Credit Card Fraud Detection — Final Report (EDA + Modeling Summary)

Problem Introduction

Credit card fraud has become one of the most critical challenges in the financial sector. **The goal of this project was to build a machine learning model that predicts whether a transaction is fraudulent or not based on transaction details**. The dataset for this project was sourced from GitHub, containing transaction-level details with labels indicating fraud or non-fraud. Since fraudulent transactions are rare compared to genuine ones, this classification problem is highly imbalanced and requires careful preprocessing and evaluation.

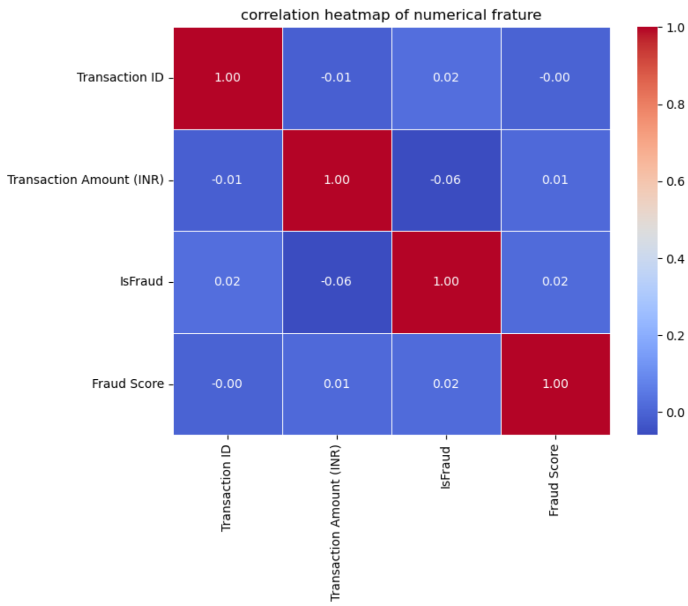
Section 1 — Exploratory Data Analysis (EDA)

1.1 Schema

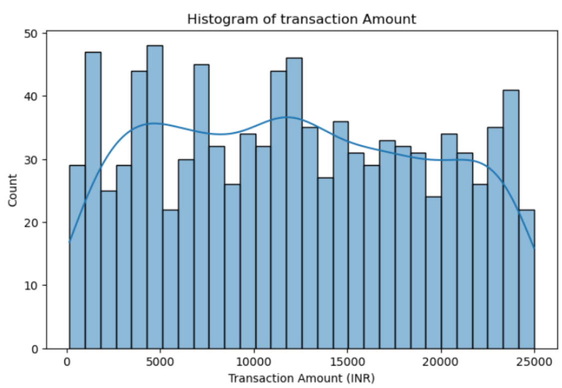
Column	Type
Transaction ID	int64
Customer Name	object
Merchant Name	object
Transaction Date	datetime64[ns]
Transaction Amount (INR)	int64
Fraud Risk	object
Fraud Type	object
State	object
Card Type	object
Bank	object

1.2 Key Plots

Correlation heatmap (Numeric Columns):

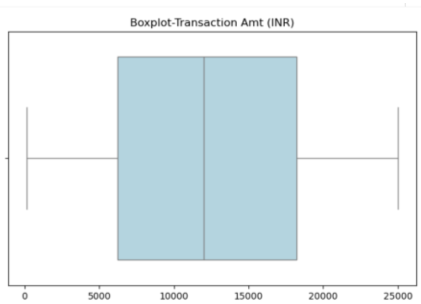


Transaction Amount distribution:




```
# correlation heatmap only numeric feature
```

Transaction Amount box plot (IQR view):




```
Outliers count in Transaction Amount: 0
Empty DataFrame
Columns: [Transaction ID, Customer Name, Merchant Name, Transaction Date, Transaction Amount (INR), Fraud Risk, Fraud Type, State, Card Type, Bank, IsFraud, Fraud Score, Transaction Category, Merchant
Index: []
```

1.3 EDA Insights (brief)

• Linear correlations between numeric features and IsFraud are weak; non-linear models may be advantageous.

• Transaction amounts span a broad range; distribution is fairly spread with few IQR-flagged outliers.

• Typical class imbalance expected in fraud; use stratified CV and recall/PR-AUC for evaluation.

Section 2 — Concepts Applied

To address the problem, the following machine learning and preprocessing techniques were applied:-

Algorithms evaluated: Logistic Regression, Decision Tree (with ColumnTransformer, One-Hot Encoding, StandardScaler).

Model selection: GridSearchCV; post-training threshold tuning to favor fraud recall.

Evaluation metrics: Accuracy, Precision, Recall, F1-Score, ROC-AUC, with focus on fraud recall.

These concepts were chosen to balance model interpretability (Logistic Regression) and predictive power (Decision Tree)

2.1— Results

Model	Accuracy	Fraud Recall	Non-Fraud Recall	ROC-AUC
Logistic Regression (tuned)	**72.50%**	0.65	0.76	0.73
Decision Tree (tuned)	**87.50%**	0.84	0.89	0.95

2.2 Interpretation & Takeaways

Logistic Regression: Simpler, interpretable, moderate performance.-
Decision Tree: Stronger predictive power, higher recall and ROC-AUC.-

If interpretability is required → Logistic Regression.-

If predictive power is required → Decision Tree.

2.3 Recommendations

Strategies

1. Deploying Ensemble Models (RandomForest, XGBoost, LightGBM).

2. Increasing the dataset size

3. Feature Engineering (transaction frequency, anomalies, time features)

. 4. Real-Time Monitoring for instant blocking of suspicious activities.

Conclusion

This project successfully demonstrated the use of machine learning to predict fraudulent transactions. While Logistic Regression offered interpretability, the Decision Tree emerged as the best-performing solution. Future work includes experimenting with ensemble models, feature engineering, and cross-validation to further enhance fraud detection accuracy