Project Documentation: Payment Fraud Detection using Logistic Regression

Introduction

Project Title: Payment Fraud Detection Using Logistic Regression

Objective: The goal of this project is to build a machine learning model to detect fraudulent payment transactions. By analyzing historical transaction data, the model aims to distinguish between legitimate and fraudulent transactions, helping to prevent financial losses and enhance security.

Data Description

Dataset: The dataset used in this project is named `payment_fraud.csv`, which includes various features related to payment transactions.

- **paymentMethod**: The method of payment (e.g., Credit Card, PayPal, etc.).
- **Other features**: Additional features related to the transaction (e.g., amount, location, etc.).
- **label**: The target variable, indicating whether a transaction is legitimate (`0`) or fraudulent (`1`).

Steps and Methodology

1. **Data Loading and Exploration**:
   - Loaded the dataset using `pandas`.
   - Displayed the first few rows of the dataset using `df.head()`.
   - Checked for null values using `df.isnull().sum()`.
2. **Data Visualization**:
   - Visualized the distribution of the payment methods using a bar plot.
   - Counted the number of legitimate and fraudulent transactions using `df.label.value_counts()`.
3. **Data Preprocessing**:
   - **Label Encoding**: Converted the `paymentMethod` column into numerical labels using dictionary mapping.
   - **Correlation Analysis**: Used a heatmap to visualize the correlation between features.
   - **Standardization**: Standardized the independent features using `StandardScaler` to ensure that all features contribute equally to the model.
4. **Splitting the Data**:
   - Split the dataset into training and testing sets using `train_test_split`, with 25% of the data reserved for testing.
5. **Model Building**:
   - Chose Logistic Regression as the model for its simplicity and effectiveness in binary classification problems.

    - Trained the model using the training data (`lg.fit(X_train, y_train)`).

6. **Model Evaluation**:
    - Made predictions on the test set using `lg.predict(X_test).`
    - Evaluated the model's performance using accuracy score, classification report, and confusion matrix.

Results

- **Accuracy**: The model achieved an accuracy score on the test set, indicating the percentage of correctly predicted transactions.
- **Classification Report**: Provided detailed metrics including precision, recall, and F1-score for both legitimate and fraudulent classes.
- **Confusion Matrix**: Visualized the true positives, true negatives, false positives, and false negatives, helping to understand the model's performance in different scenarios.

Conclusion

The logistic regression model effectively identified fraudulent transactions with a reasonable accuracy. The detailed evaluation metrics demonstrated the model's ability to distinguish between legitimate and fraudulent transactions, which is crucial for preventing financial fraud.

Future Work

1. **Feature Engineering**:
    - Incorporate additional features such as user behavior patterns, time-based features, and external data sources to improve model accuracy.
2. **Advanced Models**:
    - Experiment with more complex models such as Random Forest, Gradient Boosting, and Neural Networks to potentially enhance performance.
3. **Real-Time Detection**:
    - Implement the model in a real-time fraud detection system, integrating it with transaction processing systems for instant fraud prevention.
4. **Imbalanced Data Handling**:
    - Address the class imbalance problem using techniques like SMOTE (Synthetic Minority Over-sampling Technique) to improve the model's performance on the minority class (fraudulent transactions).