MSDS692: Data Science Practicum Report

Nahian Arnob
Master in Data Science
Anderson College of Business and Computing
Regis University, Denver, CO, USA
narnob@regis.edu

Card Fraud Detection using Machine Learning Model

Abstract

Card fraud has become a major problem for the current generation. Financial institutions lose millions of currency globally, affecting more than thousands of people. It has become increasingly difficult to distinguish genuine transactions from fraudulent merchants before making online transactions. In between millions of transactions, fraudulent transactions only represent a very small number; this creates a highly imbalanced dataset for such cases. Therefore, there is a very high importance of such data-driven solutions that can identify fraudulent transactions in real-time.

To identify fraudulent credit card transactions using previous transaction data, the project suggests a machine learning-based fraud detection system. To differentiate between legitimate and fraudulent transactions, it inspects supervised learning techniques, which include logistic regression and random forest classifiers. These models are conditioned and assessed using transaction-level features acquired from a publicly available dataset, with performance assessed using estimation of metrics such as precision, recall, and ROC-AUC. Previous research shows that machine learning algorithms can substantially enhance fraud detection accuracy by capturing complex, nonlinear transaction patterns (Dornadula and Geetha, 2019; Awoyemi et al., 2017).

The system includes a full data science pipeline consisting of data acquisition, preprocessing, exploratory data analysis, feature evaluation, model training, and performance comparison. By maintaining operational efficiency, the target is to develop a scalable analytical framework that improves fraud detection precision. The project emphasises the detection of fraudulent activities efficiently, reducing the losses and promoting trust in digital payment systems.

I. Introduction/Background

Digital financial services are rapidly expanding globally. Consumers are making billions of transactions globally with credit or debit card payments from online banking; mobile payment systems have become an essential foundation of the financial ecosystem nowadays. While these systems provide convenience and efficiency for daily tasks, they have also raised severe vulnerability of the financial systems. Fraudulent activities have increased and have become a global challenge, as billions of dollars have been lost or stolen from such fraudulent activities. As the world modernises and becomes cashless and card transactions increase rapidly, detecting fraudulent transactions has become highly important for consumers and financial institutions.

The common fraud detection system works on rule-based approaches. Predefined rules and thresholds are used to flag fraudulent transactions. Although these systems can identify known fraud patterns, they often fail to identify the evolving fraud strategies used by scammers. Fraudulent people continuously modify their techniques to bypass these traditional detection systems. Additionally, the traditional detection system identifies many real transactions as fraud, making high false positives and incorrectly labelling legitimate transactions as fraud. These, however, lead to tremendous dissatisfaction among consumers and increased operational costs for financial institutions.

As data science advances, machine learning models have created vast opportunities in this field to resolve these issues. Machine learning models can analyse large volumes of transactional data and learn complex transaction patterns to diagnose and identify fraudulent behaviour. Unlike traditional approaches,

machine learning algorithms can capture nonlinear relationships between variables. These can improve the fraud detection performance as more data is fed into training. These unique capabilities make the machine learning approach more suitable for fraud detection tasks.

In this project, we develop a prototype of a machine learning-based credit card fraud detection system using a publicly available dataset. The project demonstrates a complete data science pipeline, including data preprocessing, exploratory data analysis (EDA), feature analysis, model development, model evaluation, and deployment.

In this project, we tested two machine learning models – logistic regression and random forest. We used the dataset to evaluate these two models by applying the transactional dataset to these models. These models were then evaluated using precision, recall, and ROC-AUC in order to understand their effectiveness in identifying fraudulent transactions.

Throughout the project, multiple challenges were also discussed regarding class imbalance for the nature of transaction datasets, as there were very, very few transactions that were fraudulent to train the overall data on. Therefore, special attention was paid to evaluating the metrics that identified fraud detection performance, like recall for fraudulent transactions and precision for minimizing false alerts.

Overall, this project represents how machine learning techniques can improve card fraud detection systems. By using advanced analytical methods and transactional data, the system aims to detect the fraud transaction probability for consumers and financial institutions.

II. Problem Statement

The increasing growth of digital payment systems and online financial transactions has remarkably increased the risk of credit card fraud. Using the traditional monitoring system, financial institutions process a vast amount of transactions every day, which makes it difficult to identify fraudulent activities. As it is almost insignificant compared to total transactions, it creates an imbalanced dataset for which the accurate detection is challenging. This sometimes causes legitimate transactions to be fraudulent.

Previously, fraud detection systems relied on predefined patterns or thresholds to detect suspicious activity. This causes the system to detect only the known fraud patterns; as a result, the new strategies get past the system. Moreover, this system causes inconvenience for customers and increases the cost for financial institutions.

There's a clear need for a data-driven approach to catching fraud as it happens, especially when dealing with large volumes of card transactions. This project tackles that challenge by building a machine learning model that can separate genuine purchases from fraudulent ones with high accuracy while also boosting overall detection performance and minimizing false alerts.

III. Literature Review

In a published journal, Dornadula and Geetha (2019) proposed a card fraud detection system that had clusters of cardholders dataset of transactions and their amounts, which they applied a sliding window technique to extract their transaction behavior from the streaming data. Afterwards, they implemented multiple machine learning models upon the dataset using SMOTE to handle class imbalance and applied machine learning models like, logistic regression, decision tree and random forest and found out random forest achieved 99.98% accuracy and high precision and recall. [1]

Another article by Awoyemi et al. (2017) conducted a similar study using Naïve Bayes, KNN and Logistic Regression using a dataset of 284,807 transactions of European Cardholders. They used a hybrid approach that combined under sampling and oversampling to address the class imbalance of the dataset. Their results showed that, KNN achieved the best with 97.9% accuracy in detecting card fraud. [2]

A real-time fraud detection system was created by Thennakoon et al. (2019) using Logistic Regression, Naïve Bayes, KNN and Support Vector Machine. Their approach used PCA-based feature reduction, resampling and 10-fold cross validation. Finally, they achieved and accuracy between 74% and 91% for real-time fraud detection system. [3]

Varmedja et al. (2019) executed a study that compared multiple ML models including Logistic Regression, Random Forest, Naïve Bayes and Multilayer Perceptron on Credit Card fraud dataset from Kaggle. Their methodology included preprocessing, feature engineering and training the European Cardholder dataset using ensemble models. Their models were evaluated using precision, recall, F-1 score, ROC-AUC, proving that, their Random Forest model improved fraud detection accuracy more than other model achieving an accuracy of 99.96%. [4]

An article ensemble machine learning model for the detection of fraud transactions by the application of multiple classifiers to boost the performance of the model on unbalanced datasets of transactions was proposed by Khalid et al. (2024). The methodology for the proposed model included preprocessing the data, creating features, and training the ensemble model with precision, recall, F1-score, and ROC-AUC evaluation. Their results showed that an ensemble approach can significantly improve card fraud detection accuracy compared to single models. [5]

IV. Methodology / Proposed Approach

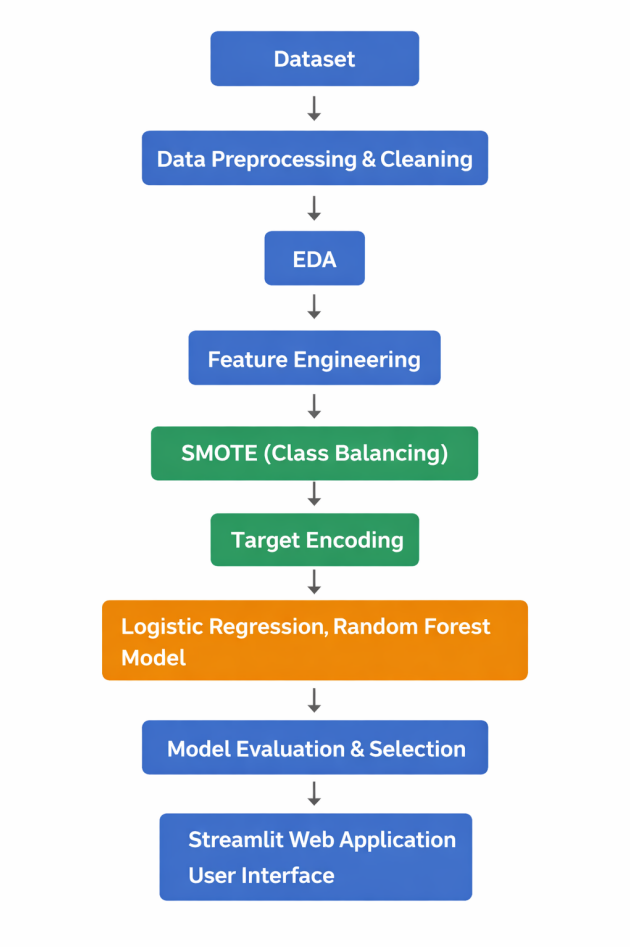

Fig. 1. Methodology / Proposed Approach [6]

Our project follows a structured machine learning pipeline with an aim to detect fraudulent card transaction probability. Our architecture integrates data pre-processing, exploratory data analysis (EDA), feature engineering, class balancing (SMOTE) technique, model development, model evaluation and selection, finally model deployment using an interactive web application.

A. Data Collection and Overview

Our project begins with sourcing the dataset from Kaggle Kaggle dataset repository. It is a publicly available credit card transaction dataset of USA across 52 states. The dataset includes 555,719 Entries of Transaction, in which 2145 Transactions are classified fraud. The dataset contains 13 variable features including target variable (is_fraud=1)

The dataset has only 0.38% of its data, being fraud making it highly imbalanced data which is normal for such dataset as very few transactions are flagged fraudulent from geuine transaction. This extreme imbalance is a common characteristic of fraud detection datasets and presents a significant challenge for machine learning models.

B. Data Preprocessing

The data set was evaluated and understood before pre-processing. In the pre-processing phase, the structure of the data was analysed, and potential missing values were looked for, verifying feature types and confirming the distribution of the target variable.

The dataset was notable, as it was already heavily preprocessed before publication. Due to prior annotation and preprocessing, the data set had no missing values and did not require additional cleaning. Therefore, no additional data cleaning was necessary for us to proceed with our project requirements. As transaction patterns have a wide range of variables, capping out the outliers would create a bias in our project outcome. Therefore, instead of aggressive cleaning, the pre-processing process focused on verifying the data quality and ensuring that the features were ready for project implementation. Notably, the **date of birth of the consumer column was converted to their age** to identify and relate features with the target variable. Additionally, the categorical variables were converted to numerical value variables via target encoding in order to train every suitable feature of the dataset that influenced the categorisation of fraudulent transactions.

C. Exploratory Data Analysis (EDA)

We conducted exploratory data analysis on the dataset in order to understand the dataset further before implementing the project. This EDA gives us an overview of the dataset and helps us guide the project more accurately and professionally. Initial analysis involved examining the class distribution of the dataset.

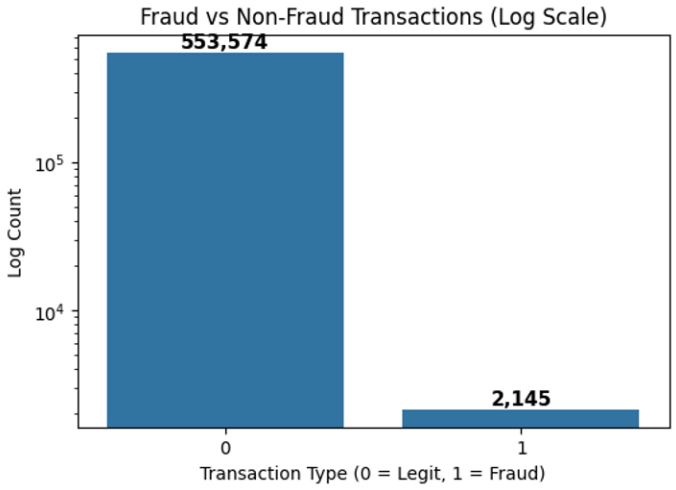

Fig. 2. Class Distribution of the Dataset

This confirms the highly imbalanced class distribution and helps us navigate the use of SMOTE (class imbalance handling technique) to be implemented to create a better-performing machine learning model.

The top 10 transaction categories with the highest fraud rates in the dataset show that online shopping has the highest fraud rate, followed by the category of transactions. This indicates high correlation with fraudulent transactions that an online transactions tend to be more vulnerable to fraudulent activity compared to other categories.

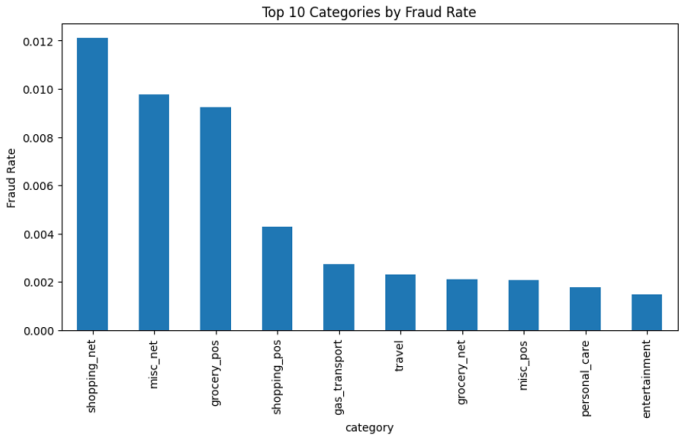

Fig. 3. Fraud Rate by Category

The following EDA shows 10 states with the highest fraud rates in the dataset. Alaska (AK) has the highest fraud rate, followed by states like Connecticut (CT) and Idaho (ID). This represent that state and ZIP code are high cardinality features where fraud occurrences occur. Fraud is not evenly distributed geographically, suggesting that regional transaction patterns may influence fraud risk.

Fig. 4. Fraud Rate by Location

One of the most important EDA, average transaction amount between legitimate and fraudulent transactions, shows that fraudulent transactions have a much higher average amount compared to normal transactions. Meaning fraud cases often involve larger transaction values, making the transaction amount an important indicator for detecting suspicious activity.

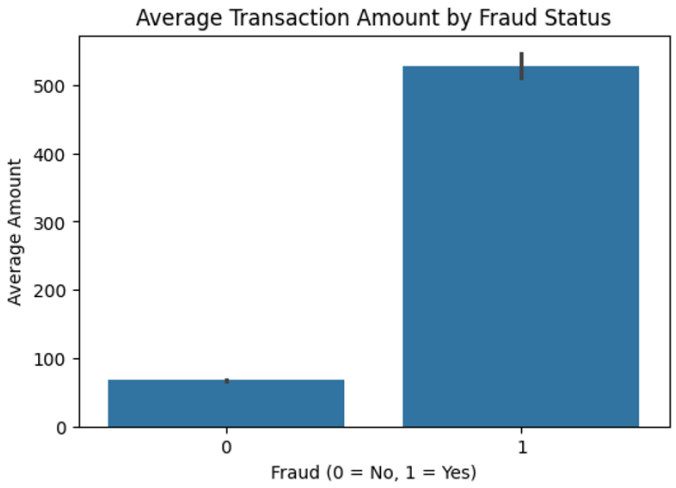

Fig. 5. Transaction Amount Distribution

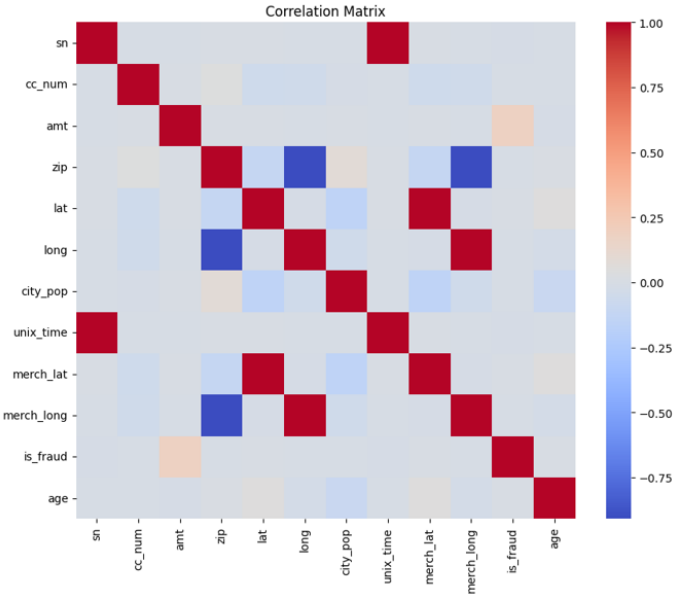

Fig. 6. Correlation Matrix

The correlation heatmap shows the relationship between different transaction features in the dataset. Location-related variables such as latitude, longitude, and merchant location show higher correlations because they represent related geographic information.

D. Feature Engineering

We used feature engineering to identify the best variables for the model. In the exploratory data analysis phase, various variables like transaction category, amount, merchant name, location, and zip code were analyzed to identify the pattern of fraudulent transactions.

Based on the exploratory data analysis (EDA) and correlation matrix, various variables that show a significant relationship with fraudulent transactions. Other variables that show little or no relationship with fraudulent transactions were ignored to keep the model efficient.

Some of the variables chosen for the model were categorical variables. The categorical values needed to be converted to numerical variables for the machine learning model. We implemented, **target encoding**.

In target encoding, the categorical variables are replaced by the average fraud rate for the given category. This method is and efficient approach that creates a clean dataset and has an effective impact for model training by representing its avarage fraud value. Moreover, the using target encoding prevents making many columns in the dataset.

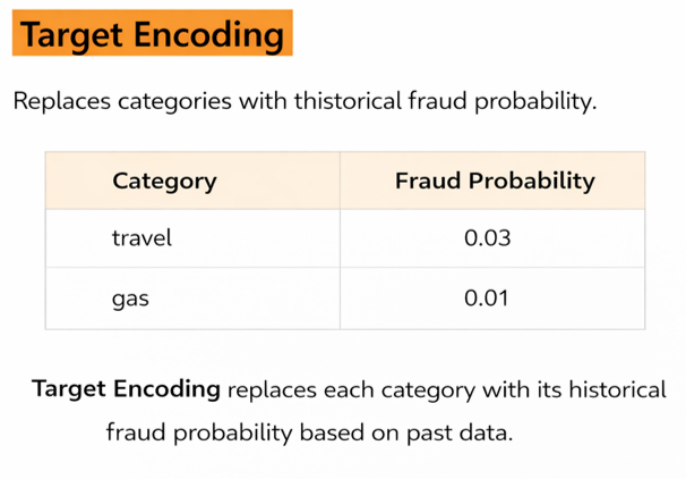

Fig. 7. Target Encoding

E. Machine Learning Model Implementation

The raw dataset after feature selection was tested initially with two machine learning models. **Logistic Regression** and **Random Forest** model. Our aim was to compare the performance of both ML models on raw data before applying any techniques like class balancing. We focused on evaluation metric like **precision, recall, ROC-AUC** more than accuracy.

1) Logistic Regression: Logistic Regression was implemented. When trained on the original dataset, the model achieved a **fraud precision of 0.0235** and a **fraud recall of 0.5758**. The results indicate that the model was able to identify a reasonable proportion of fraudulent transactions, but with extremely low precision.

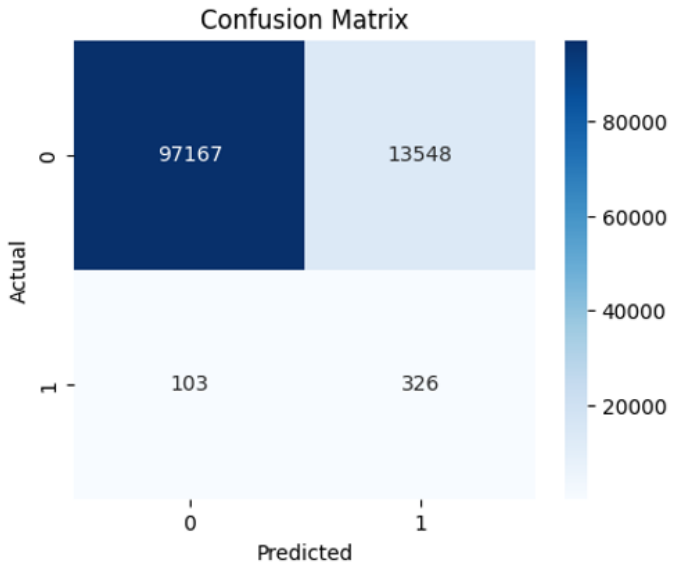

Fig. 8. Confusion Matrix Logistic Regression

Logistic Regression correctly detected 326 fraud cases (Recall ≈ 0.58) but produced 13,548 false positives, resulting in very low precision (≈ 0.0235).

2) Random Forest: The Random Forest model demonstrated significantly stronger performance when tested on same raw dataset. Compared to Logistic Regression, the model achieved a **fraud precision of 0.9693** and a **fraud recall of 0.5152**. High precision indicates that when the model predicts a transaction as fraudulent, it is very likely to be correct.

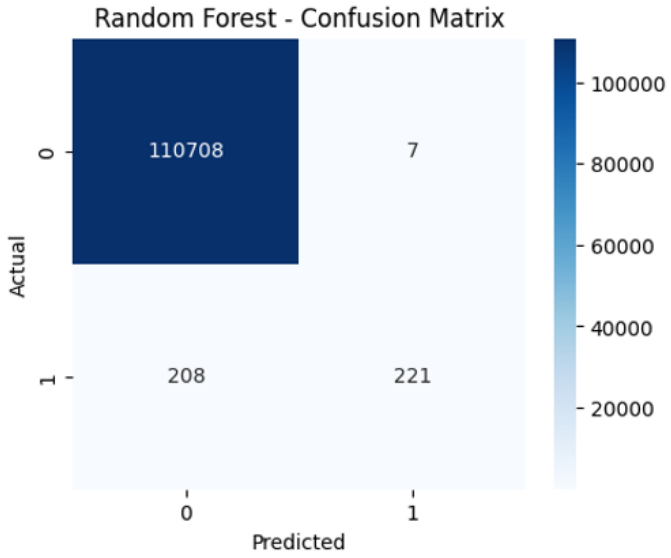

Fig. 9. Confusion Matrix Random Forest

Random Forest correctly detected 221 fraud cases (Recall≈ 0.52) while producing only 7 false positives, achieving very high precision (≈ 0.97) and more reliable fraud predictions.

F. Model Evaluation and Selection

TABLE I
MACHINE LEARNING MODEL EVALUATION COMPARISON

Machine Learning Models	Fraud Precision	Fraud Recall	ROC-AUC
Logistic Regression	0.0235	0.5758	0.9089
Random Forest	0.9693	0.5152	0.9962

From the above evaluation result, the Random Forest model was chosen as the final model for the fraud detection system. Although the model based on the Logistic Regression approach has a good recall rate, but very poor precision rate resulted in a very high false positive rate. However, the Random Forest model has a good balance between the precision rate and the recall rate, along with the high ROC-AUC score.

The high precision rate of the Random Forest model makes it more appropriate for the fraud detection system because accusing the customer of fraud with a legitimate transaction might harm the customer.

G. Streamlit Web Application Implementation

1) Handling Class Imbalance for Final Model Training: Before the deployment of the final model for web application, we implemented Synthetic Minority Oversampling Technique (SMOTE) in order to handle the class imbalance within dataset. It is because, if the model was trained on raw dataset, the trained model might be biased, as very few number of data was classified as fraud within the training dataset.

```
Before SMOTE:
is_fraud
0    442859
1      1716
Name: count, dtype: int64

After SMOTE:
is_fraud
0    442859
1     22142
Name: count, dtype: int64
```

Fig. 10. SMOTE (Handling Class Imbalance)

Using SMOTE, the training data was adjusted 95% legitimate transaction to have at least 5% fraudulent transactions. This oversampling method generated synthetic data of fraudulent transactions based on existing patterns of fraud, enabling the model to better learn the patterns of fraud transaction.

2) Deployment: Web Application: To demonstrate the practical application of the fraud detection system, we deployed the model using a web based application the Streamlit library. After assessing the performance of the trained final trained random forest model after including SMOTE the final fraud detection model was saved fraud_web_model_v3.pkl.

The saved model file was loaded into Streamlit using Python. It served as the main prediction engine of the application. When a user enters transaction-related inputs through the web interface, the application processes the values, formats them to match the structure expected by the trained model, and then passes them into fraud_web_model_v3.pkl to generate a prediction.

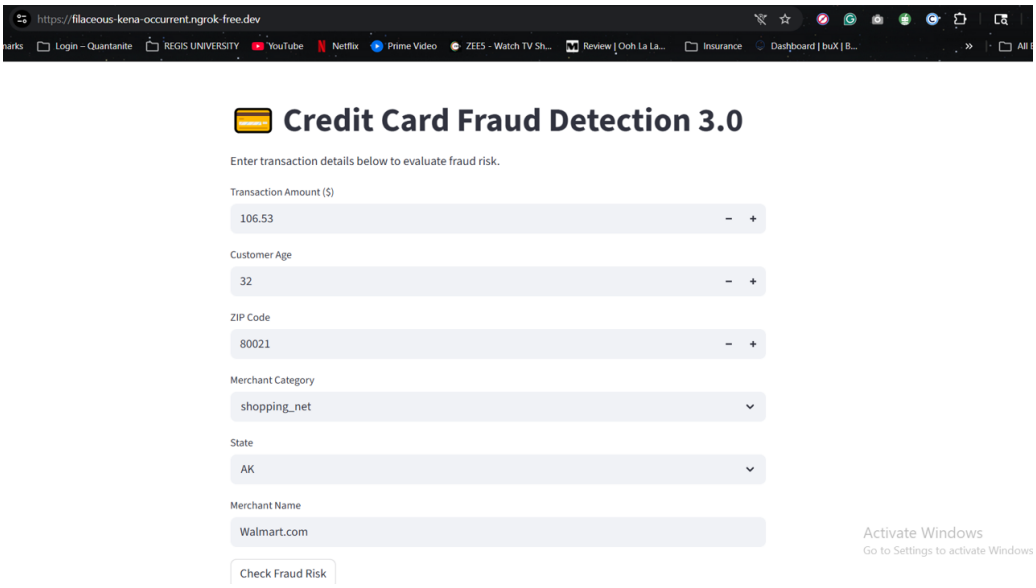

Fig. 11. Web Application User Interface

This demonstrates the user interface of the Web Application design, where users can input data of their transaction to verify whether their transaction would be fraudulent or not.

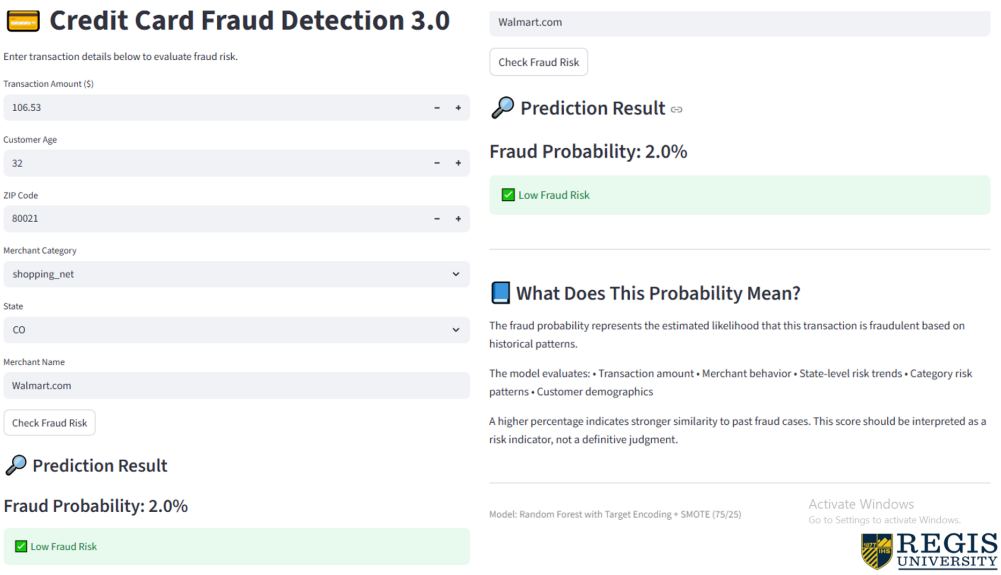

Fig. 12. Web Application UI (Low Fraud Risk)

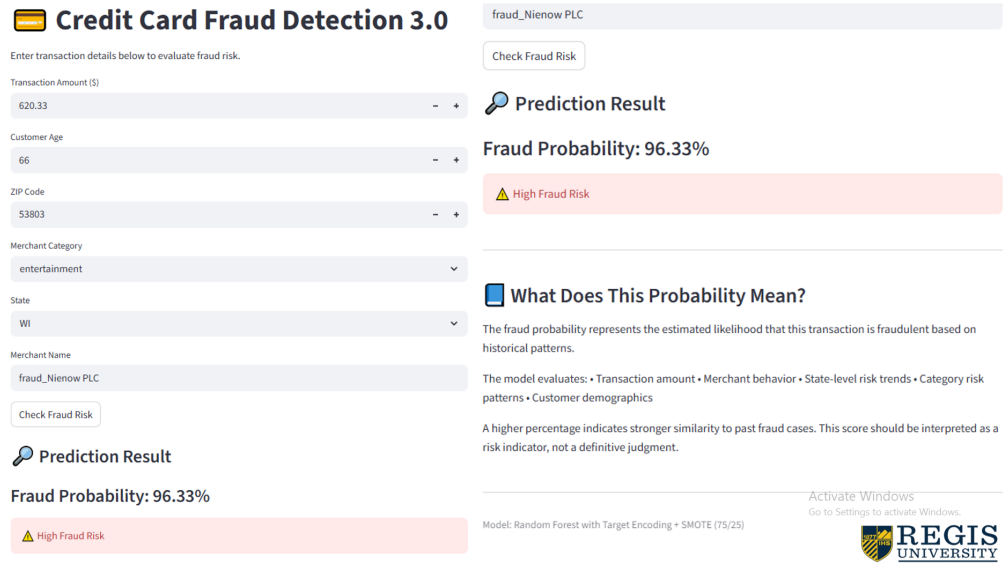

Fig. 13. Web Application UI (High Fraud Risk)

V. Expected Outcomes

The primary expected outcome of this project is the development of an efficient machine learning-based system that has the potential to use the transactional data to identify between fraudulent and genuine transactions. The primary objective of this study is to identify patterns that are fraudulent transaction. At the same time illustrate the potential that machine learning has for the purpose of improving fraud detection significantly compared to conventional methods through the use of supervised learning algorithms such as Random Forest and Logistic Regression.

VI. Timeline

The project was conducted over eight weeks from January 15, 2026 to March 05, 2026. The initial phase focused on data collection, understanding the dataset, and planning the project structure. This was

followed by data preprocessing and exploratory analysis to understand the characteristics of the dataset. In the middle phase of the project, machine learning models were implemented and evaluated, and feedback was incorporated to further refine the approach. The later stages focused on expanding the project by developing the Streamlit web application, summarising the findings, and preparing the final presentation and report.

TABLE II
PROJECT TIMELINE

Week	Project Activities
Week 1	Project initiation, dataset collection, and sourcing data collection from Kaggle.
Week 2	Dataset Selection, Project planning, defining project structure, research on fraud detection methods, and literature review preparation.
Week 3	Data understanding, data preprocessing, dataset inspection, and exploratory data analysis (EDA).
Week 4	Implementation of machine learning models including Logistic Regression and Random Forest.
Week 5	Model evaluation, comparison of results, and selection of the most suitable model.
Week 6	Application of SMOTE for class balancing and preparation of the final model for deployment.
Week 7	Development of the Streamlit web application and integration of the trained model into the interface.
Week 8	Final project execution, preparation of presentation slides, and writing of the final practicum report.

VII. CONCLUSION

With the help of the available public data set, the project has successfully implemented a prototype system of credit card fraud detection using machine learning algorithms. It is important to note that the data science pipeline involves collecting data, preprocessing the data, exploratory data analysis, feature development, and developing the classification model.

In the proposed project, the performance of the machine learning algorithms, mainly Random Forest and Logistic Regression, is evaluated to detect credit card fraud. Based on the results obtained from the evaluation process, it is determined that the performance of the Random Forest model is significantly better compared to the performance of the Logistic Regression model, with the accuracy being almost perfect in nature, to prove the efficiency of the detection process of patterns related to credit card fraud.

VIII. FUTURE SCOPE

There are still a number of opportunities to improve and develop the fraud detection system in the future. For example, investigating more advanced machine learning techniques, such as Gradient Boosting, XGBoost, and Deep Learning, could be a potential direction to improve the performance of the system in detecting more complex forms of fraud.

In addition, increasing the size and variety of data sets from different financial institutions could also improve the robustness of the system. Another potential direction to improve the system is to integrate it with real-time transaction monitoring systems, through which fraud detection occurs in real-time as transactions are being performed. Hence, it could be more useful in real-world applications.

REFERENCES

[1] Dornadula, V. N., and 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

[2] 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

[3] Real-time credit card fraud detection using machine learning. (2019, January 1). IEEE Conference Publication — IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/8776942
[4] Credit Card Fraud Detection - Machine Learning methods. (2019, March 1). IEEE Conference Publication — IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/8717766
[5] Khalid, A. R., Owoh, N., Uthmani, O., Ashawa, M., Osamor, J., and Adejoh, J. (2024). Enhancing Credit Card Fraud Detection: An Ensemble Machine Learning Approach. Big Data and Cognitive Computing, 8(1), 6. https://doi.org/10.3390/bdcc8010006
[6] ChatGPT. (2026). ChatGPT (March 5 version) [Large language model]. OpenAI. https://chat.openai.com/ChatGPT (OpenAI, 2026) was used for generating illustrative images for the methodology.
[7] Credit Card Fraud Detection Platform Market Report, 2033. (n.d.). https://www.grandviewresearch.com/industry-analysis/credit-card-fraud-detection-platform-market-report
[8] Credit card fraud detection using machine learning. (2020, May 1). IEEE Conference Publication — IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9121114
[9] Tiwari, P., Mehta, S., Sakhuja, N., Kumar, J., and Singh, A. K. (2021, August 23). Credit Card Fraud Detection using Machine Learning: A Study. arXiv.org. https://arxiv.org/abs/2108.10005
[10] H. Najadat, O. Altiti, A. A. Aqouleh and M. Younes, ”Credit Card Fraud Detection Based on Machine and Deep Learning,” 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2020, pp. 204-208, doi: 10.1109/ICICS49469.2020.239524.
[11] OpenAI. (2026). ChatGPT (GPT-5.3) [Large language model]. Used as an assisting tool for outlining the structure of the project report. https://chat.openai.com