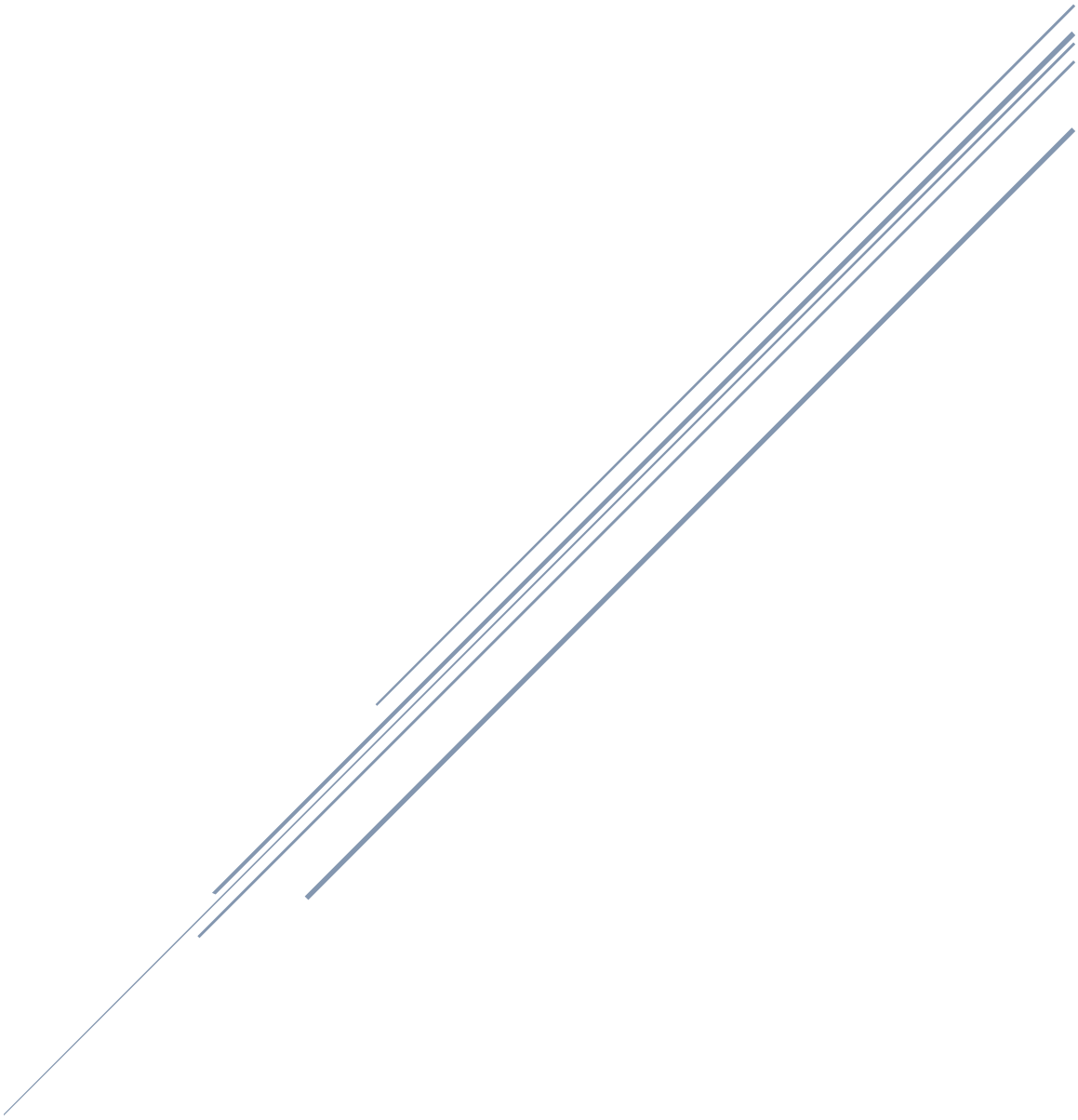
Fraud Shield | Securing Credit Card Transactions

MILESTONE 3 – PROJECT 1



Bellevue University

DSC680-T302 Applied Data Science (2241-1)

Business Problem

With the growing prevalence of credit card fraud, businesses, and financial institutions face a critical challenge. We need effective solutions to detect and prevent fraudulent transactions in real time. This project aims to use machine learning to build a robust fraud prediction system that distinguishes legitimate from fraudulent credit card transactions. The goals include enhancing security, reducing false positives, improving the customer experience, and staying ahead of evolving fraud tactics. This endeavour is vital for financial security, customer trust, and the industry's overall well-being.

Background/ History

Credit card fraud detection has evolved alongside the growth of credit card usage. Initially relying on manual methods, it transitioned to electronic authorization with magnetic stripes. As fraudsters adapted to exploit vulnerabilities, rule-based systems emerged. Recent advances in machine learning and real-time analytics have revolutionized the field, with systems now using behavioural analytics to identify fraud. Collaboration and data sharing play a crucial role in combating evolving fraud tactics, ensuring the ongoing evolution of fraud detection methods.

Datasets

The datasets used in this project is sourced from Kaggle (Shenoy, K. (2020, August 5)), a public-domain dataset. This dataset contains stimulated credit card data with legitimate and fraudulent credit card transactions from Jan 2019 to Dec 2020.

Date: 9/14/2023
By: Aarti Ramani

Data Preparation

The following steps were performed to prepare the data for modelling.

i. Checked for null rows/columns in the data.

ii. Performed check for duplicates.

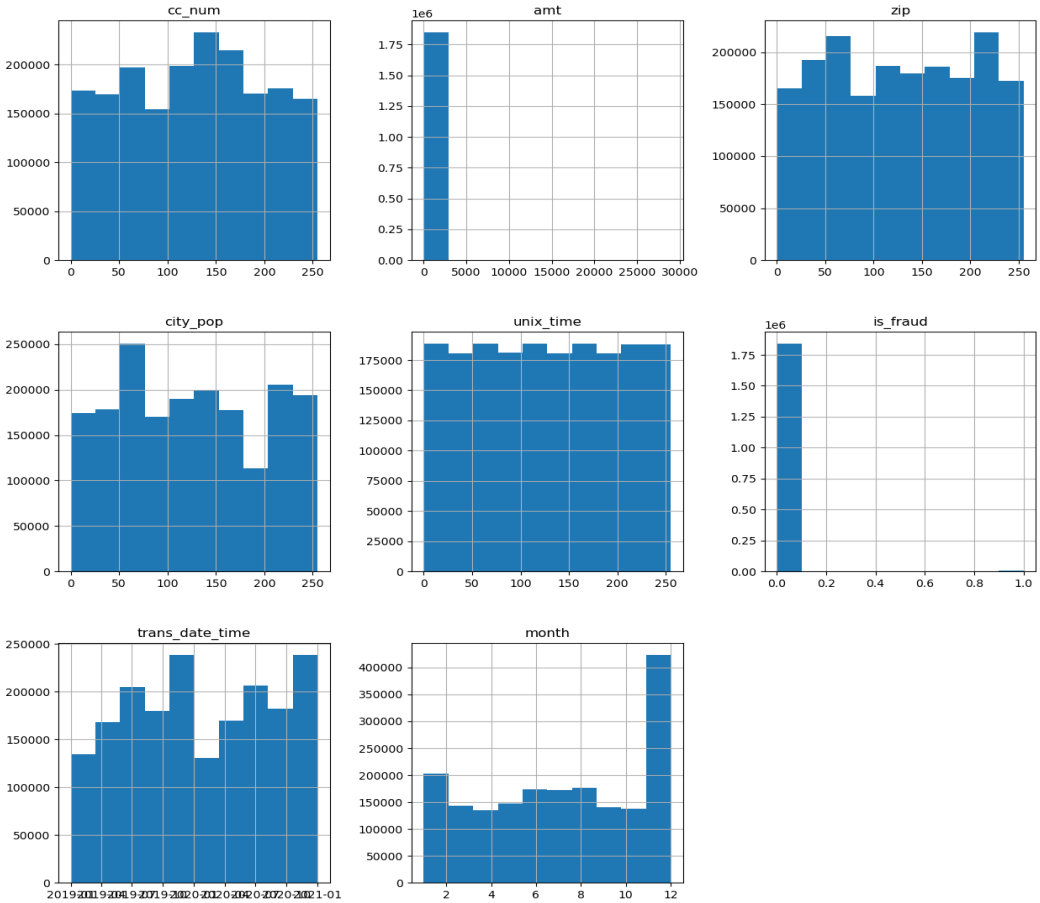
iii. Converted datetime string to a datetime datatype.

iv. Added a month column for visualization.

v. Dropped columns (‘Unnamed: 0' and 'trans_date_trans_time’)

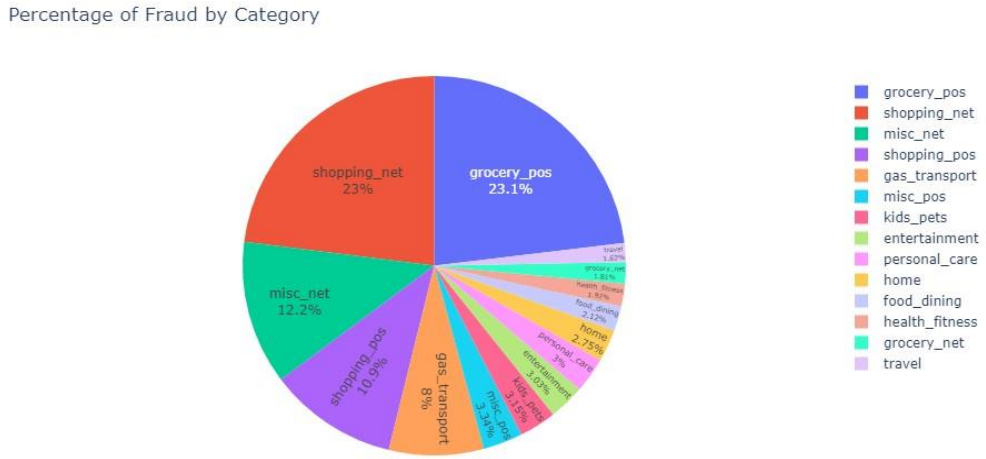
Visualizations

Numeric Variables Distribution:

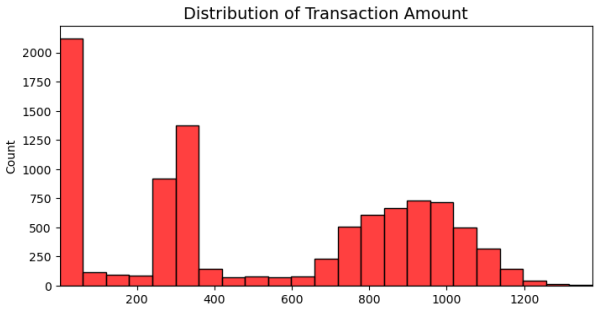


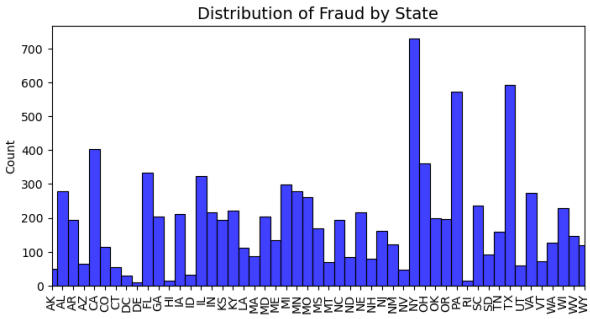
Date: 9/14/2023
By: Aarti Ramani

Percentage of Fraud by Category

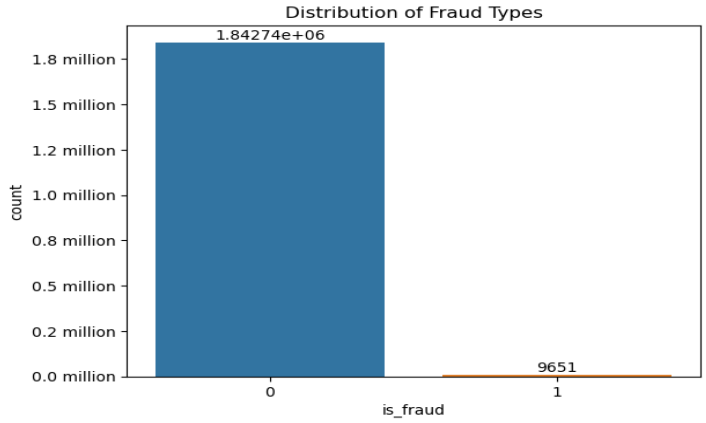


Distribution of Amount & State





Distribution of Fraud Types



Date: 9/14/2023
By: Aarti Ramani

To construct an effective model, it was essential to address the dataset's imbalance. To achieve this, the Synthetic Minority Oversampling Technique - SMOTE (Appendix (ii)) was employed.

Methods

With the balanced dataset, the subsequent step was to partition the data into training and testing datasets.

The following models were developed, with their respective outcomes recorded.

i. Logistic Regression: Using a Logistic Regression model for credit card fraud detection has benefits including interpretability, computational efficiency, and suitability for smaller datasets. It provides insights into feature impacts, acts as a baseline, and offers transparent probability estimations. However, its linear nature might limit performance on complex, non-linear data patterns.

ii. Random Forest: Using a Random Forest model for credit card fraud detection offers benefits like ensemble learning, robustness against noise and outliers, handling imbalanced data, and capturing complex patterns. Its ability to handle non-linearity, easy tunability, and feature importance analysis make it a strong choice.

iii. Gradient Boosting: Gradient Boosting is a powerful method for credit card fraud detection due to its high predictive accuracy, ensemble nature that reduces overfitting, and the ability to handle complex patterns in the data. It also offers insights into feature importance, aiding in identifying relevant variables for fraud detection.

Both Random Forest and Gradient Boosting are ensemble methods that excel in handling credit card fraud detection tasks. Random Forest builds diverse trees independently, while Gradient Boosting sequentially improves the errors of previous trees.

iv. Neural Networks: Neural Network (NN) models offer advantages in credit card fraud detection due to their ability to capture intricate non-linear patterns in data. They excel in feature learning from raw data, making them adept at detecting complex and evolving fraud behaviours.

Analysis

Since the dataset consists of categorical features, it is essential to represent the categorical data in a numerical format. For this, an encoding technique (Label Encoder) was implemented. Additionally, the StandardScaler preprocessing technique was used to standardize or normalize numerical features in the dataset.

The models were then built, and the outcomes were recorded as follows.

Random Forest Classifier

Fraud Type	Precision	Recall	F1-Score
0 – non-Fraud	1.00	1.00	1.00
1 - Fraud	0.79	0.72	0.75

Accuracy	ROC-AUC Score
99.75%	0.86

Logistic Regression

Fraud Type	Precision	Recall	F1-Score
0 – non-Fraud	1.00	0.94	0.97
1 - Fraud	0.06	0.77	0.12

Accuracy	ROC-AUC Score
93.99%	0.85

Gradient Boosting

Fraud Type	Precision	Recall	F1-Score
0 – non-Fraud	1.00	1.00	1.00
1 - Fraud	0.50	0.89	0.64

Accuracy	ROC-AUC Score
99.4%	0.94

Neural Network Model

Fraud Type	Precision	Recall	F1-Score
0 – non-Fraud	1.00	0.99	0.99
1 - Fraud	0.26	0.75	0.38

Accuracy	ROC-AUC Score
98.7%	0.87

Conclusion:

In summary, this project focused on developing and implementing effective credit card fraud detection models. We explored different machine-learning techniques, including logistic regression, random forests, gradient boosting, and neural networks.

Through data preprocessing, feature engineering, and model evaluation, we aimed to create accurate and efficient fraud detection systems. The models demonstrated promising results, with some outperforming others in specific areas.

Model	Random Forest Classifier	Logistic Regression	Gradient Booster	Neural Network
Accuracy	99.75%	93.99%	99.4%	98.7%

Date: 9/14/2023
By: Aarti Ramani

With an accuracy of 99.75%, the Random Forest Classifier is the best-performing model for this dataset. However, the dataset is highly imbalanced, with a significantly larger number of non-fraudulent transactions. This can impact the model's performance and interpretation of metrics.

While all models demonstrated high accuracy and impressive performance in identifying non-fraudulent transactions, there is a trade-off between precision and recall for detecting fraudulent transactions.

Assumptions

Simulated data often assumes an imbalanced class distribution, with a small proportion of transactions representing fraud. This project assumes that features used for modelling are independent or have certain dependencies that mimic real-world relationships.

Limitations

Credit card fraud detection systems have the risk of false positives, where legitimate transactions are mistakenly flagged as fraud, and false negatives, where some fraudulent transactions go undetected. Imbalanced data and model complexity can affect performance, while the threat of adversarial attacks and the need for data privacy are ongoing concerns.

Challenges

Ensuring security for simulated data and achieving real-world applicability could be areas of concern. Simulated datasets might not fully capture the complexities of actual fraud scenarios, potentially limiting the model's effectiveness in real-world situations.

Date: 9/14/2023
By: Aarti Ramani

Future Uses/Additional Applications

The applications of fraud detection techniques are expanding across various industries and sectors as organizations seek to protect themselves from evolving threats and optimize their operations. For example, fraud detection techniques can be extended to other payment methods, such as mobile wallets, digital currencies (cryptocurrencies), and peer-to-peer payment systems.

Recommendations

- Incorporating behavioural analysis to detect anomalies in customer transaction behaviour over time.
- Ensuring Fraud detection systems comply with financial regulations and data privacy laws.

Implementation Plan

The implementation plan for credit card fraud detection involves defining clear objectives, collecting, and preparing transaction data, selecting appropriate machine learning models, addressing class imbalance, evaluating model performance, and integrating it into a real-time system with continuous monitoring and alerts.

Ethical Assessment

Several ethical considerations are crucial for this project:

i. Data Privacy: Despite the simulation, safeguarding privacy is essential. Ensuring simulated data does not resemble real customer information avoids accidental exposure of PII (Appendix (i)).

ii. Informed Consent: Transparency builds trust and addresses concerns. Understanding if consent was obtained for real-based simulations, even if anonymized, adds ethical integrity.

iii. Intent and Use: Ethical utilization of simulated data is paramount, prohibiting any malicious or harmful intent.

iv. Data Security: Robust security measures should be applied to the simulated dataset, treating it with the same importance as real customer data.

v. Stakeholder Implications: The viewpoints of stakeholders like credit card issuers, customers, and regulators to ensure ethical alignment should be considered.

References:

Shenoy, K. (2020, August 5). *Credit Card Transactions Fraud Detection Dataset*. Kaggle. https://www.kaggle.com/datasets/kartik2112/fraud-detection?select=fraudTrain.csv

Appendix:

i. PII stands for Personally Identifiable Information. It refers to any data that can be used to identify a specific individual. Protecting PII is crucial to safeguard individuals' privacy and prevent identity theft, fraud, and unauthorized access to sensitive information.

ii. SMOTE - Synthetic Minority Oversampling Technique (SMOTE), generates synthetic instances for the minority class, thus balancing the dataset.

Questions

1. What is Credit card fraud?
2. What is Credit card simulated data?
3. Why do we need to simulate the credit card data?
4. What is a Neural Network Model? Why was this selected for the fraud detection project?
5. What is the difference between the Random Forest and Gradient Boost model?
6. Which model performs better between the Random Forest and the Gradient Boost models?
7. Will these models withstand vast datasets?
8. Can we provide real-time fraud detection using these implemented models?
9. Will these implemented models work with data from various credit card firms?
10. Can these models be extended for fraud detection in any other sector?

Date: 9/14/2023
By: Aarti Ramani