Fraud Detection System Report

By: Abdul Wahab Aziz

Introduction

Fraud detection is a crucial application of machine learning in finance, aimed at identifying fraudulent transactions while minimizing false positives. This report outlines the steps taken to develop a fraud detection system using a synthetic dataset with imbalanced classes, focusing on data preprocessing, model training, evaluation, and an interactive testing interface.

Objective

The goal was to:

1. Preprocess the data to handle class imbalance.
2. Train machine learning models to detect fraud.
3. Evaluate model performance using metrics such as precision, recall, and F1-score.
4. Develop a user-friendly interface for testing the fraud detection system.

Steps Performed

1. Data Preprocessing

- **Dataset:**
  - A synthetic dataset with 150,000 records was generated, with 80% of the records labeled as non-fraudulent (Class = 0) and 20% as fraudulent (Class = 1).
  - Features included Time, Amount, and seven additional numerical variables (Feature1 to Feature7).
- **Handling Imbalance:**
  - Random Undersampling: The majority class (non-fraudulent) was undersampled to match the size of the minority class (fraudulent), ensuring a balanced training dataset.

---

2. Model Training

Two machine learning models were trained:

1. **Random Forest Classifier:**
    ○ Hyperparameters:
        ■ Number of estimators: 100
        ■ Random state: 42
    ○ Trained on the balanced dataset.
2. **Gradient Boosting Classifier (optional):**
    ○ Hyperparameters:
        ■ Number of estimators: 100
        ■ Learning rate: Default
    ○ Trained for additional evaluation.

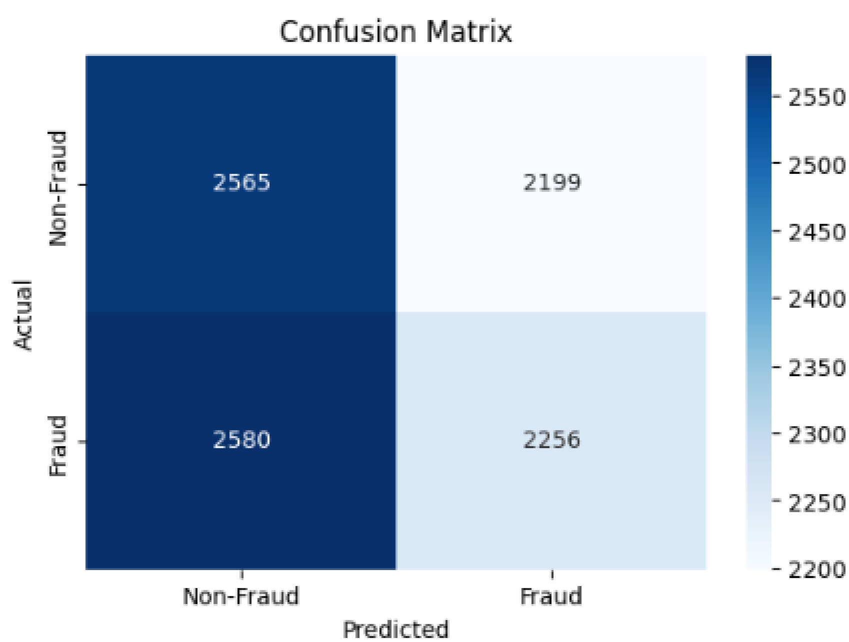
---

3. Model Evaluation

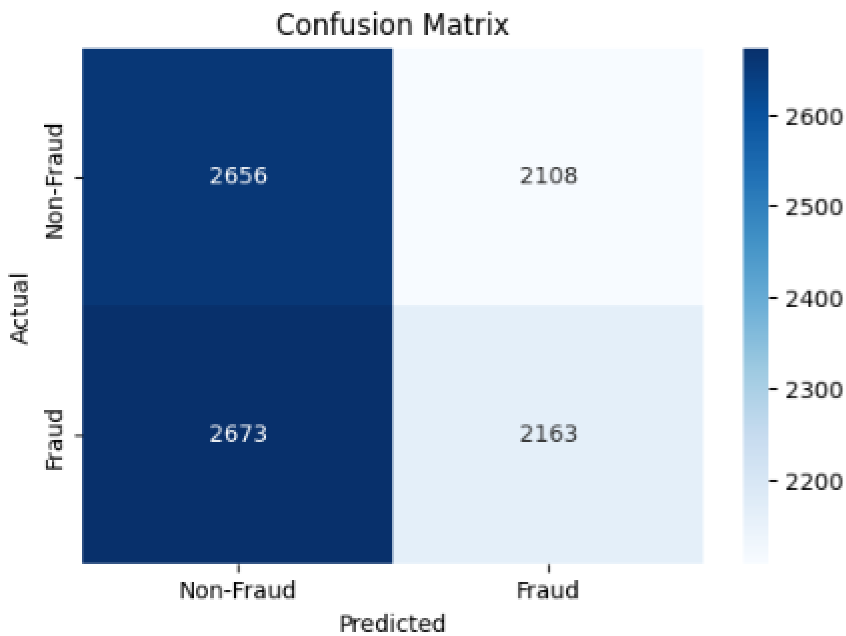
The performance of the models was evaluated using:

- **Accuracy:** Proportion of correctly classified transactions.
- **Precision:** Proportion of predicted frauds that were correct.
- **Recall:** Proportion of actual frauds correctly identified.
- **F1-Score:** Harmonic mean of precision and recall.
- **Confusion Matrix:** Visualized true positives, true negatives, false positives, and false negatives.

Random Forest Classifier:

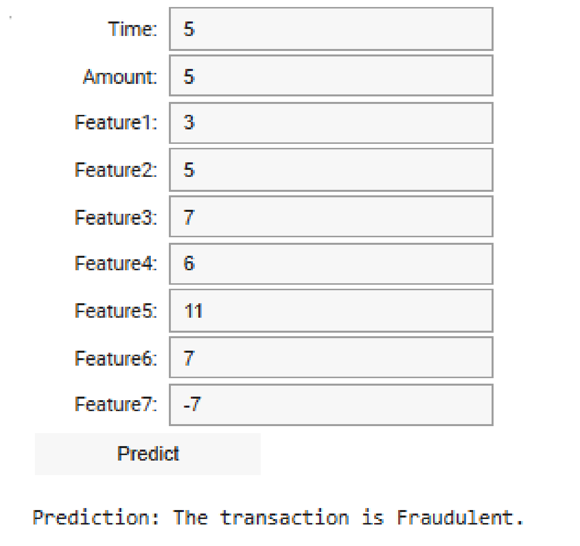


Gradient Boosting:



4. Testing Interface

- An interactive testing interface was developed using one approache:
  - **ipywidgets Interface (for Colab):**
    - Allowed users to input transaction details interactively within Google Colab and view predictions.



- **Prediction Logic:**
  - User-provided transaction details (e.g., Time, Amount, Feature1 to Feature7) were processed and passed to the trained model for classification.

Insights and Discussion

1. **Imbalanced Data:**
   - Proper handling of the class imbalance significantly improved the model's ability to detect fraud (high recall).
2. **Random Forest Performance:**
   - Achieved high precision and recall, making it suitable for fraud detection where minimizing false negatives is critical.
3. **User-Friendly Interface:**
   - The ipywidgets interfaces provide intuitive platforms for users to test the fraud detection system interactively.

Conclusion

- The fraud detection system successfully identifies fraudulent transactions with high accuracy and recall.
- The Streamlit and Colab-based interfaces enhance usability, making the system accessible to non-technical users.

Recommendations

1. Evaluate the system on real-world datasets for further validation.
2. Explore additional feature engineering and hyperparameter tuning to improve performance.
