SFWR TECH 4DA3 - Course Project  
Comparing Classifiers

Joe Choi(choih42)

Sonali Gupta(gupts17)

Satnam Singh Sandhu (sandhs68)

Credit Card Fraud Detection

Logistic Regression and Decision Tree

# Summary of Credit Card Fraud Detection Project

## Dataset Description

Data Source: <https://www.kaggle.com/code/gpreda/credit-card-fraud-detection-predictive-models/notebook>

The dataset used for this project is sourced from Kaggle and contains credit card transactions made by European cardholders in September 2013. It includes 284,807 transactions, with 492identified as fraudulent. This creates a highly imbalanced dataset where fraudulent transactions represent only 0.172% of the total.

The data has undergone PCA transformation for confidentiality, resulting in the features V1 to V28. The only original features retained are "Time" (seconds elapsed since the first transaction) and "Amount" (transaction amount). The target variable is "Class," where 1 indicates fraud and 0 indicates a legitimate transaction.

## Libraries Used

This project utilizes several Python libraries for data analysis, preprocessing, model training, and evaluation:

* **pandas:** Used for data manipulation and analysis.
* **sklearn:** Tools for model selection, preprocessing, training, and evaluation
* **time:** Used to measure the training and testing time for each model.

## Data Preprocessing

The dataset was loaded into a pandas DataFrame. The features were identified as all columns except the 31st which was extracted as the target variable. A preprocessing pipeline was created to standardize the numeric features using StandardScaler. The dataset was then split into training and testing sets with a 75%-25% ratio (213,605 training samples and 71,202 testing samples) to evaluate the models on unseen data.

## Model Training and Evaluation

Two classification models were trained and evaluated: Logistic Regression and Decision Tree.

### Logistic Regression:

* **Training Time:** Approximately 0.298 seconds.
* **Testing Time:** Approximately 0.011 seconds.
* **Confusion Matrix:**

|  |  |
| --- | --- |
| **False positive**  Incorrectly predicted fraudulent transactions when they were negative | **True positive**  Correctly predicted the fraudulent transactions |
| 12 | 69 |
| **True negative**  correctly predicted the non-fraudulent transactions | **False negative**  Incorrectly predicted non-fraudulent transactions when they were positive |
| 71077 | 44 |

This indicates that the Logistic Regression model correctly identified 71,077 legitimate transactions (TN) and 69 fraudulent transactions (TP). However, it misclassified 12 legitimate transactions as fraudulent (FP) and failed to identify 44 fraudulent transactions (FN).

### Decision Tree:

* **Training Time:** Approximately 8.063 seconds.
* **Testing Time:** Approximately 0.016 seconds.
* **Confusion Matrix:**

|  |  |
| --- | --- |
| **False positive**  Incorrectly predicted fraudulent transactions when they were negative | **True positive**  Correctly predicted the fraudulent transactions |
| 37 | 83 |
| **True negative**  correctly predicted the non-fraudulent transactions | **False negative**  Incorrectly predicted non-fraudulent transactions when they were positive |
| 71052 | 30 |

The Decision Tree model correctly identified 71,054 legitimate transactions (TN) and 82 fraudulent transactions (TP). It misclassified 35 legitimate transactions as fraudulent (FP) and failed to identify 30 fraudulent transactions (FN).

## Analysis of Confusion Matrix:

Both models demonstrate high accuracy in classifying the majority class (non-fraudulent transactions). However, the Decision Tree model shows slightly better performance in identifying the minority class (fraudulent transactions). This is evident in the confusion matrices:

* Logistic Regression: Correctly identified 69 fraudulent transactions (TP) and missed 44 (FN), resulting in a ~61% success rate in detecting actual fraud.
* Decision Tree: Correctly identified 83 fraudulent transactions (TP) and missed 30 (FN), resulting in a ~73% success rate in detecting actual fraud.

The Decision Tree's higher true positive rate and lower false negative rate suggest that it is more sensitive to fraudulent patterns in the data compared to Logistic Regression.

## Conclusion

Both models performed well in classifying non-fraudulent transactions. However, the Decision Tree model demonstrated a slightly better ability to detect fraudulent transactions, which is crucial in fraud detection applications. While the Logistic Regression model was significantly faster to train and test, the Decision Tree's improved performance in identifying fraudulent transactions makes it a strong contender, especially when the accurate detection of fraud is a priority.