Card Fraud Detection using Machine Learning model

Nahian Arnob

Regis University

MSDS 692: DATA SCIENCE PRACTICUM

Submitted to: Dr. Ghulam Mujtaba

REGIS UNIVERSITY

ABSTRACT

- The Project is Card Fraud Detection System using Machine Learning model is used to identify fraudulent transaction that have been made before and help prevent future fraudulent transactions.
- Based on specific features, the ML model analyzes and gives you a potential feedback, weather or not the transaction you will be doing is fraud or not.
- The model was trained on big dataset of credit card transactions from **Kaggle**. features such as transaction amount, merchant name, state, category, and customer demographics influence Card Fraud Detection Model.

Problem statement

- Card Fraud is a **growing financial and cybersecurity problem** affecting every individual and financial institution worldwide.
- Fraud Transaction are **very difficult to detect** because they often resemble real customer behavior.
- Transaction **datasets are highly imbalanced**, because fraud cases represent only 1% or less of total transaction.
- High-cardinality features such as **merchant names and geographic locations** make fraud detection more complex.
- An intelligent system that can **identify suspicious transactions accurately while minimizing false positives.**
- An interactive tool is required to allow users or analysts to **input transaction details and quickly assess the probability of fraud** using machine learning models and AI

Project Objectives

- Identify patterns associated with fraudulent transactions.
- Build a machine learning model to detect fraud.
- Handle class imbalance in the dataset.
- Evaluate model performance using appropriate metrics.
- Develop a user-friendly interface to evaluate fraud risk.

This Photo by Unknown Author is licensed under

REGIS UNIVERSITY

DATASET OVERVIEW

- Source: Kaggle (https://www.kaggle.com/datasets/chetanmittal033/credit-card-fraud-data/data)
- **555,719** Entries of Transac
- **13 Features**
  - serial number
  - transaction date and t
  - merchant name
  - category of transactio
  - amount
  - first name of customer
  - last name customer
  - city
  - state
  - zip
  - Target variable (is_fraud)

```
df = pd.read_csv('/content/fraudTest.csv')
display(df.head())
```

sn	trans_date_trans_time	cc_num	merchant	category	amt	first	last	gender	s
0	21-06-2020 12:14	2.291160e+15	fraud_Kirlin and Sons	personal_care	2.86	Jeff	Elliott	M	D
1	21-06-2020 12:14	3.573030e+15	fraud_Sporer-Keebler	personal_care	29.84	Joanne	Williams	F	
2	21-06-2020 12:14	3.598220e+15	fraud_Swaniawski, Nitzsche and Welch	health_fitness	41.28	Ashley	Lopez	F	Val
3	21-06-2020 12:15	3.591920e+15	fraud_Haley Group	misc_pos	60.05	Brian	Williams	M	M
4	21-06-2020 12:15	3.526830e+15	fraud_Johnston-Casper	travel	3.19	Nathan	Massey	M	Ap

METHODOLOGY

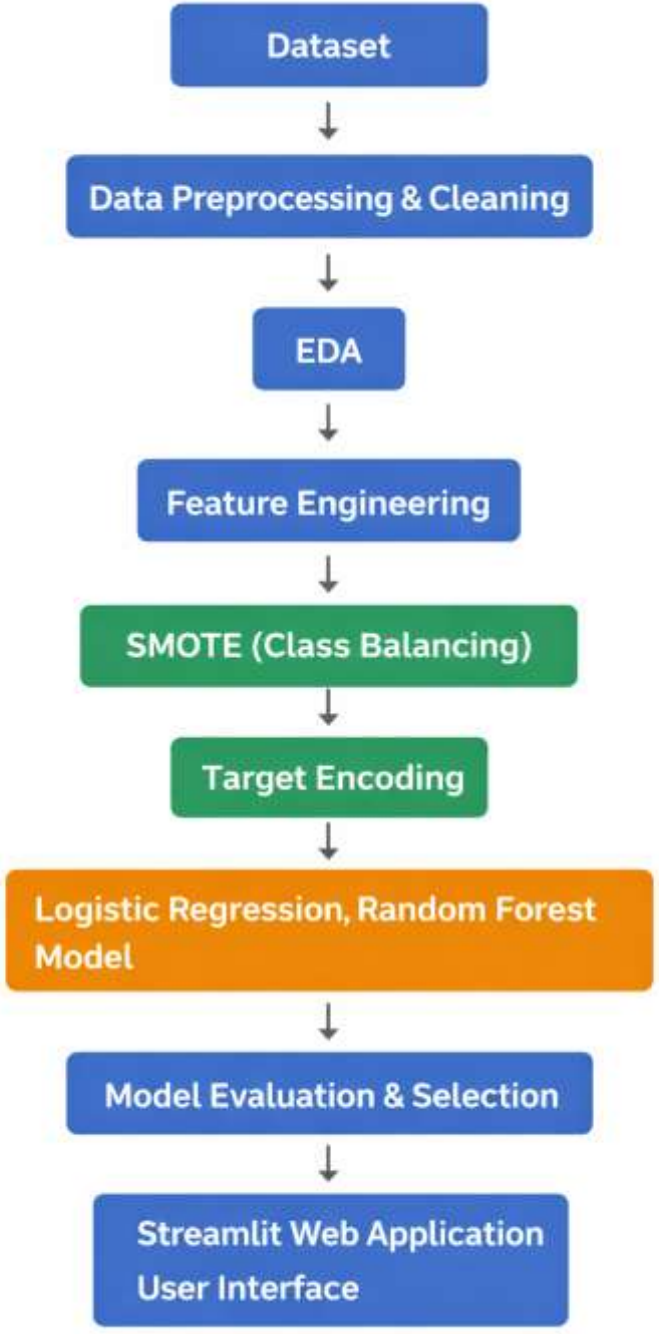
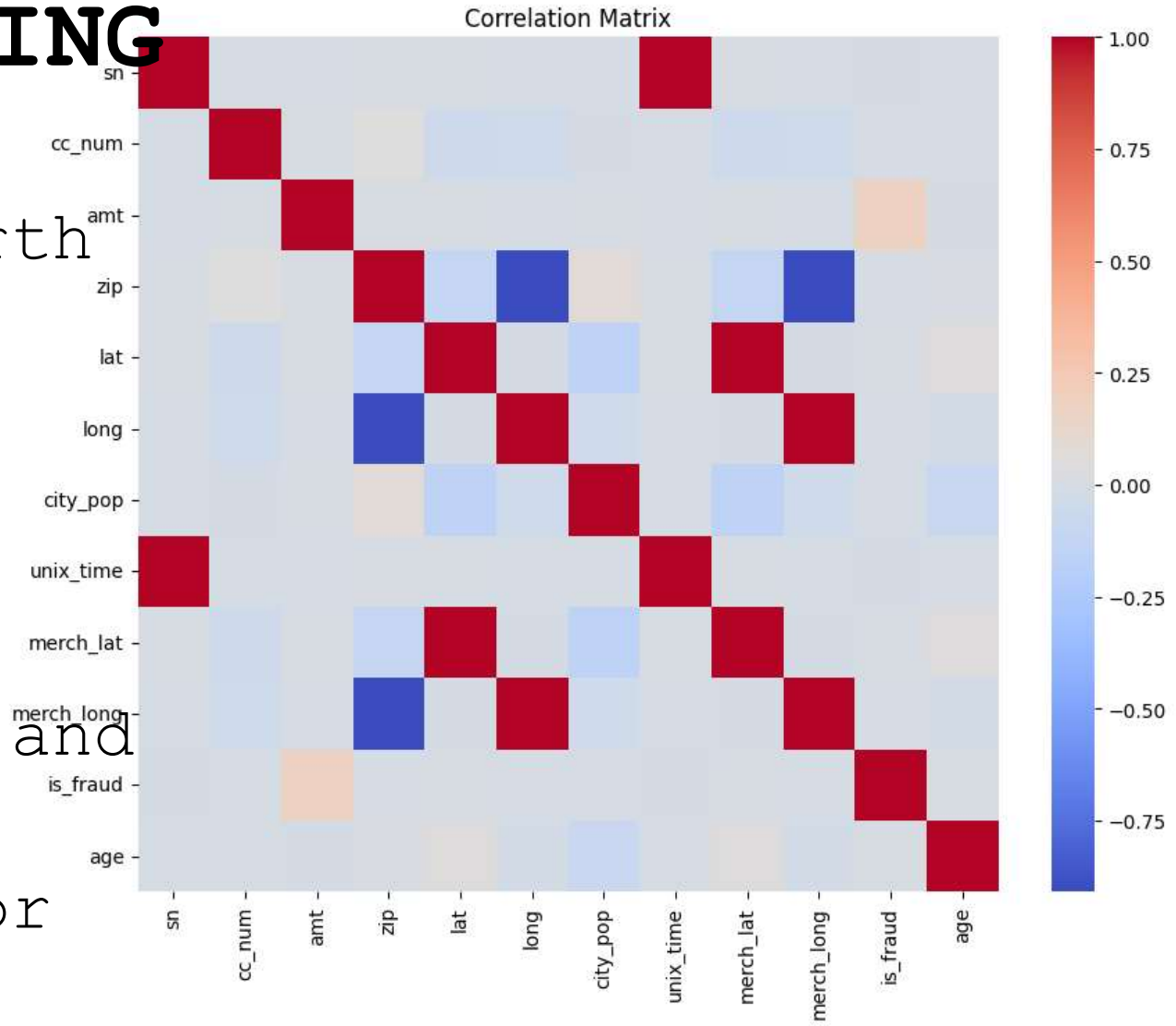


Image generated using ChatGPT (OpenAI, [illegible])

REGIS UNIVERSITY

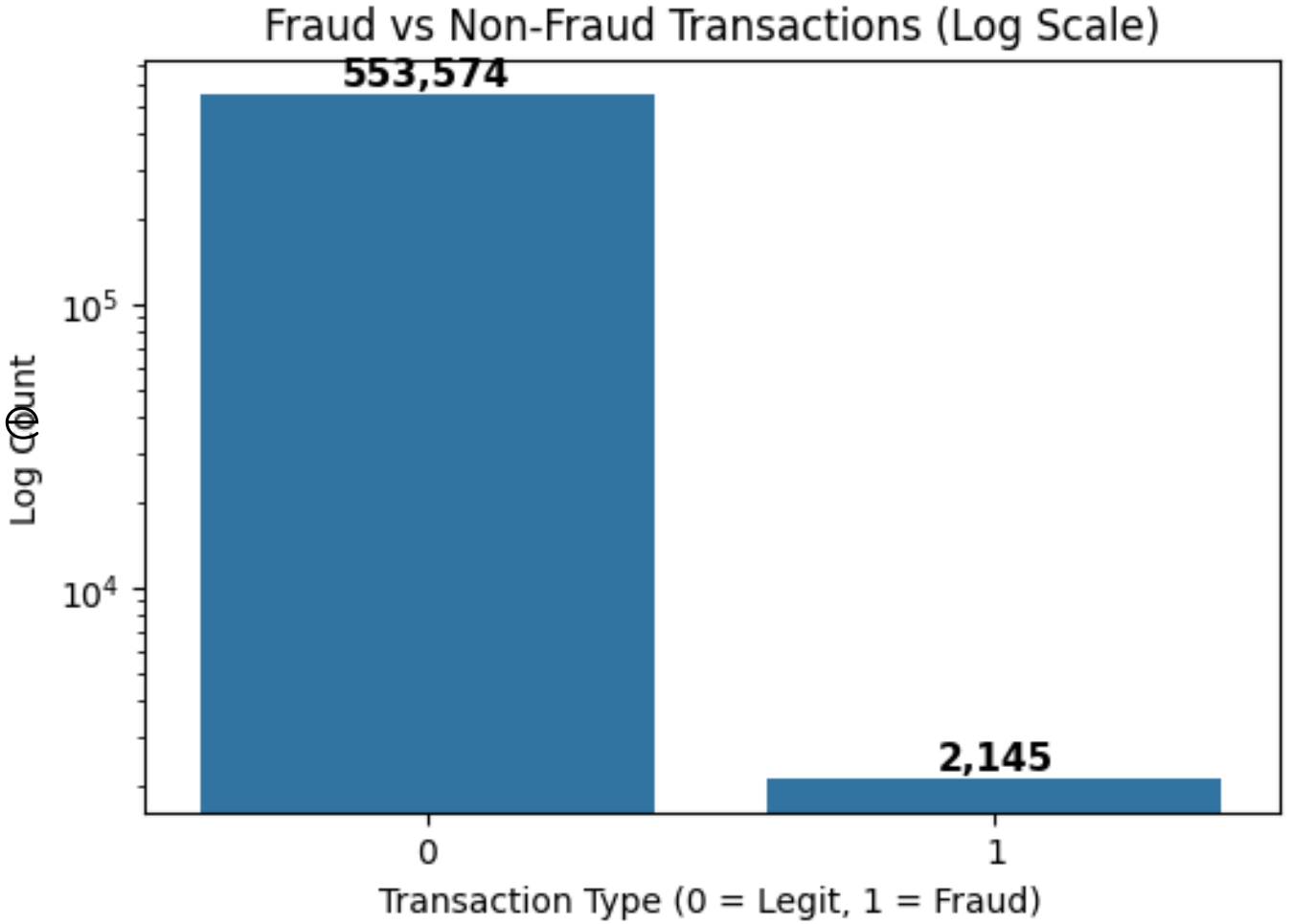
DATA PREPROCESSING

- Converted date of birth to **customer age**
- Selected important predictive features based on correlation matrix
- Prepared categorical and numerical features
- Structured dataset for machine learning pipeline



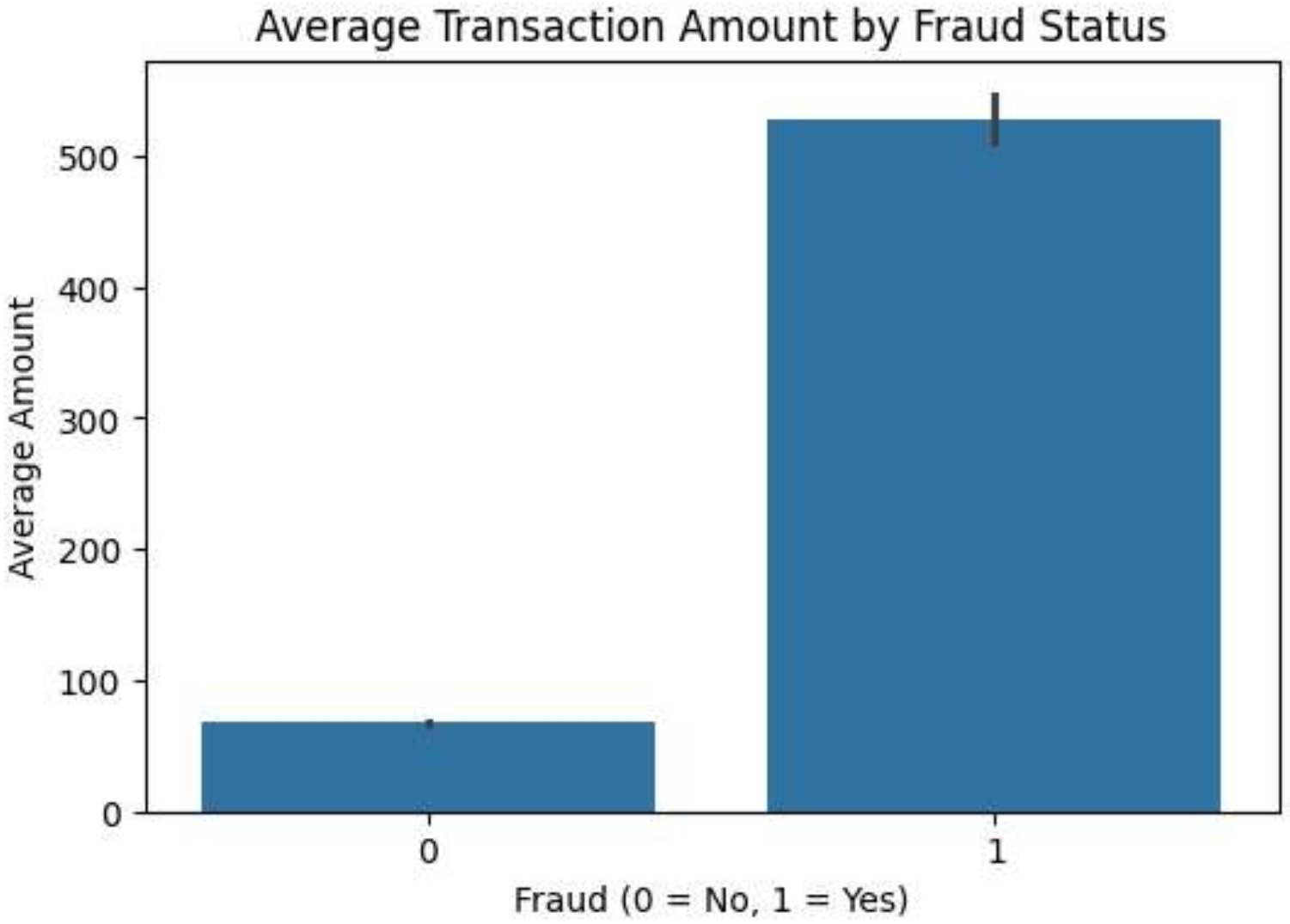
EDA (Exploratory Data Analysis)

- Fraud vs Non-Fraud count
- Fraud transactions represent a very small portion of the dataset.



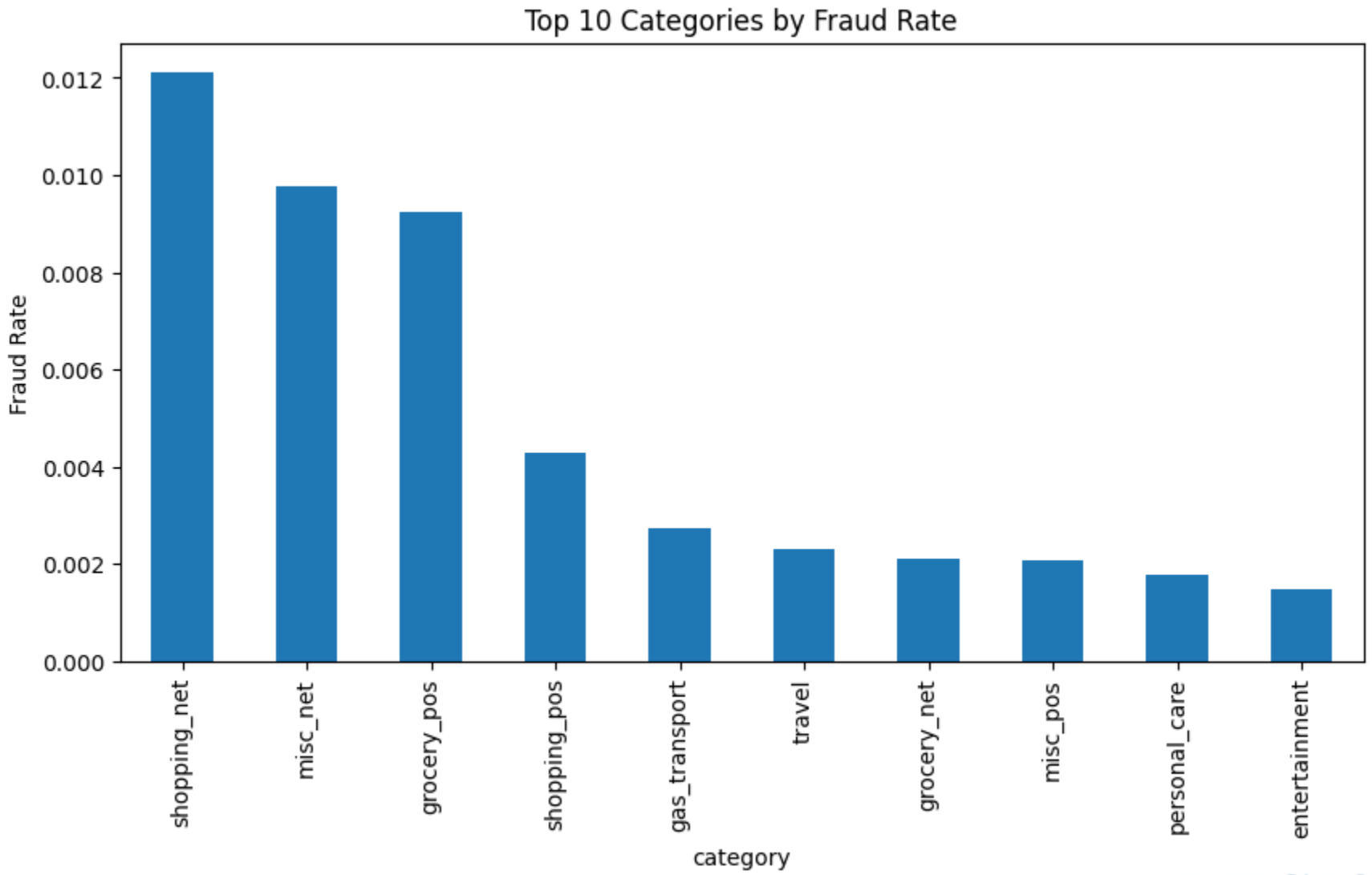
EDA (Average Transaction Amount)

- Transaction Amount
- Fraudulent transactions often occur in unusual spending patterns.
- Usual Average of Fraud Transaction is more than average



EDA (Top 10 Fraud Rate by Category)

- Shopping Category has been the top category in terms of Fraud Rate.

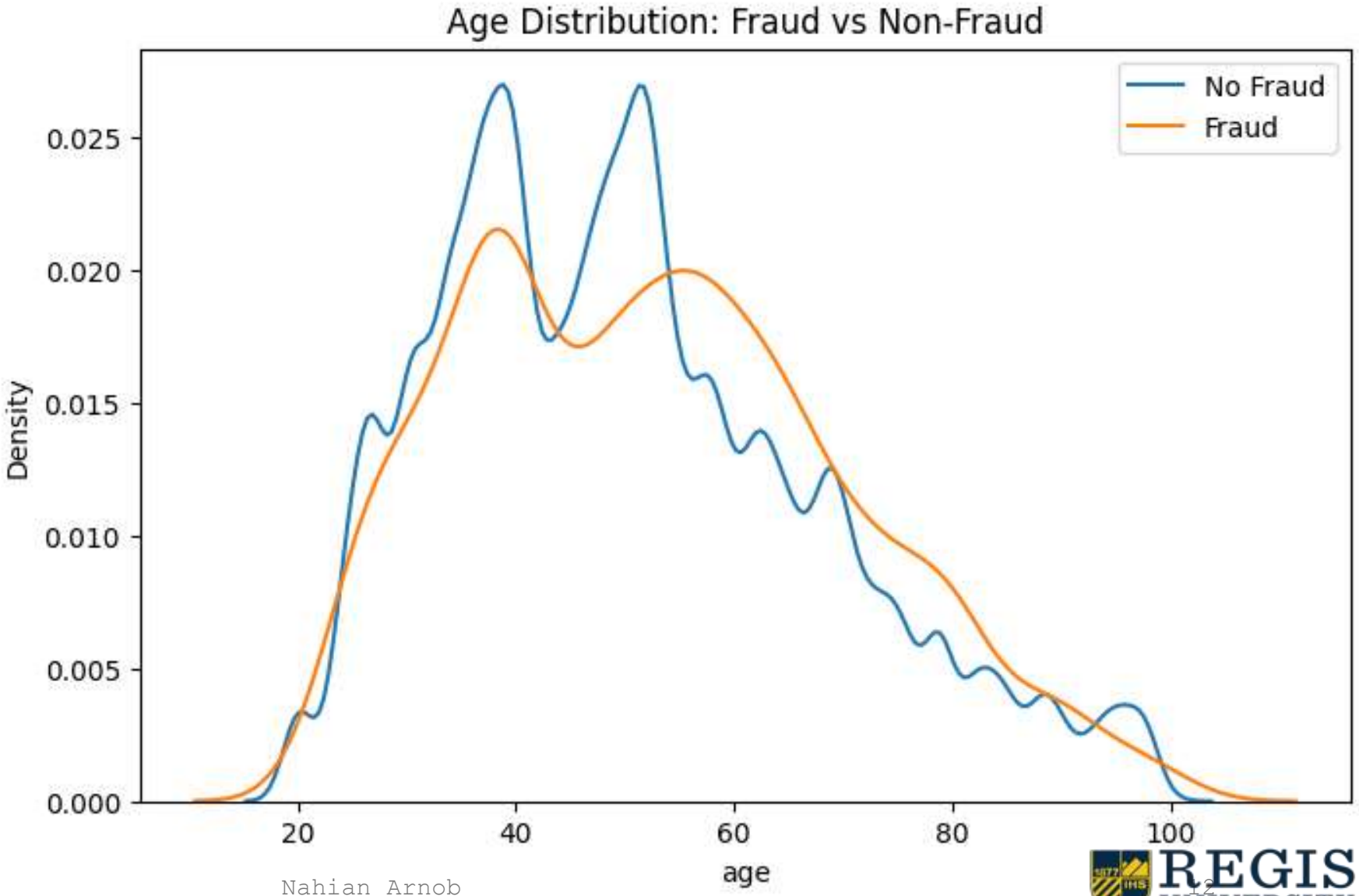


EDA (Top 10 Fraud Rate by State)

- AK (Alaska) has the highest fraud rate in USA

EDA (Age Curve)

- Fraud and non-fraud transactions occur between **30 and 60 years old.**
- The distributions largely overlap, meaning **age alone is not a strong indicator of fraud.**



Feature Engineering

- Features used for our Machine Learning model
  - Transaction Amount
  - Merchant Name
  - Merchant Category
  - State
  - Customer Age
  - Zip Code

Encoding Technique:

- Target Encoding for important categorical Variable

Target Encoding

Replaces categories with thistorical fraud probability.

Category	Fraud Probability
travel	0.03
gas	0.01

Target Encoding replaces each category with its historical fraud probability based on past data.

REGIS UNIVERSITY

Class Imbalance Handling (SMOTE)

- SMOTE (synthetic oversampling)
  We intend to apply SMOTE to train our final chosen Model for deployement, to train the model accurately so that, it can identify each Fraud Transaction accurately.

is_fraud	count
0	553574
1	2145

is_fraud	proportion
0	99.614014
1	0.385986

But, First, we will Evaluate our Models on real dataset to identify and select model which performs the best in our case

Machine Learning Models Test and Evaluation

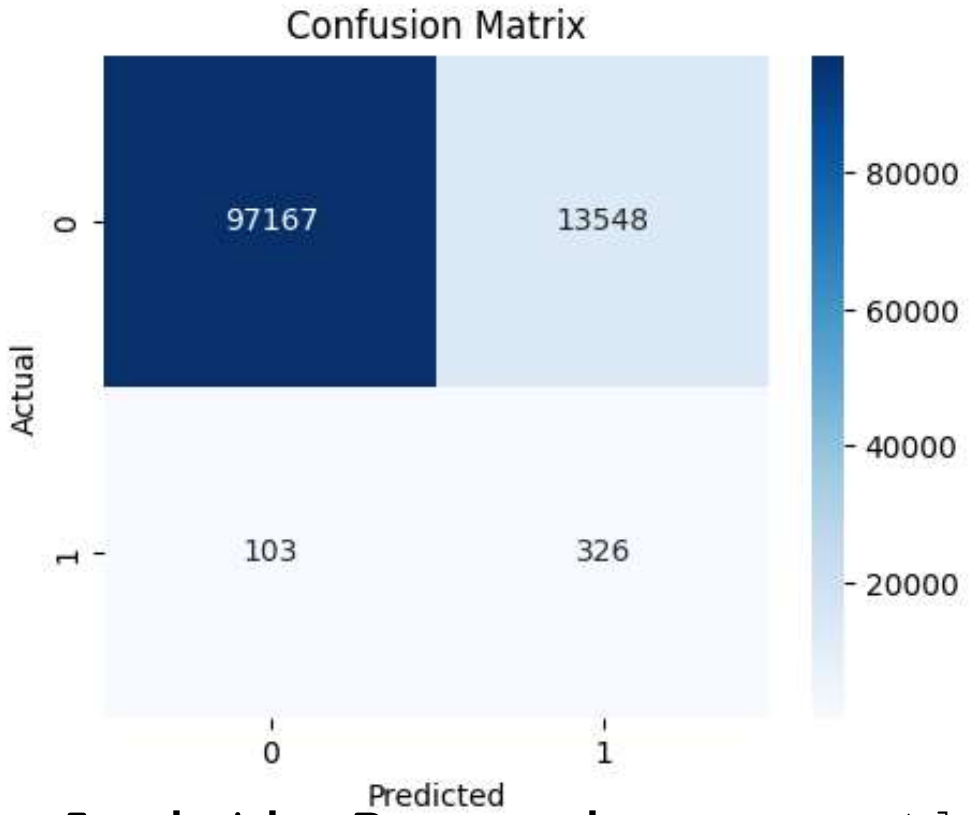
Models Tested

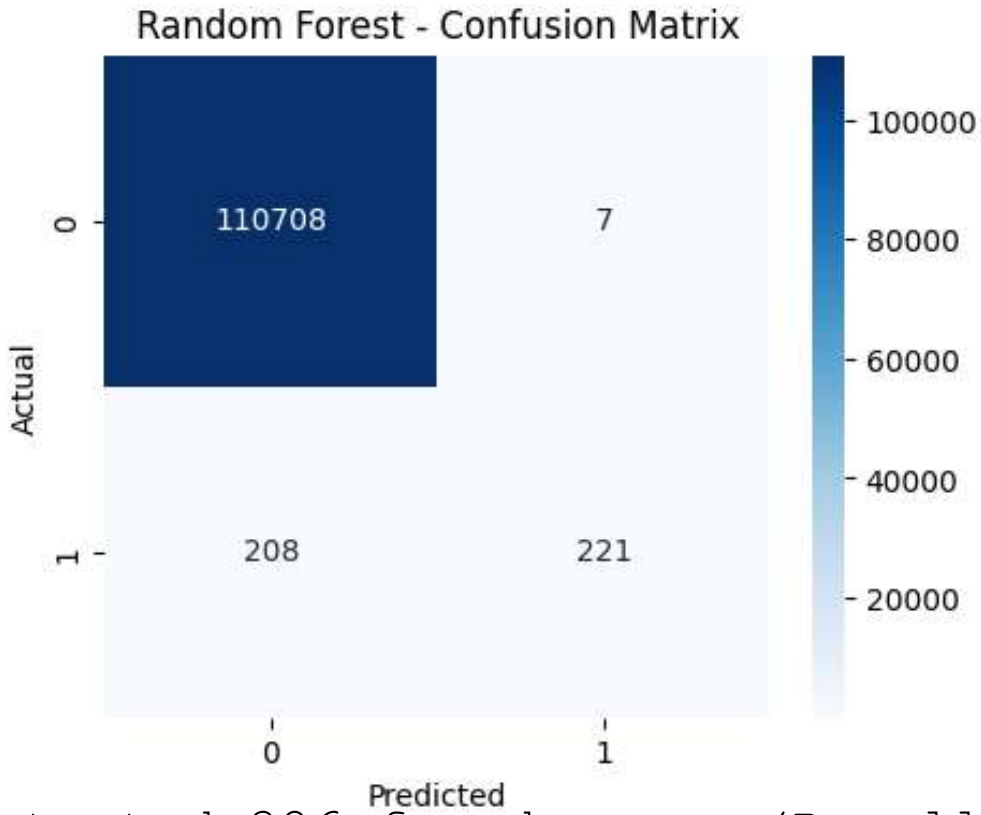
- Logistic Regression
- Random Forest

Model	Fraud Precision	Fraud Recall
Logistic Regression	0.0235	0.5758
Random Forest	0.9693	0.5152

- **Logistic Regression** achieves higher recall but very low precision, meaning it detects more fraud cases but produces many false positives.
- **Random Forest** gives high precision with slightly lower recall, making it better at accurately identifying fraudulent transactions with fewer false alarms.

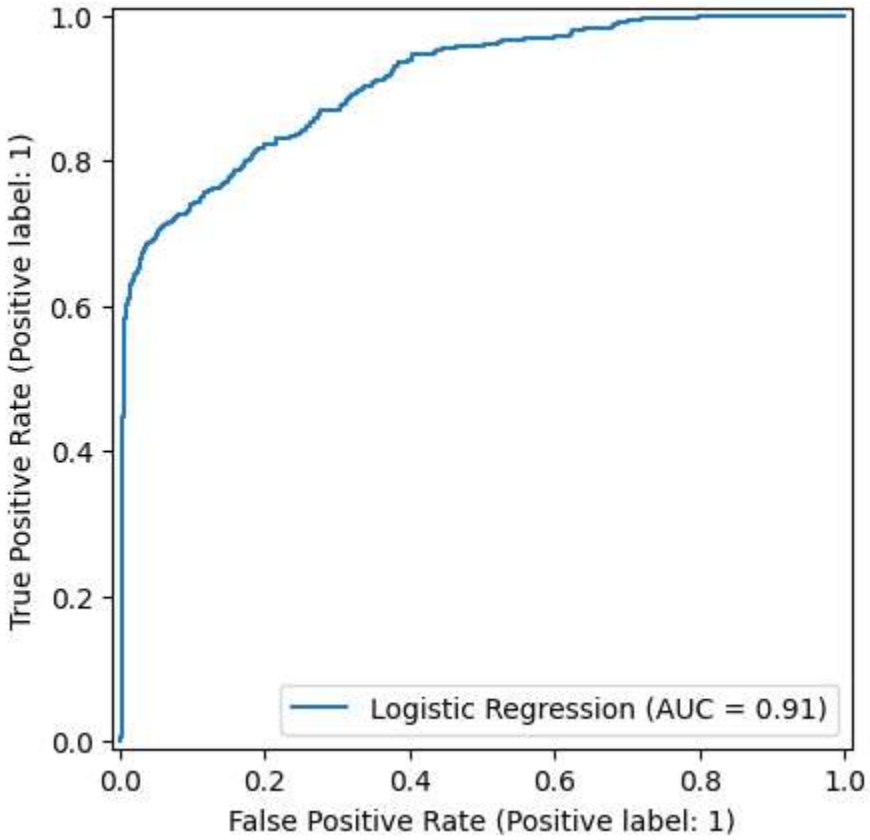
Confusion Matrix (Logistic Regression vs Random Forest)

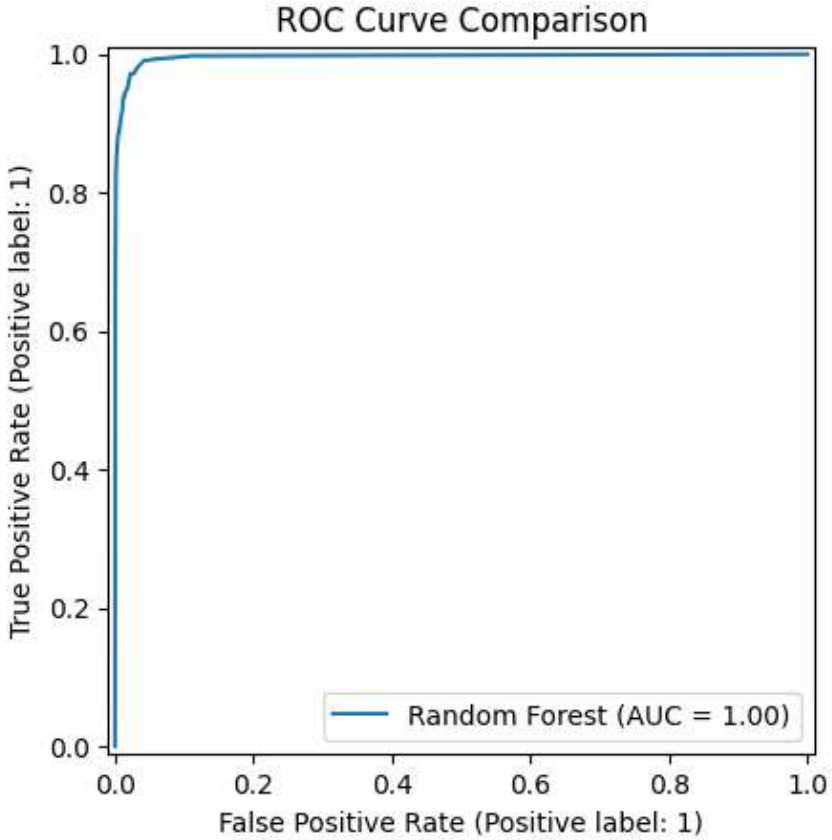




- **Logistic Regression** correctly detected 326 fraud cases (Recall ≈ 0.58) but produced 13,548 false positives, resulting in very low precision (≈ 0.0235).
- **Random Forest** correctly detected 221 fraud cases (Recall ≈ 0.52) while producing only 7 false positives, achieving very high precision (≈ 0.97) and more reliable fraud predictions.

ROC-AUC Compraison (Logistic Regression vs Random Forest)





- Random Forest (ROC-AUC = **0.9962**) significantly outperforms Logistic Regression (ROC-AUC = **0.9089**), indicating superior ability to capture complex patterns and distinguish fraudulent from legitimate transactions.

RANDOM FOREST MODEL

So, we finally choose to use RANDOM FOREST Model for our WEB Application Deployment.

Now We will apply SMOTE to train our Model Again for Web Application (User Interface Purposes)

Applying SMOTE (Synthetic Oversampling)

We applied SMOTE to Training Data only for using in our Final Deployment Model RANDOM FOREST.

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

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

- We used 95% to 5% Ratio, because Transaction Frauds are highly less prevalent in terms of Fraud Detection.
- Random Forest model achieved a fraud **recall of 72%** and **precision of 67%**. The balanced performance makes this model well-suited for deployment in the interactive web application.

REGIS UNIVERSITY

Web Application (User-Interface)

- Saved the Model (fraud_web_model_v3.pkl)
- Used Streamlit (an open-source Python framework for data scientists and AI/ML engineers) to deploy our model and test.

Web Application Features:

- User inputs transaction amount and details
- Select Category, State (Location of Transaction)'
- Enter Merchant Name
- System predicts Fraud Probability
- Display risk explanation

System Workflow Pipeline

User Input
↓
Target Encoding
↓
Random Forest Model
↓
Fraud Probability
↓
Risk Explanation

Image generated using ChatGPT (OpenAI, 2026).

https://filaceous-kena-occurrent.ngrok-free.dev

marks | Login – Quantanite | REGIS UNIVERSITY | YouTube | Netflix | Prime Video | ZEE5 - Watch TV Sh... | Review | Ooh La La... | Insurance | Dashboard | buX | B... | All B

Credit Card Fraud Detection 3.0

Enter transaction details below to evaluate fraud risk.

Transaction Amount ($)

106.53

Customer Age

32

ZIP Code

80021

Merchant Category

shopping_net

State

AK

Merchant Name

Walmart.com

Check Fraud Risk

Activate Windows
Go to Settings to activate Windows.

REGIS UNIVERSITY

💳 Credit Card Fraud Detection 3.0

Enter transaction details below to evaluate fraud risk.

Transaction Amount ($)

106.53 − +

Customer Age

32 − +

ZIP Code

80021 − +

Merchant Category

shopping_net

State

CO

Merchant Name

Walmart.com

Check Fraud Risk

🔎 Prediction Result

Fraud Probability: 2.0%

✅ Low Fraud Risk

Walmart.com

Check Fraud Risk

🔎 Prediction Result

Fraud Probability: 2.0%

✅ Low Fraud Risk

📘 What Does This Probability Mean?

The fraud probability represents the estimated likelihood that this transaction is fraudulent based on historical patterns.

The model evaluates: • Transaction amount • Merchant behavior • State-level risk trends • Category risk patterns • Customer demographics

A higher percentage indicates stronger similarity to past fraud cases. This score should be interpreted as a risk indicator, not a definitive judgment.

Model: Random Forest with Target Encoding + SMOTE (75/25)

REGIS UNIVERSITY

💳 Credit Card Fraud Detection 3.0

Enter transaction details below to evaluate fraud risk.

Transaction Amount ($)

620.33 − +

Customer Age

66 − +

ZIP Code

53803 − +

Merchant Category

entertainment

State

WI

Merchant Name

fraud_Nienow PLC

Check Fraud Risk

🔎 Prediction Result

Fraud Probability: 96.33%

⚠ High Fraud Risk

fraud_Nienow PLC

Check Fraud Risk

🔎 Prediction Result

Fraud Probability: 96.33%

⚠ High Fraud Risk

📘 What Does This Probability Mean?

The fraud probability represents the estimated likelihood that this transaction is fraudulent based on historical patterns.

The model evaluates: • Transaction amount • Merchant behavior • State-level risk trends • Category risk patterns • Customer demographics

A higher percentage indicates stronger similarity to past fraud cases. This score should be interpreted as a risk indicator, not a definitive judgment.

Model: Random Forest with Target Encoding + SMOTE (75/25)

REGIS UNIVERSITY

CONCLUSION

- Machine learning models can effectively detect fraudulent transactions
- Target Encoding improves handling of high-cardinality features
- SMOTE helps address class imbalance
- Random Forest provides strong predictive performance with high precision
- Web application demonstrates practical fraud detection system

REGIS UNIVERSITY

FUTURE SCOPE

Potential Improvements:

- Create a real-time card fraud detection system using global dataset
- Integrate Deep learning models to the system
- Use Explainable AI techniques (SHAP)
- Using Artificial Intelligence (AI) models like GPT Models, XGBoost (Anomaly Detection), LSTM (Sequence Models) to enhance Card Fraud Detection
- Cloud Deployment
- Integration with banking transaction systems

References:

- 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 and UI workflow diagrams.
- *Credit Card Fraud Detection Platform Market Report, 2033. (n.d.). https://www.grandviewresearch.com/industry-analysis/credit-card-fraud-detection-platform-market-report*
- *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*
- *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*
- *Real-time credit card fraud detection using machine learning. (2019, January 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/8776942*
- *Credit card fraud detection using machine learning. (2020, May 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9121114*
- *Credit card fraud detection using State-of-the-Art machine learning and deep learning algorithms. (2022). IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9755930*
- *Tiwari, P., Mehta, S., Sakhuja, N., Kumar, J., & Singh, A. K. (2021, August 23). Credit Card Fraud Detection using Machine Learning: A Study. arXiv.org. https://arxiv.org/abs/2108.10005*
- *Credit Card Fraud Detection - Machine Learning methods. (2019, March 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/8717766*
- *Khalid, A. R., Owoh, N., Uthmani, O., Ashawa, M., Osamor, J., & 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*
- *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.*

REGIS UNIVERSITY

THANK YOU

Nahian Arnob

International Grad Student in Data Science, Regis University

+1(303)-505-4899 | +8801679039190 (WhatsApp)

arnob.nahian@gmail.com | narnob@regis.edu