Detecting Fraudulent Financial Transactions Using Machine Learning

Ansh Rana (arr238)
Sahil Sanghvi (sss426)
Omar Ghanima (omg29)

1 Research Question and Motivation

Research Question

Can machine learning algorithms be effectively applied to detect fraudulent financial transactions within a highly imbalanced dataset?

Motivation and Background

Financial fraud incurs significant losses annually, and traditional rule-based systems often fail to adapt to evolving fraudulent behaviors. Leveraging machine learning could provide a more accurate solution by identifying hard-to-find patterns in transactional data.

2 Data Sources

Primary Data Source

We plan to use the publicly available *Credit Card Fraud Detection* dataset from Kaggle, which contains transaction data for European cardholders.

Data Acquisition and Preprocessing

- **Acquisition:** Download the dataset from Kaggle.
- **Cleaning:** Identify and handle missing values, remove anomalies, and normalize numerical features.
- **Preprocessing:**
  - Address class imbalance using techniques such as SMOTE or undersampling.
  - Apply feature scaling (e.g., StandardScaler) for model stability.
  - Consider dimensionality reduction methods (e.g., PCA) if necessary.

3 Methodology and Analysis Plan

Data Science Techniques and Approach

- **Exploratory Data Analysis (EDA):** Use visualization tools (e.g., Matplotlib, Seaborn) to analyze feature distributions and class imbalances.

- **Modeling:** Start with baseline models (e.g., logistic regression, decision trees) and then explore ensemble methods like Random Forests and Gradient Boosting Machines.
- **Handling Imbalance:** Implement resampling techniques and cost-sensitive learning methods.

Tools and Libraries

- **Programming Language:** Python
- **Data Manipulation:** Pandas, NumPy
- **Visualization:** Matplotlib, Seaborn
- **Machine Learning:** Scikit-Learn, Imbalanced-learn
- **Optional Deep Learning:** TensorFlow or PyTorch

4 Expected Outcomes and Evaluation

Anticipated Findings

- A robust predictive model that effectively detects fraudulent transactions with high recall.
- Identification of key features that contribute to accurate fraud detection.
- A comparative analysis of different modeling techniques in handling imbalanced datasets.

Evaluation Criteria

- **Performance Metrics:** Precision, Recall, F1-Score, and ROC-AUC.
- **Robustness:** Testing model performance using various resampling strategies.
- **Scalability and Efficiency:** Assessing the model's potential for real-time fraud detection.

Potential Extensions

- Integration of additional data sources (e.g., geographical data) to enhance detection robustness.
- Development of a real-time fraud detection system.
- Exploration of advanced deep learning techniques to further improve model performance.