Multi Model CC Fraud Detection Comparison

A project by Md Shamse Tabrej

Introduction: Credit card fraud poses a significant financial threat to both individuals and institutions. With the increasing volume of online transactions, the need for robust and efficient fraud detection systems has become paramount. This project aims to investigate the effectiveness of different machine learning models in detecting fraudulent credit card transactions.

Data: This project utilizes a real-world dataset comprising credit card transactions made by European cardholders in September 2013. The dataset encompasses two days of transactions, with a total of 284,807 instances. Notably, the dataset exhibits a significant class imbalance, with only 492 instances (approximately 0.172%) representing fraudulent transactions.
Link: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

The dataset features:
Time: The number of seconds elapsed between each transaction and the first transaction in the dataset.
Amount: The transaction amount.
V1-V28: Principal components obtained through PCA transformation. The original features are not disclosed due to confidentiality reasons.
Class: The target variable, indicating whether a transaction is fraud (1) or not fraud (0).

Data Splitting:
To address the significant class imbalance, two distinct approaches were employed:

Balanced Subset: A balanced subset was created by sampling 1000 instances from the non-fraudulent class and using all 492 fraudulent instances. This sampling strategy aimed to mitigate the potential bias towards the majority class during model training.

Full Dataset: The entire dataset (284,315 non-fraudulent and 492 fraudulent transactions) was utilized for training and evaluation. This approach aimed to assess model performance on a more realistic and challenging scenario.

Model Selection:

We will explore and compare the performance of four prominent models:
Logistic Regression: A simple yet powerful linear model for binary classification.
Support Vector Machines (SVM): A versatile model capable of handling high-dimensional data and complex decision boundaries.

Random Forest: An ensemble learning method that combines multiple decision trees to improve accuracy and robustness.
Neural Networks: A powerful class of models that can learn complex patterns and representations from data.

Model Training and Evaluation

Balanced Subset:

Logistic Regression: Achieved a training accuracy of 95.22% and a test accuracy of 97.66%. This indicates good generalization performance on the balanced subset.
SVM: Demonstrated a training accuracy of 91.70% and a test accuracy of 93.65%. While the training accuracy is lower compared to Logistic Regression, the model still exhibits reasonable performance on the test set.
Random Forest: Achieved perfect training accuracy (100%) and a test accuracy of 96.99%. This suggests that the model is well-suited for the balanced dataset, but there might be a slight risk of overfitting due to perfect training accuracy.
Neural Network: After 10 epochs of training, the model achieved a loss of 2.9371 and an accuracy of 72.00%. This performance is significantly lower compared to the other models, indicating that the neural network may require further tuning or a different architecture to effectively learn from the balanced subset.

Full Dataset:

Logistic Regression: Achieved a training accuracy of 99.92% and a test accuracy of 99.93%. This indicates excellent performance on the full dataset, demonstrating the model's ability to generalize well to the imbalanced data distribution.
SVM: Demonstrated a training accuracy of 99.87% and a test accuracy of 99.87%. Similar to Logistic Regression, the SVM model exhibits high accuracy on both training and test sets, suggesting strong performance on the full dataset.
Random Forest: Achieved perfect training accuracy (100%) and a test accuracy of 99.95%. This further highlights the effectiveness of Random Forest in handling the imbalanced dataset, with excellent performance on both training and testing.
Neural Network: After 10 epochs of training, the model achieved a loss of 0.0155 and an accuracy of 99.81%. This demonstrates a significant improvement compared to the balanced subset, showcasing the neural network's ability to learn effectively from the larger, more representative dataset.

Model Comparison:

On the balanced subset, Logistic Regression and SVM exhibited strong generalization performance, while Random Forest showed potential overfitting despite high test accuracy. The Neural Network underperformed on this subset, likely requiring further optimization. However, on the full dataset, all models demonstrated excellent performance, with Logistic Regression, SVM, and Random Forest consistently achieving high accuracy. The Neural Network also showed significant improvement on the full dataset, highlighting the importance of appropriate model architecture and hyperparameter tuning for handling imbalanced data. Overall, Logistic Regression and SVM proved to be robust and effective across both datasets, while Random Forest excelled in handling the imbalanced data.

Conclusion:

The analysis reveals varying model performance across the balanced and full datasets. On the balanced subset, Logistic Regression and SVM demonstrated strong generalization with high accuracy, while Random Forest exhibited potential overfitting despite good test accuracy. The Neural Network underperformed on the balanced subset, suggesting the need for further optimization.

However, when evaluated on the full, imbalanced dataset, all models showed significant improvement. Logistic Regression and SVM consistently achieved high accuracy, showcasing their robustness in handling imbalanced data. Random Forest maintained its strong performance, further emphasizing its effectiveness in this scenario. The Neural Network also demonstrated a marked improvement, highlighting the importance of appropriate model architecture and hyperparameter tuning for effectively learning from imbalanced datasets.

Overall, this analysis suggests that Logistic Regression and SVM are robust and effective models for credit card fraud detection across different data scenarios. Random Forest also exhibits strong potential, particularly in handling imbalanced data. While the Neural Network required further optimization on the balanced subset, its performance improved significantly on the full dataset, indicating its potential with appropriate adjustments.