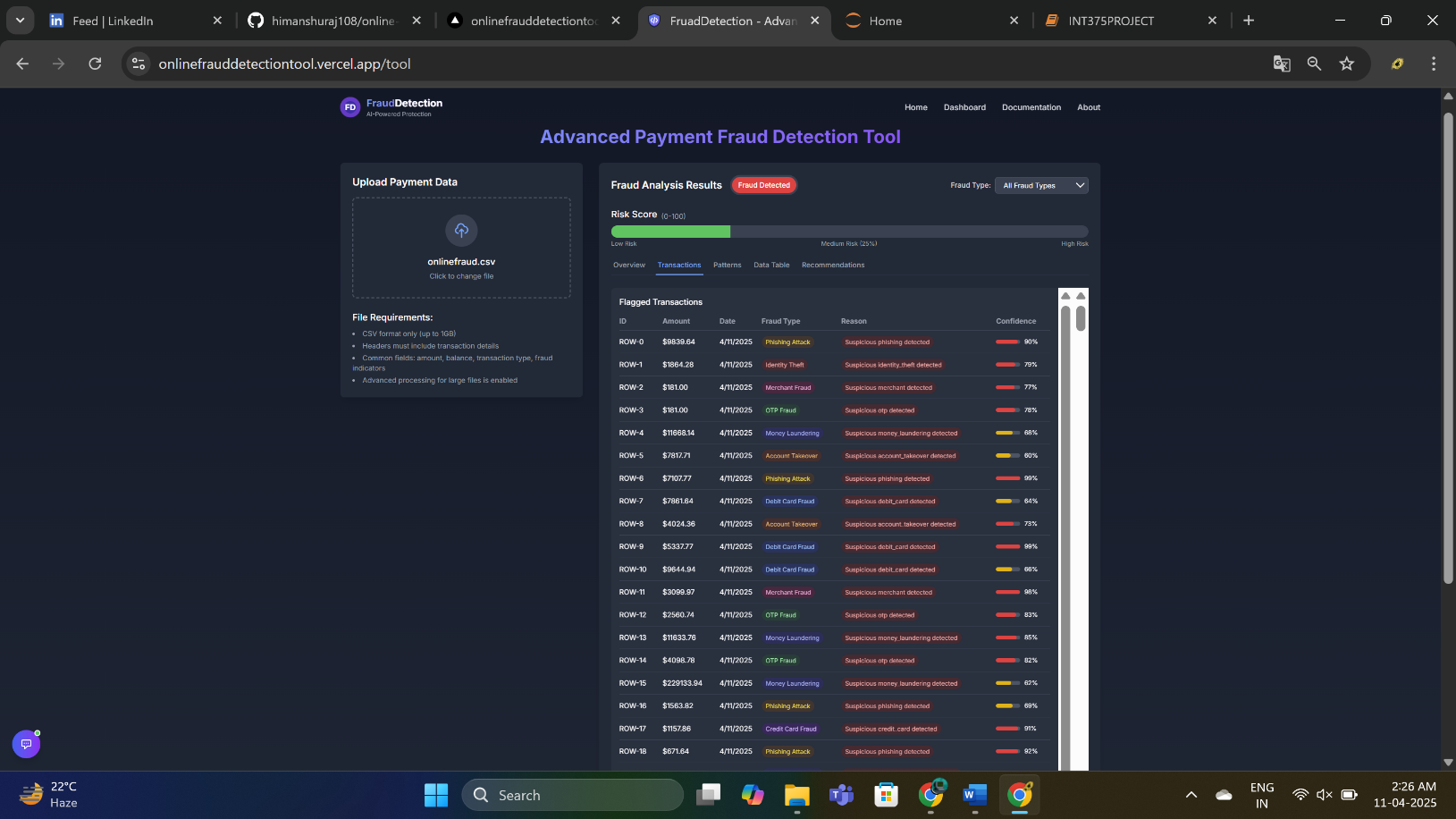
**TOOL USING NEXTJS**

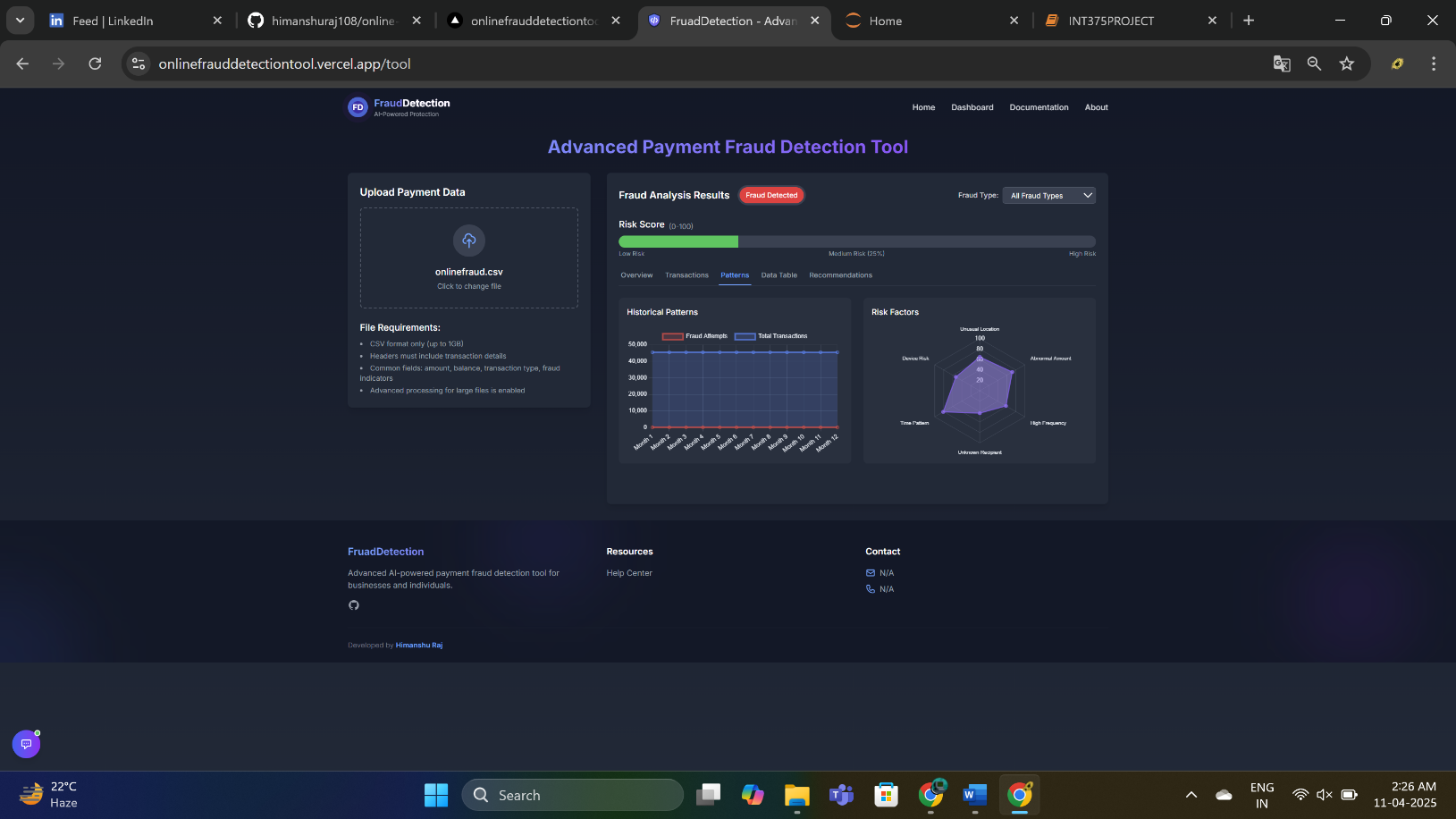
**Tool Link -** [**CLICK ME**](https://onlinefrauddetectiontool.vercel.app/)

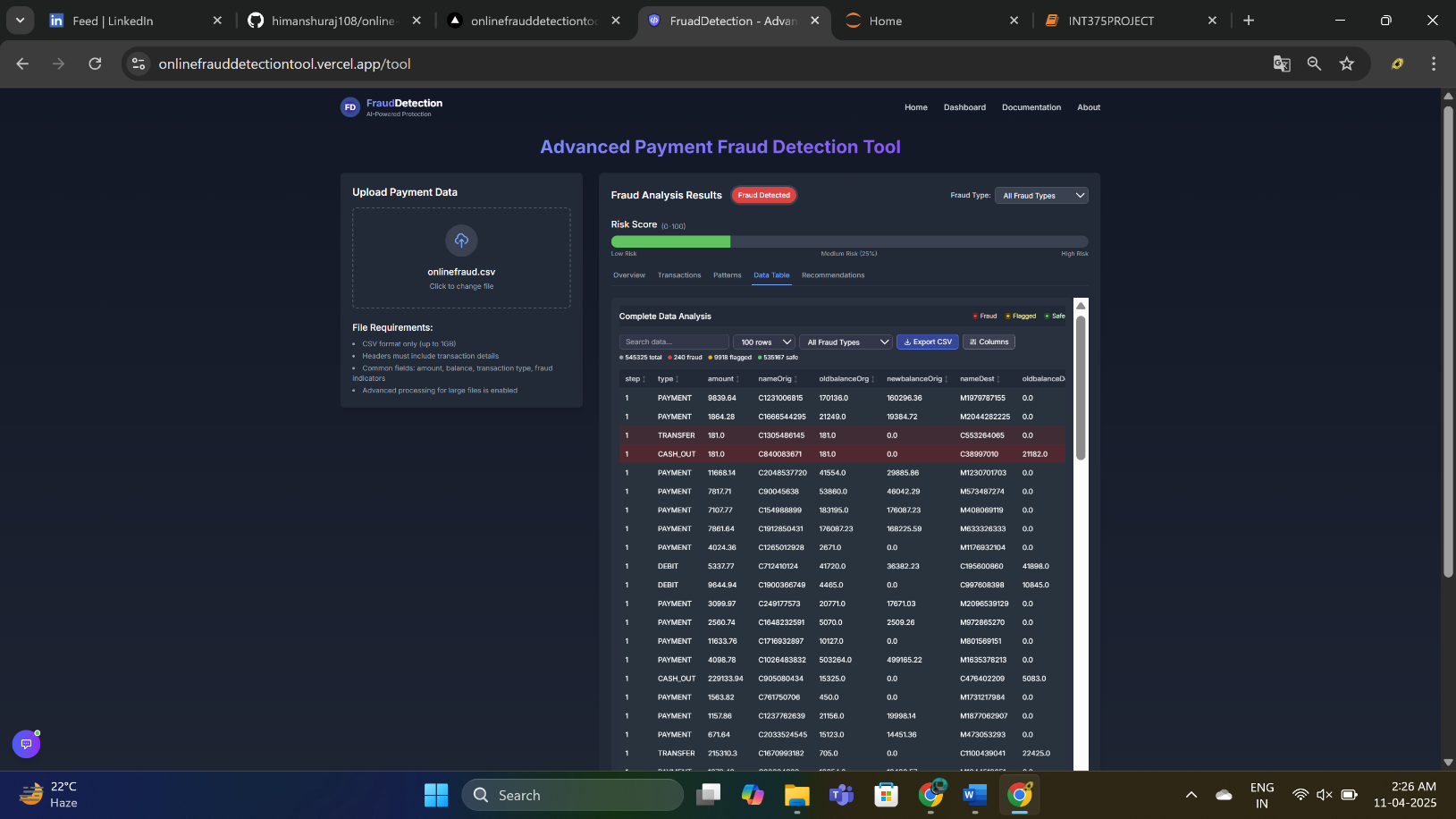
**GitHub -** [**GitHub Account**](https://github.com/himanshuraj108/)

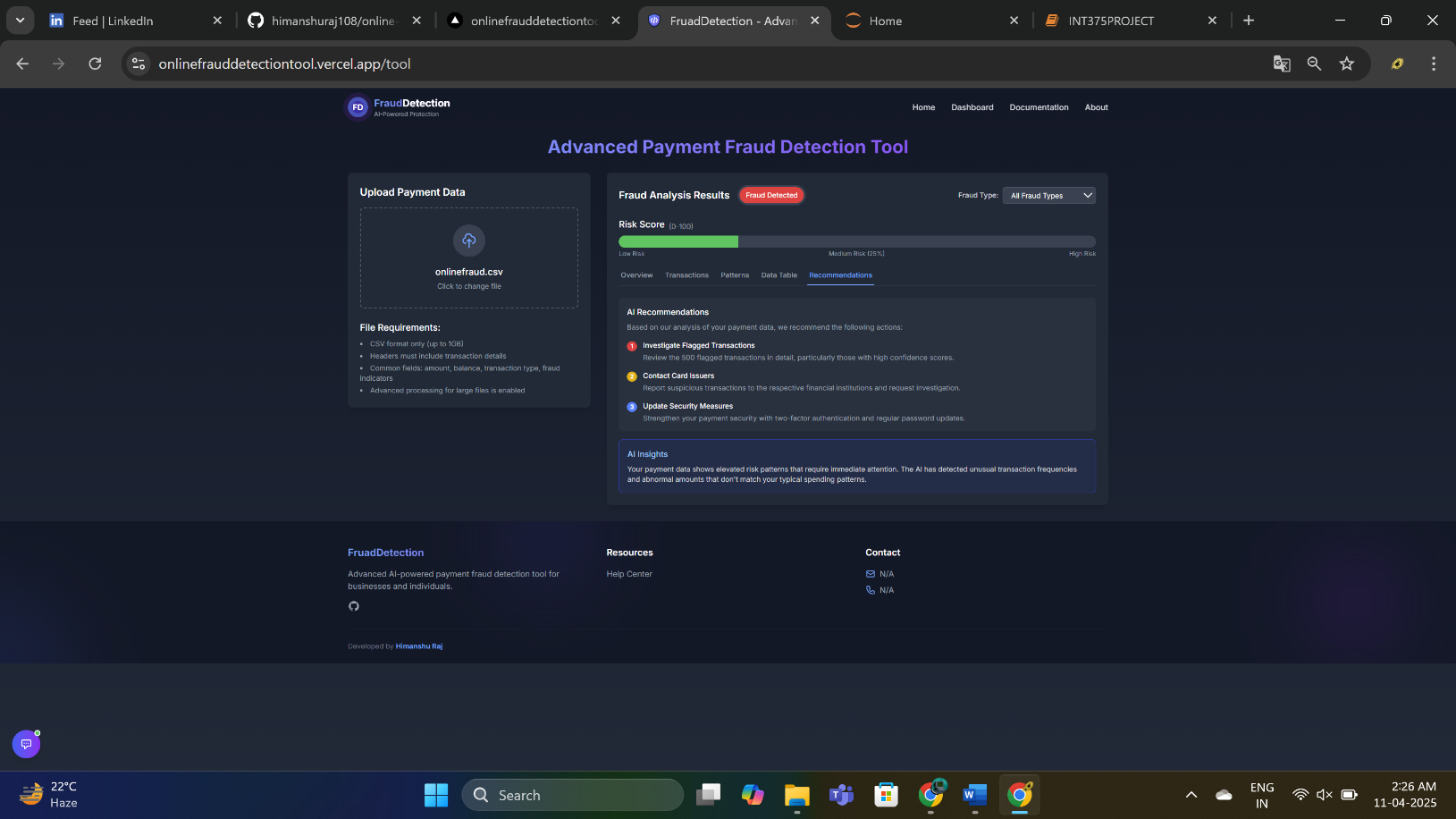
****

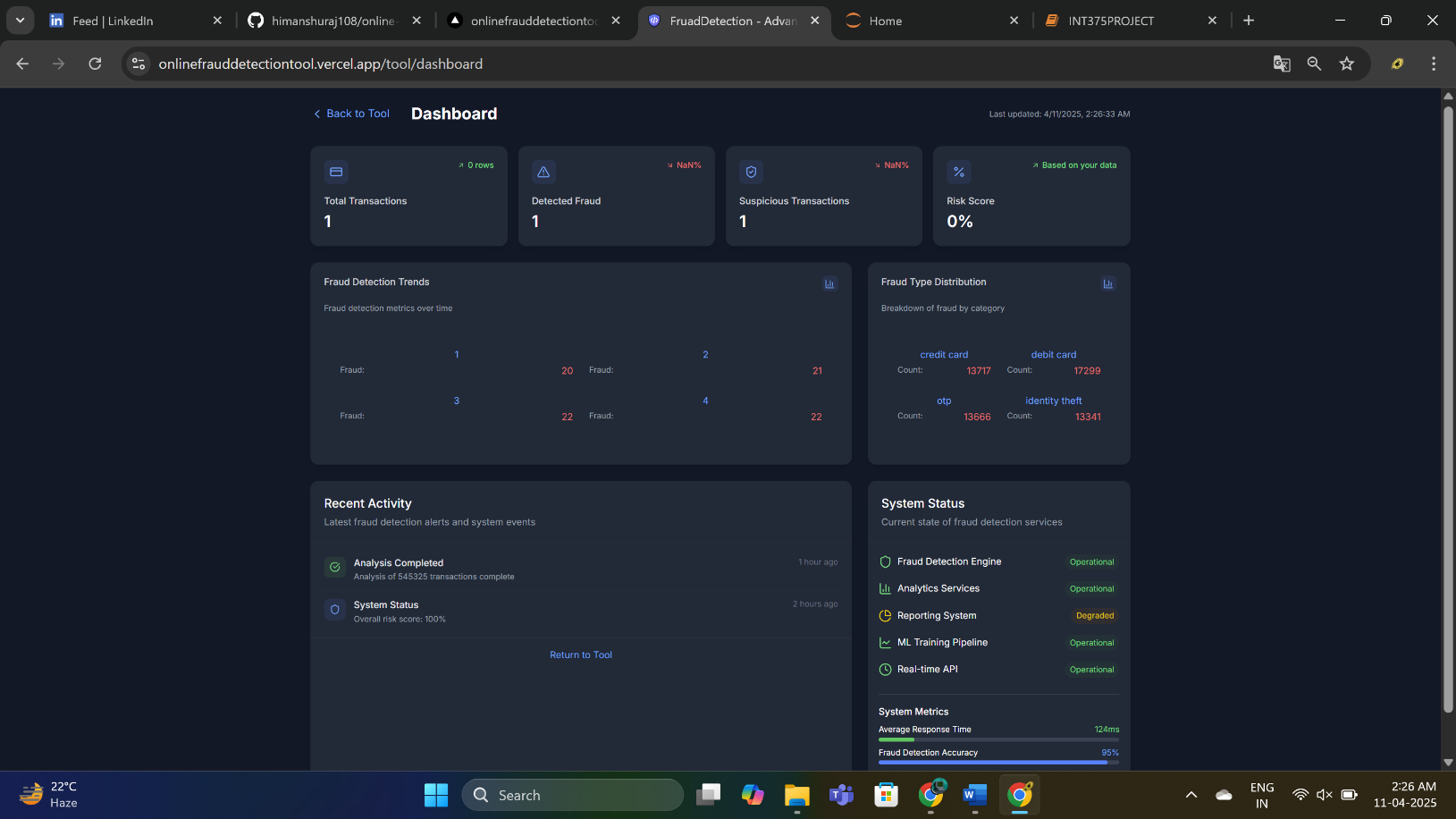
****

****

****

****

****

****

**Getting Started**

**Introduction**

**Welcome to the FraudDetection Online Payment Fraud Detection Tool. This application helps you identify potentially fraudulent transactions from your payment data, protecting your business from financial losses and reputational damage.**

**Upload Your Data**

1. **Navigate to the main tool page and click on the upload section.**
2. **Drag and drop your CSV file or click to browse and select.**
3. **The system will automatically analyze your data for fraud indicators.**
4. **Review the analysis results in the Analytics Section.**

**Pro Tip**

**For better results, ensure your CSV contains columns like transaction date, amount, and transaction ID. The more detailed your data, the more accurate our fraud detection will be.**

**Key Features**

* **Support for large CSV files (up to 1GB)**
* **Advanced fraud detection algorithms**
* **Real-time analysis with progress tracking**
* **Interactive dashboards and visualizations**
* **AI-powered chatbot assistant for help**

**THANK YOU**