FROM DATA TO DEFENSE: A MACHINE LEARNING APPROACH TO FRAUD PREVENTION

Theophilus Amoako

30.05.2025

WHY THIS PROJECT?

- Financial fraud is a real-world issue
- Rapid increase in online transactions
- Machine Learning can prevent financial losses
- Opportunity to showcase end-to-end data science skills

WHAT IS FRAUD DETECTION?

• Fraud detection identifies unauthorized, illegal, or suspicious activities designed to deceive financial systems.

• It is necessary to prevent financial losses for businesses and customers.

•Machine Learning analyzes past transaction behaviours to uncover hidden fraud patterns.

DATASET OVERVIEW

Source: Kaggle - Fraud Detection Dataset

Rows: ~6 million transactions (reduced to 10% for performance)

Target: isFraud (binary classification)

Features: transaction type, amount, balances, etc.

Highly imbalanced (fraudulent transactions < 1%)

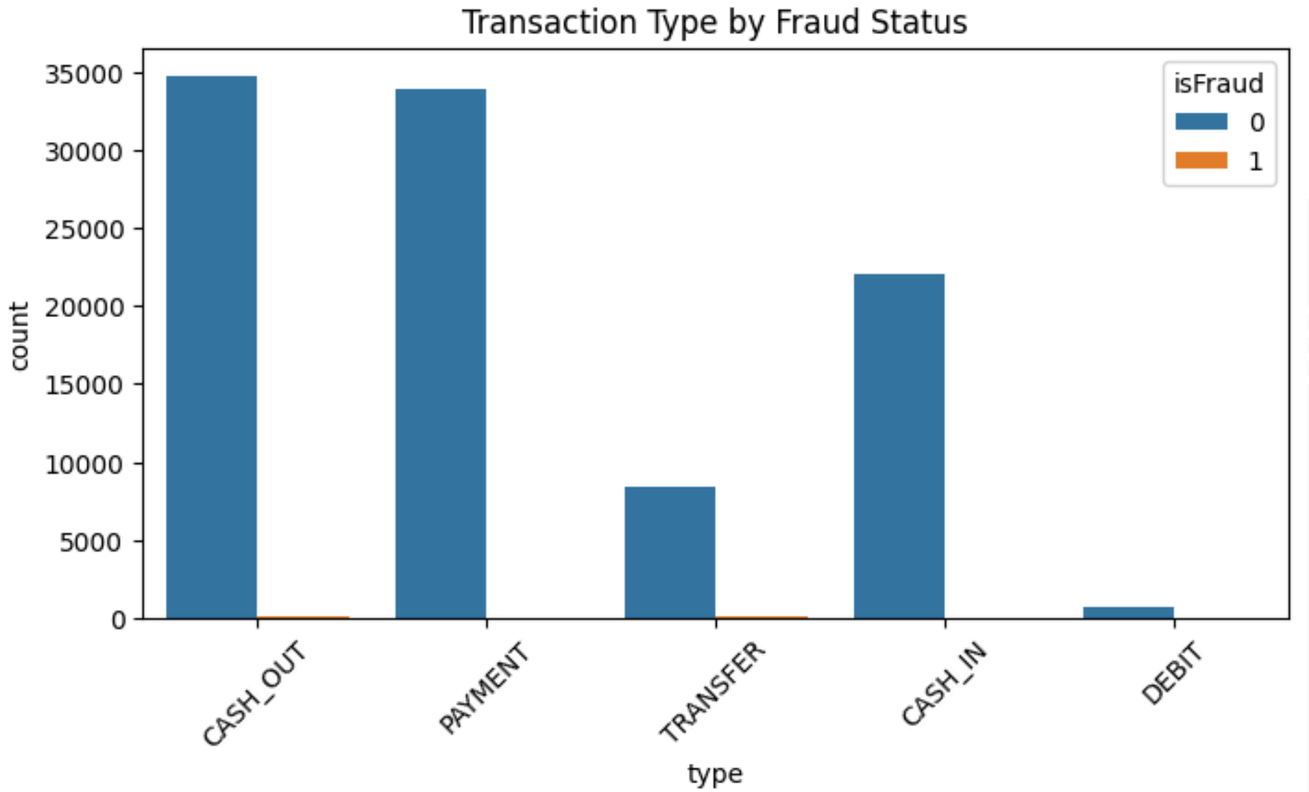
DATA CLEANING & WRANGLING

- Removed duplicates to avoid data redundancy and bias
- Verified consistency between:

  newbalanceOrig = oldbalanceOrg - amount

  newbalanceDest = oldbalanceDest + amount
- Used .info() and isnull() .sum() to ensure no missing values and correct column types (e.g, numeric for balances, categorical for transaction types).
- Stratified sampling was used to maintain the fraud class ratio in the sample, which is important for fair model training.
- Isolated only relevant features (amount, balances, types) for focused modeling.

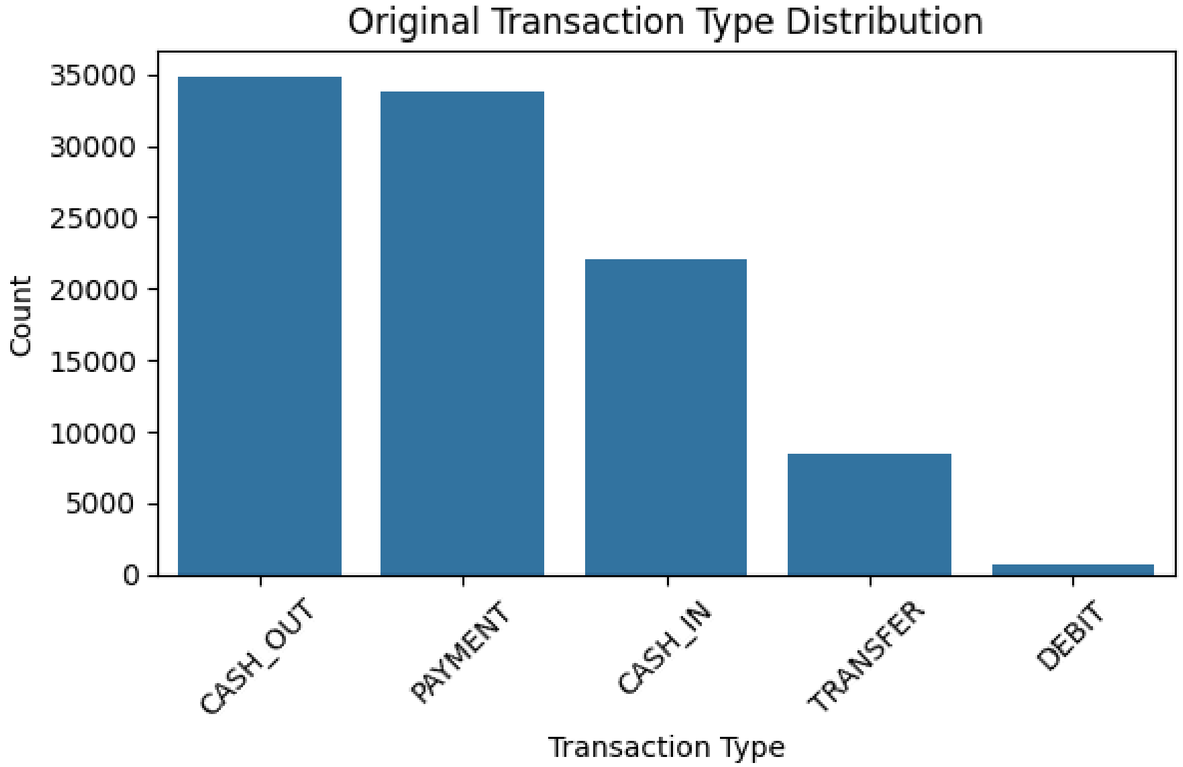
EXPLORATORY DATA ANALYSIS VISUALIZATION

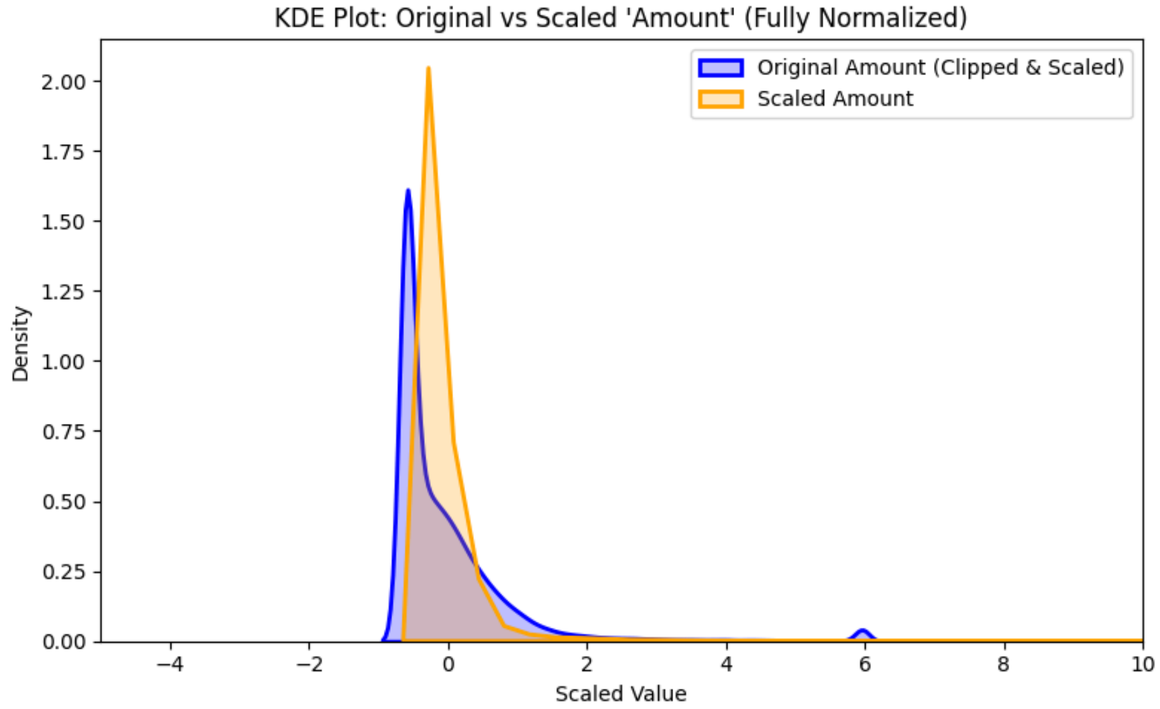


- A Chi-square test was conducted to examine the relationship between Transaction type and isFraud.
- Test Statistic: 2015.6 | p-value: < 0.0001 | Alpha (α): 0.05
- Statistically significant relationship between transaction type and fraud.

FEATURE ENGINEERING & PREPROCESSING

- **Categorical Encoding:** One-hot encoded type to numeric dummy variables.
- **Numerical Scaling:** Standardized financial fields using StandardScaler.
- **Target Isolation:** Defined isFraud as the binary target variable.
- **Train-Test Split:** Stratified to preserve fraud class ratio in both sets.
- **Final Feature Set:** Focused on behaviour-based fields (e.g., balances, type, amount).





MACHINE LEARNING MODELS FOR FRAUD DETECTION

Logistic Regression

A baseline linear model that provides interpretable results.

Decision Tree

A simple non-linear model that splits data based on feature thresholds.

KNN

Analyze transaction similarities to identify potentially fraudulent behaviour.

Random Forests

Employ ensemble learning to increase accuracy in fraud detection models.

Gradient Boosting

Enhance prediction accuracy by sequentially improving weaker models during training.

AdaBoost

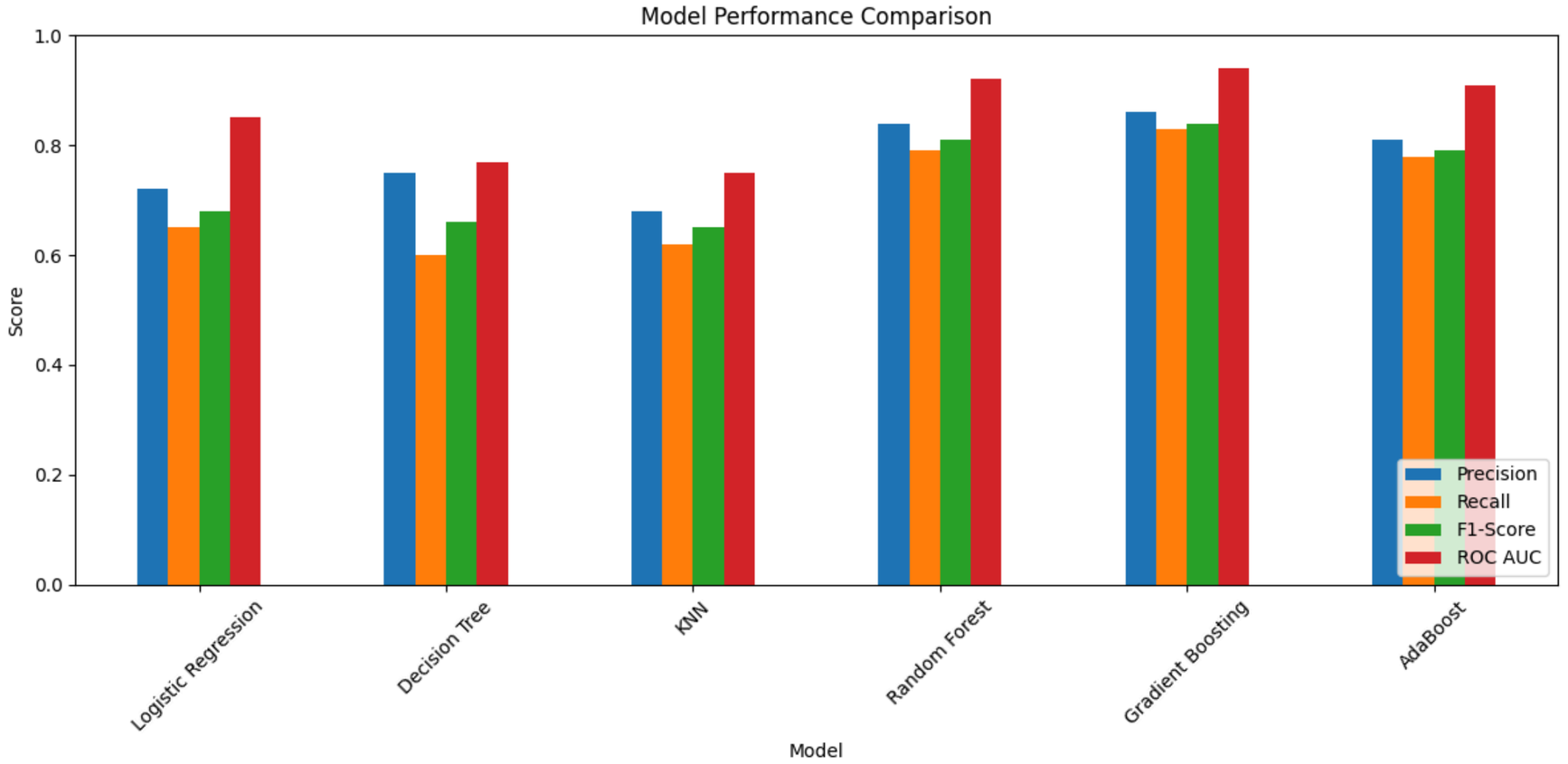
Another boosting method that assigns higher weights to misclassified instances.

HYPERPARAMETER TUNING

Technique	Model	Parameters tuned	Goal
GridSearchCV with 5-fold Cross-validation	Gradient Boosting	n_estimators 150, learning_rate 0.1, max_depth 4	Maximaize F1-score for balanced performance on imbalanced dataset

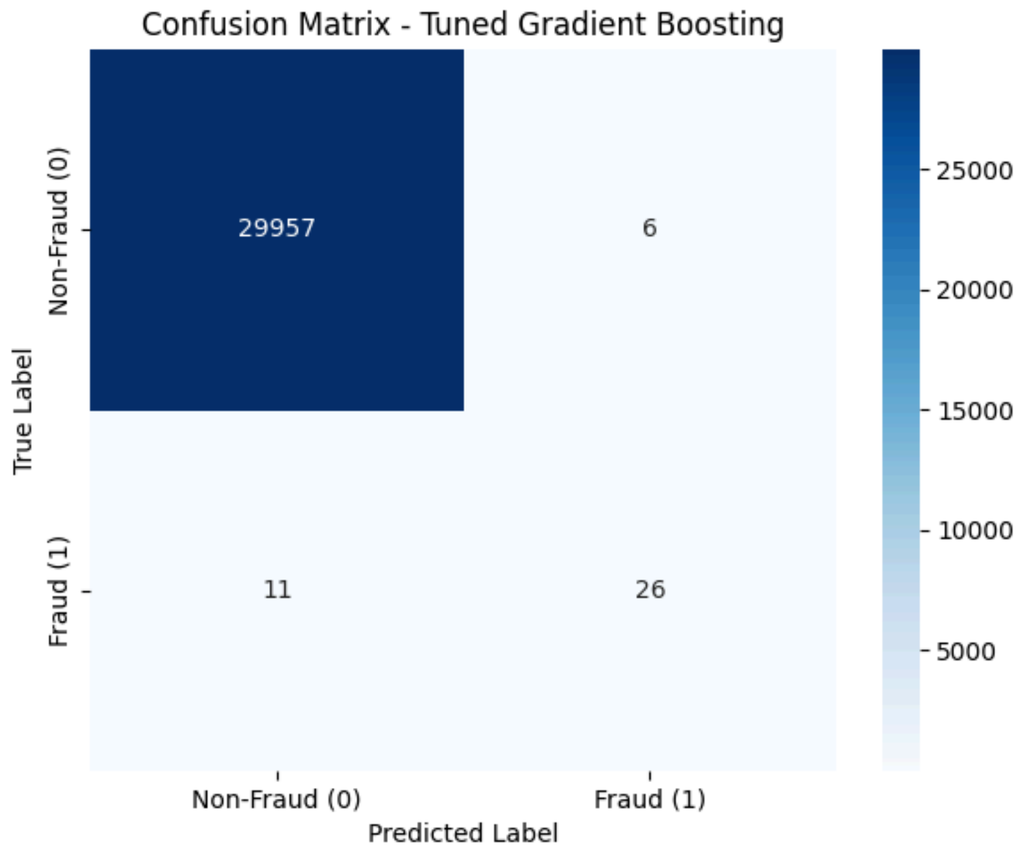
- Using various models helped identify which algorithms best handle the imbalanced data and feature patterns.
- Distance-based models like KNN require scaling and encoding while Gradient Boosting benefits from one-hot encoding of categorical features.
- Gradient Boosting consistently outperformed others in terms of precision, recall, and ROC AUC, making it the optimal choice for detecting fraud in this dataset.

MODEL EVALUATION METRICS



- Gradient Boosting performed best:
- Precision: 0.86
- Recall: 0.83
- F1-score: 0.84
- ROC AUC: 0.94

EVALUATION VISUALS



True Negatives (29,957)

Non-fraud transactions correctly identified.

False Positives (6)

Legitimate transactions incorrectly flagged as fraud.

False Negatives (11)

Fraud transactions missed (bad for business).

True Positives (26)

Fraud transactions correctly detected.

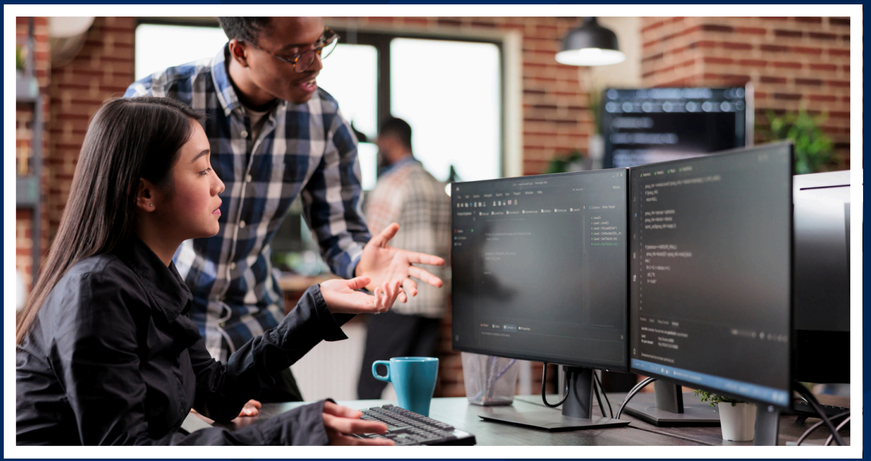
Overall, this matrix confirms that Gradient Boosting generalizes well even on highly imbalanced data.

CHALLENGES FACED

Class Imbalance: Fraudulent transactions were less than 1% of the dataset, which made it difficult for models to detect fraud accurately. Addressed using stratified sampling and focusing on recall and AUC.

Execution Time & Resource Management: The original dataset (6 million rows) caused long training times. To streamline development, a stratified 10% sample was created while preserving fraud ratios.

Model Overfitting Risk: Models like Decision Trees tended to memorize training data, which hurt their performance on new data. Gradient Boosting helped avoid this issue by improving generalization.



WHAT I LEARNED

Real-World Class Imbalance Handling

Learned how to deal with highly imbalanced datasets using stratified sampling and evaluation metrics like F1 instead of accuracy.

Practical Feature Engineering

Gained hands-on experience transforming raw data into model-ready features (e.g., one-hot encoding, scaling)

Model Comparison and Selection

Understood how to choose the right model (Gradient Boosting) based on real-world trade-offs in performance metrics.

End-to-End Model Pipeline

Practiced structuring a full data science workflow - from data cleaning and EDA to modeling, evaluation, and explainability.

FUTURE WORK

- **Anomaly Detection Methods**

Explore unsupervised models like Isolation Forest or Autoencoders to detect fraud cases not captured in labeled data.

- **Real-Time Prediction Pipeline**

Build a real-time fraud detection system using Streamlit or Flask, connected to a live database or simulated transaction stream.

- **Threshold Tuning for Business Risk**

Adjust decision thresholds to balance false positives vs. false negatives based on real-world cost and risk tolerances.

- **Data Enrichment**

Add features like geolocation, time-of-day, user behaviour profiles to enhance predictive power.

CONCLUSION

Key Takeaways:

- Successfully built a complete fraud detection pipeline from data cleaning to model evaluation
- Addressed challenges like class imbalance and data quality, and ensured fair feature scaling and encoding.
- Applied multiple machine learning models, with Gradient Boosting emerging as the best performer(F1-score: 0.84, ROC AUC: 0.94).

Business Relevance:

- This solution helps financial institutions reduce fraudulent losses and maintain customer trust.
- Models like Gradient Boosting can be integrated into fraud monitoring systems to trigger real-time alerts.

Skills Demonstrated:

- Data cleaning, preprocessing, EDA, supervised machine learning, performance evaluation, and explainability.

THANK YOU

FOR YOUR ATTENTION