Predicting Fraudulent Transactions Using Machine Learning

1. Problem Statement

- Zelle is a popular digital payment platform, and as its usage grows, so does the incidence of fraudulent transactions. The goal of this project is to develop a machine learning model to predict fraudulent Zelle transactions, thereby assisting in early detection and prevention.

2. Objectives

- Develop a machine learning model to predict fraudulent transactions.
- Identify key features and patterns associated with fraudulent behavior.
- Reduce false positives while ensuring a high rate of fraud detection.

3. Conclusions:

- The **Decision Tree** model is clearly superior in detecting fraud, as it has significantly higher precision and a balanced F1 score compared to Logistic Regression.
- Logistic Regression has high recall but very low precision for fraud, leading to an unacceptable number of false alarms, which is not practical for a business use case.
- We Choose a **Decision Tree model** to deploy.

5. Data Set:

The dataset titled "Fraudulent Transactions Data" on Kaggle contains simulated transaction data used for the detection of fraud. The features include:

- **step**: Represents a time unit (in hours) for a 30-day simulation.
- **type**: Transaction type (e.g., CASH-IN, CASH-OUT, DEBIT, PAYMENT, TRANSFER).
- **amount**: Transaction amount.

- **nameOrig**: The customer initiating the transaction.
- **oldbalanceOrg** and **newbalanceOrig**: Balance of the originator before and after the transaction.
- **nameDest**: The recipient of the transaction.
- **oldbalanceDest** and **newbalanceDest**: Recipient's balance before and after the transaction.
- **isFraud**: Indicates if the transaction is fraudulent.

Note: Only 0.12% of the data is fraud. Thus we use a stratified sampling method to ensure the balance in both training and testing sets.

4. Results

- **We use two models:**
- **Logistic Regression:** Logistic regression is a binary classification model that predicts one of two outcomes, such as fraudulent or non-fraudulent transactions. It estimates the probability of an input belonging to a particular class using the sigmoid function, which maps values to a range between 0 and 1. A threshold, typically 0.5, is used to classify inputs: below the threshold is non-fraudulent, above is fraudulent.
    - From a business standpoint, logistic regression offers interpretability and ease of implementation. It provides coefficients indicating the importance of each feature in predicting fraud, which is valuable for understanding predictions and meeting regulatory requirements. The model's simplicity, efficiency, and effectiveness for approximately linear relationships make it suitable for applications requiring transparent decision-making.
- **Decision Tree:** Decision trees are classification models that split data into branches, forming a tree-like structure for decision-making. Internal nodes represent features, branches represent decision rules, and leaf nodes represent outcomes. The model begins at the root node and makes splits based on the most informative features until reaching a stopping criterion.

    - In a business context, decision trees are highly interpretable, mimicking human decision-making processes. They can handle both numerical and categorical data, making them flexible for capturing complex patterns in fraud detection. The model's ability to provide a clear decision path from root to leaf node makes results easily explainable to stakeholders and auditors.
- Both models offer distinct advantages in fraud detection, with logistic regression providing simplicity and linear interpretability, while decision trees offer flexibility and the ability to capture non-linear relationships in data.

Comparative Study:

	Prediction	Logistic Regression	Decision Tree
Precision	Fraud	0.03	0.89
	Non - Fraud	1.00	1.00
Recall	Fraud	0.89	0.86
	Non - Fraud	0.96	1.00
F1 Score	Fraud	0.06	0.87
	Non - Fraud	0.93	1.00

- **Precision**: Precision measures the accuracy of positive predictions. Specifically, it is the ratio of correctly identified positive cases (e.g., fraudulent transactions) to the total number of cases predicted as positive. In fraud detection, high precision means fewer false alarms, which is important for not mistakenly flagging legitimate transactions as fraud.
- **Recall**: Recall measures how well the model identifies all actual positive cases. It is the ratio of correctly identified positive cases to the total number of actual

positive cases. High recall means the model is good at capturing most of the fraud cases, which helps minimize missed fraudulent activities.

- F1 Score: The F1 score is the harmonic mean of precision and recall, providing a balance between them. It is particularly useful when the data is imbalanced, as it accounts for both false positives and false negatives. A high F1 score indicates that both precision and recall are reasonably high, which is crucial in fraud detection to catch as many frauds as possible without many false alarms.

Comparing Logistic Regression and Decision Tree Models

Fraud Prediction:

- Precision:
  - Logistic Regression: 0.03 (very low), meaning it has many false positives and is not reliable in predicting fraud accurately.
  - Decision Tree: 0.89, meaning it has fewer false positives and performs much better at correctly predicting fraudulent transactions.
- Recall:
  - Logistic Regression: 0.89, indicating it catches most of the actual frauds.
  - Decision Tree: 0.86, slightly lower but still effective at identifying most fraudulent activities.
- F1 Score:
  - Logistic Regression: 0.06, indicating poor performance as precision is too low despite high recall.
  - Decision Tree: 0.87, indicating a good balance between precision and recall.

Non-Fraud Prediction:

- **Precision and Recall**: Both models have perfect or near-perfect scores for non-fraud cases, meaning they are both effective at correctly identifying legitimate transactions.