Capstone Three: Credit Card Fraud Detection Project Documentation

Introduction

Objective: The primary goal of this project is to develop a neural network-based system capable of detecting fraudulent transactions in real-time. This initiative, spearheaded by Company ABC, aims to enhance security measures in the financial sector, leveraging advanced machine learning techniques to adapt to and identify emerging fraudulent tactics promptly.

Data Wrangling and Exploratory Data Analysis (EDA)

Data Cleaning:

- Missing Values: There were no missing values in the datasets used, which included cardholder information and transaction logs.
- Outlier Detection: Outlier detection and handling were implicitly considered through robust data processing techniques.

Exploratory Data Analysis:

- Visualizations and Insights:
    - Transaction amounts and their distributions were visualized using histograms to understand spending patterns.
    - Transaction activities were explored by days of the week, revealing any particular trends and consistencies in transaction volumes.
- Correlation Analysis: A correlation matrix was generated for features such as transaction amount, longitude, and latitude to identify any significant relationships.

Feature Engineering:

- Time Features: Extracted day of the week, hour of the day, and transactions in the last hour from the date field, adding temporal context to the data.

Pre-processing Work and Modeling

Data Pre-processing:

- Feature Scaling: Applied StandardScaler to normalize the transaction dollar amounts, ensuring model sensitivity to small deviations in spending behavior.
- Categorical Encoding: Utilized one-hot encoding for the day of the week to transform this categorical data into a machine-readable format.

Model Development and Evaluation:

- Model Architecture: The model consisted of an input layer adjusted to 18 features, followed by two hidden layers with 16 neurons each, and a sigmoid activation function in the output layer for binary classification.
- Model Training: Trained using a batch size of 32 over 20 epochs, showing gradual improvement in accuracy.

- Model Performance: The final model achieved an accuracy of approximately 95.11% on the validation set, indicating strong performance in identifying fraudulent transactions.

Results and Conclusion

- Effectiveness: The model effectively identifies fraudulent transactions, with a high accuracy rate on unseen data, demonstrating its potential utility in a real-world environment.
- Future Work: Future improvements could include experimenting with different architectures, more sophisticated feature engineering (e.g., incorporating merchant categories), and deploying the model in a live environment to further validate its effectiveness.

Appendices

- Code Listings: Complete code for this project is available on GitHub at hyserena/DataScienceCapstone-CreditCardFraudDetection.
- Data Sources: Data utilized for this project is sourced from Kaggle's Credit Card Fraud Detection dataset.

References

- Kaggle Dataset for Credit Card Fraud Detection.
- Python libraries such as Pandas, Numpy, Scikit-Learn, TensorFlow, and Matplotlib were extensively used for data manipulation, analysis, and model building.