Project Documentation: Payment Fraud Detection using Logistic Regression

Introduction

The goal of this project is to build a machine learning model to detect payment fraud using logistic regression. We will perform data preprocessing, visualization, feature scaling, model training, and evaluation.

Steps Involved

1. Importing Libraries

We start by importing the necessary libraries for data manipulation, visualization, and machine learning.

2. Loading and Exploring the Data

Load the dataset from a CSV file and perform an initial exploration to understand its structure and contents.

3. Data Visualization

Visualize the distribution of payment methods to understand their frequencies.

4. Data Preprocessing

Convert categorical values in the `paymentMethod` column into numerical labels.

5. Feature Correlation

Visualize the correlation between features using a heatmap.

6. Statistical Summary

Get a statistical summary of the dataset

7. Splitting Data into Independent and Dependent Features

Separate the dataset into features (independent variables) and the target (dependent variable).

8. Feature Scaling

Standardize the features to have a mean of 0 and a standard deviation of 1.

9. Train-Test Split

Split the dataset into training and testing sets.

10. Model Training

Train a logistic regression model on the training data.

11. Model Prediction

Make predictions on the test data.

12. Model Evaluation

Evaluate the model's performance using accuracy, classification report, and confusion matrix.

Conclusion

The logistic regression model was trained and evaluated to detect payment fraud. The evaluation metrics such as accuracy, classification report, and confusion matrix provide insights into the model's performance.

Summary of Project Workflow

1. **Import Libraries**: Load necessary libraries for data manipulation, visualization, and machine learning.
2. **Load and Explore Data**: Read the dataset and perform initial exploration to understand the data structure.
3. **Data Visualization**: Visualize the distribution of payment methods and target variable.
4. **Data Preprocessing**: Convert categorical values to numerical labels.
5. **Feature Correlation**: Visualize the correlation between features.
6. **Statistical Summary**: Obtain a statistical summary of the dataset.
7. **Split Data into Features and Target**: Separate the dataset into independent and dependent variables.
8. **Feature Scaling**: Standardize the features.
9. **Train-Test Split**: Split the dataset into training and testing sets.
10. **Model Training**: Train a logistic regression model.
11. **Model Prediction**: Make predictions on the test data.
12. **Model Evaluation**: Evaluate the model's performance using various metrics.

This documentation outlines the process of building a logistic regression model to detect payment fraud, providing a clear understanding of each step involved in the project.