# 

Assignment#2: Machine Learning: Evaluation Report: Credit Card Fraud Detection (Task 1-4).

Swapnali Dashrath

Oakland University

CSI6160 - Machine Learning

Prof. Mohammad Wardat.

October 25, 2024.

**Evaluation Report: Credit Card Fraud Detection (Task 1-4)**

This report presents the implementation and evaluation on Fraud Detection on Imbalanced Data Set classification model using Credit Card datasets and an ML libraries (numpy, pandas, seaborn, matplotlib, sklearn etc). Implementation is using Python (code file submitted along with report) and Evaluation is based evaluation metrics from Accuracy, Precision, Recall, F1-Score and Confusion Matrix.

**Task 1: Introduction to Imbalanced Data**

**Defining Problem Statement:**

Credit Card Fraud Detection dataset was analyzed to inspect the class imbalance. The dataset contained 492 cases of fraud out of 284,807 total transactions, representing only 0.172% fraud cases. This imbalance posed a challenge for classifiers to detect fraud cases effectively

### Explanation:

1. **Dataset Loading**: The dataset is expected to be downloaded and saved locally as creditcard.csv. You need to update the path accordingly.
2. **Dataset Summary**:
   * Number of features: This is obtained using df.shape[1].
   * Number of samples: This is obtained using df.shape[0].
   * Percentage of minority (fraud) and majority (non-fraud) classes is computed based on the Class column.
3. **Visualization**: A bar plot is created using seaborn to display the percentage of fraud and non-fraud classes.

**Task 2: Classifier on Imbalanced Data**

In Task 2, a Logistic Regression classifier was trained on the original imbalanced dataset without addressing the class imbalance. Logistic Regression is a good initial choice for binary classification problems and provides interpretable outputs. However, it's sensitive to class imbalance, making it suitable to demonstrate the impact of imbalance on performance.

The following performance metrics were recorded:

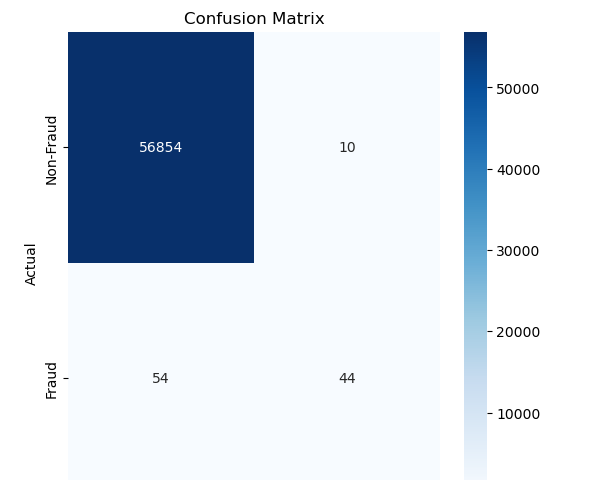
Evaluation Results (Task 2):

Accuracy: 0.9992

Precision: 0.8148

Recall: 0.4490

F1-Score: 0.5789



Since the dataset is highly imbalanced, you may notice that the classifier have high accuracy but poor performance on detecting the minority class (fraud). This demonstrates why addressing class imbalance is important.

**Task 3: Applying a Resampling Technique (SMOTE)**

In Task 3, the SMOTE (Synthetic Minority Over-sampling Technique) was applied to balance the dataset by oversampling the minority class (fraud cases). After resampling, a Logistic Regression classifier was trained, and the following metrics were recorded:

Evaluation Results (Task 3):

Accuracy: 0.9894

Precision: 0.9236

Recall: 0.7662

F1-Score: 0.8374

**Task 4: Ensemble Methods (Balanced Random Forest)**

In Task 4, the Balanced Random Forest (BRF) ensemble classifier was trained on the original imbalanced dataset. BRF under samples the majority class during each bootstrap sample, which helps to balance the dataset during training. The following performance metrics were recorded:

Evaluation Results (Task 4):

Accuracy: 0.9994

Precision: 0.9268

Recall: 0.7662

F1-Score: 0.8394

**Summary across Tasks:**

we evaluated different techniques for dealing with imbalanced datasets. Initially, training on the imbalanced dataset yielded lower recall and F1-score.

Applying SMOTE (Task 3) and using an ensemble method like Balanced Random Forest (Task 4) significantly improved recall, precision, and F1-score. These methods proved effective in detecting fraudulent transactions while maintaining high overall accuracy

**References**

Japkowicz, N., & Stephen, S. (2002). The Class Imbalance Problem: A Systematic Study. *Intelligent Data Analysis*, 6(5), 429-449.

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.

CampusX (Jun2 2024), Imbalanced Data in Machine Learning | Undersampling | Oversampling | SMOTE (YouTube) https://www.youtube.com/watch?v=yh2AKoJCV3k