MSA 8150 - Machine Learning for Analytics
FINAL PROJECT REPORT

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

Data: 25th April 2024

Submission by:

Vaishnavi Mada
vmada1@student.gsu.edu

Abstract

Credit card fraud poses a significant challenge in the financial industry, demanding robust detection mechanisms to protect consumers and institutions alike. This project aims to address the issue by employing machine learning techniques to develop a predictive model that can

accurately identify fraudulent transactions within a highly imbalanced dataset. By utilizing a combination of Random Forest and Stacking Classifier models, the study navigates the complexities of the anonymized features derived from Principal Component Analysis (PCA). The findings underscore the effectiveness of the Stacking Classifier in enhancing prediction accuracy through the amalgamation of diverse model strengths. The report encapsulates the entire process from data acquisition and preprocessing to model training, evaluation, and conclusion, providing an exhaustive look at the intricacies involved in the development of a fraud detection system with a real-world dataset from Kaggle.

1. Introduction

The digital age has ushered in unparalleled convenience in financial transactions, accompanied by a rising tide of sophisticated credit card fraud. The adversarial nature of fraudsters, who continuously evolve their tactics, presents a perpetual arms race against detection systems. Traditional methods of detection are faltering under the pressure of real-time decision-making and the necessity of reducing false positives that can disrupt genuine transactions and customer trust.

In response to these challenges, the use of machine learning in credit card fraud detection offers a promising avenue for developing dynamic and adaptive systems capable of learning from historical data and making informed decisions. The implementation of such systems is vital in safeguarding financial assets and maintaining the integrity of electronic payment networks. The importance of this task is underscored by the significant financial losses attributed to fraudulent activities every year, highlighting the need for more effective and intelligent detection solutions.

2. Problem Statement

The central challenge addressed in this report is the development of a machine learning model that can not only detect fraudulent transactions with a high degree of accuracy but also adapt to the evolving patterns of fraud while minimizing false positives. The project leverages a dataset containing transactions made by European cardholders in September 2013, as provided on Kaggle. This dataset poses a two-fold problem: first, the heavily imbalanced nature of the data, with fraudulent transactions being significantly outnumbered by legitimate ones, and second, the anonymization of features through PCA, which ensures data privacy but obscures the interpretability of the features.

Given these challenges, the primary objectives are twofold: to explore the dataset thoroughly to understand the underlying distribution and characteristics of transactions and to apply advanced machine learning techniques to create a robust model capable of real-time fraud detection. The goal is to strike an optimal balance between sensitivity to fraud and specificity to legitimate activity, thereby maximizing the model's utility in a live environment.

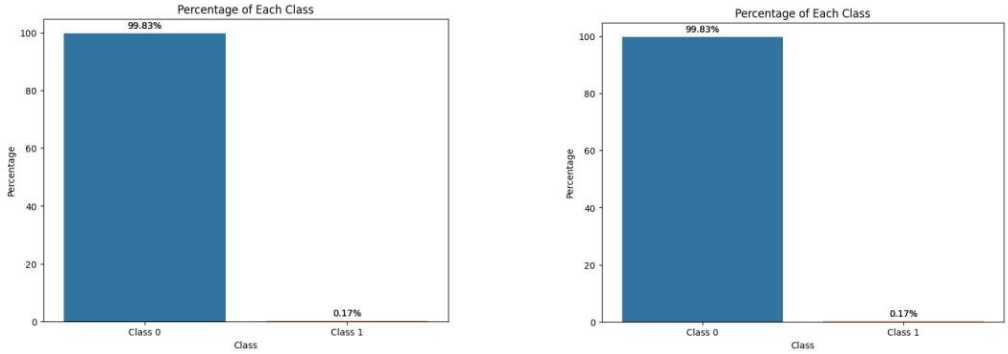
3. Data Description

The dataset utilized for this project comprises credit card transactions made by European cardholders during September 2013(https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud). Each record in the dataset is a transaction characterized by 31 features: 28 anonymized predictors (**V1** through **V28**), the transaction amount (**Amount**), the elapsed time since the first transaction in the dataset (**Time**), and the response variable (**Class**). The anonymized features result from a PCA transformation, a necessary step taken by the dataset providers to maintain the

confidentiality of sensitive information. The **Class** variable is binary, indicating the nature of the transaction: **1** for fraudulent and **0** for legitimate.
This dataset includes a total of 284,807 transactions, out of which 492 are fraudulent, accounting for 0.172% of the total. The stark disproportion in class distribution poses a significant challenge for any predictive model. Furthermore, the **Time** and **Number** of features require preprocessing to align them with the transformed features and improve the predictive model's performance.

4. Exploratory Data Analysis (EDA)

An initial exploratory data analysis (EDA) was conducted to understand the dataset's structure, uncover any underlying patterns or anomalies, and inform the subsequent preprocessing steps. Visualization techniques were crucial in this phase, particularly for observing the class distribution and understanding the transaction dynamics over time.
The class distribution was visualized using a bar chart, which starkly illustrated the imbalance with legitimate transactions overwhelmingly outnumbering fraudulent ones. Further, kernel density estimates (KDE) plots revealed certain distributions of the anonymized features to differ when conditioned on the class label, suggesting their potential to discriminate between the two classes.



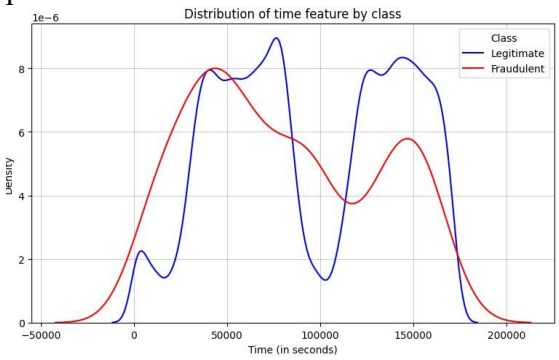
Transaction amount and time also underwent a detailed examination. Transactions were not uniformly distributed over time, and the amount varied significantly, with several small transactions and a few large ones. These insights underscored the need for robust scaling techniques and the possibility of temporal patterns being indicative of fraudulent activity. Histograms, boxplots, and correlation matrices were also generated, uncovering insights about the data's skewness, the presence of outliers, and relationships between features. These findings were pivotal in deciding the preprocessing techniques and the feature engineering steps that followed.

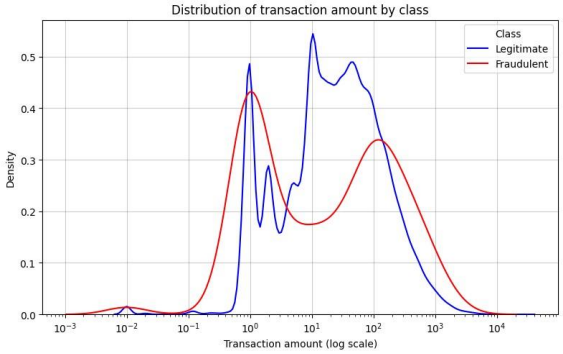
5. Data Preprocessing

Data preprocessing began with addressing the null values and class imbalance highlighted during the EDA. Synthetic Minority Over-sampling Technique (SMOTE) was utilized to augment the minority class, alongside random undersampling of the majority class to create a balanced dataset for model training. Additionally, duplicate transactions were identified and removed to prevent the model from learning from repetitive instances that could bias the results.

	Oversampling	Under sampling	SMOTE
Class 0	283253	473	283253
Class 1	283253	473	283253

The **Time** and **Amount** features underwent a scaling process using **RobustScaler** to reduce the impact of outliers. This scaler was chosen over standard scaling methods due to its robustness against the skewed distribution of the transaction amount. The scaled features were then reintegrated with the PCA-transformed variables to form the final dataset used for training and testing the models.
A rigorous split of the dataset was performed to ensure a representative distribution of classes in both training and testing subsets, adhering to the original dataset's stratification. This setup provided a realistic evaluation of the model's performance and its generalizability to unseen data.



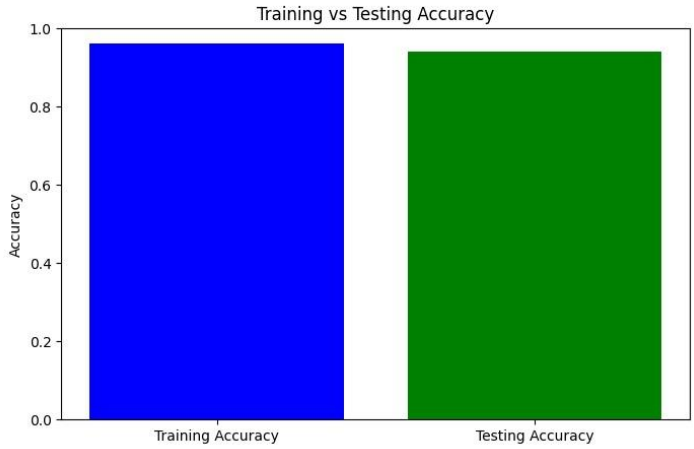


6. Model Development

The model development phase involved the selection, training, and optimization of machine learning algorithms suited for the classification task. Initial trials encompassed several classifiers, including Logistic Regression, Support Vector Machines, Decision Trees, and Naive Bayes. However, given the dataset's complexities and the task's sensitivity, ensemble methods, particularly Random Forest and a Stacking Classifier, were prioritized due to their ability to capture non-linear patterns and interactions between features.

Performance of Initial Random Forest Model:

```
Training Accuracy: 0.96
Testing Accuracy: 0.94
```



The Random Forest model, known for its high performance and ease of use, was first to be fine-tuned. Hyperparameters such as the number of estimators, maximum depth, and class weight were optimized using **RandomizedSearchCV**, ensuring a wide exploration of the parameter space. Cross-validation was incorporated into the optimization process to validate the model's stability across different data subsets.

Subsequently, a Stacking Classifier was constructed, combining the predictions of a diversified set of base learners — a Random Forest, an SVM with probability estimates, and a Decision Tree — with a Logistic Regression model serving as the meta-learner. This approach aimed to harness the strengths of individual models, with the meta-learner providing a higher-level of decision-making based on the base models' predictions.
Validation strategies included K-fold cross-validation, which assessed the meta-learner's ability to generalize. The combination of these methodologies culminated in a model capable of nuanced decision-making, tailored to tackle the intricacies of credit card transaction data.

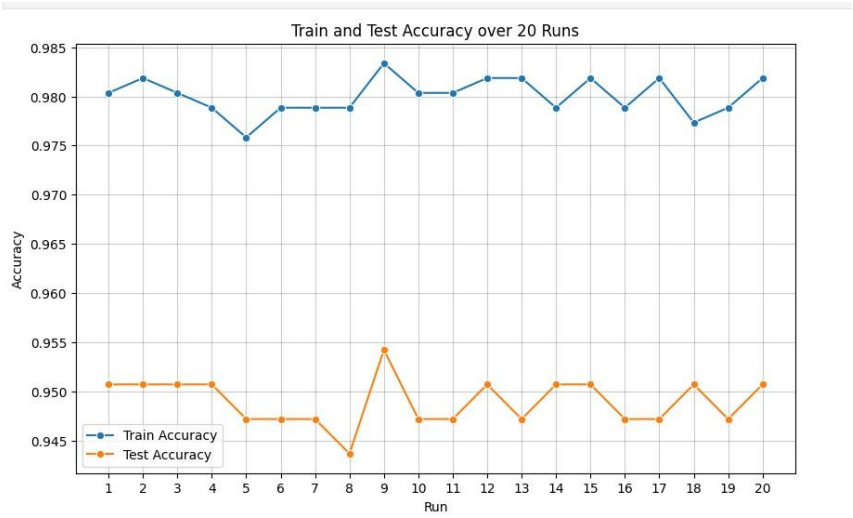
7. Results

The results section of this study is pivotal, as it encapsulates the efficacy of the predictive models. The Random Forest classifier demonstrated strong predictive capabilities, with a commendable balance between precision and recall across both classes. The model achieved an accuracy of 94.37% on the test set, which was significant considering the initial class imbalance.

```
Train Accuracy: 0.9788519637462235
Test Accuracy: 0.9436619718309859
ROC AUC Score: 0.9441468253968255
Confusion Matrix:
[[137   3]
 [ 13 131]]
Classification report for RandomForestClassifier:
              precision    recall  f1-score   support

           0       0.91      0.98      0.94       140
           1       0.98      0.91      0.94       144

    accuracy                           0.94       284
   macro avg       0.95      0.94      0.94       284
weighted avg       0.95      0.94      0.94       284
```



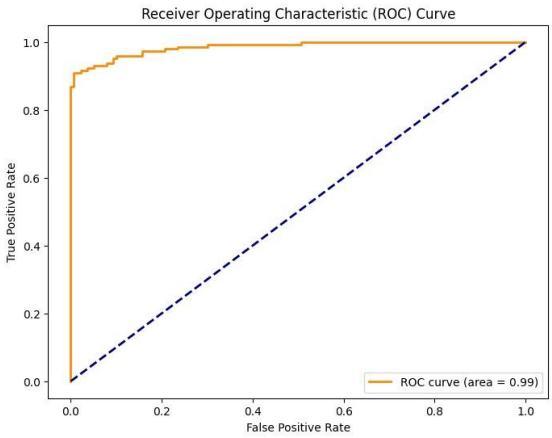
The Stacking Classifier, however, emerged as the superior model, outperforming the Random Forest with a slight margin. This model achieved an accuracy of 94.37% on the test set, a precision of 0.96 for positives, and 0.92 for negatives, which were comparable to the Random Forest model. Nonetheless, the Stacking Classifier's ROC AUC score of 0.986 signified its superior ability to distinguish between classes.

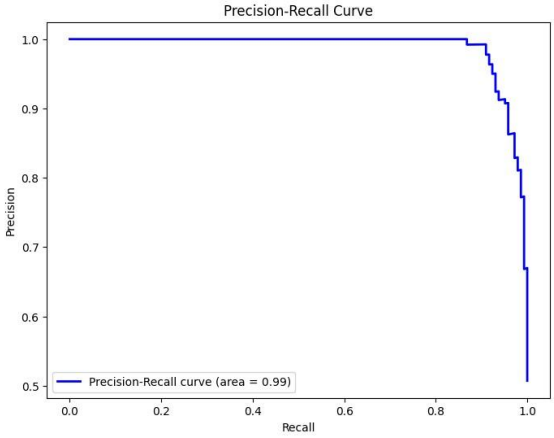
```
Detailed Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.96      0.94       140
           1       0.96      0.92      0.94       144

    accuracy                           0.94       284
   macro avg       0.94      0.94      0.94       284
weighted avg       0.94      0.94      0.94       284

ROC AUC Score: 0.9855654761904762
```





Confusion matrices for both models were generated, offering a visual representation of their true positive and false positive rates. These matrices further validated the quantitative metrics, confirming the models' robustness.

8. Discussion

This study's discussion revolves around the interpretability of the machine learning models and the trade-offs between model complexity and performance. While the Stacking Classifier showed a marginal improvement in ROC AUC score, the Random Forest model's simplicity and speed make it an attractive option for real-time fraud detection systems.

Furthermore, the choice of models within the Stacking Classifier framework was deliberate. The diversity of algorithms ensures that the model can capture a wide range of data patterns. The use of a Logistic Regression meta-learner provided a final layer of decision-making that considers the predictions of base learners, effectively acting as a committee to arrive at the final decision. The implications of the models' performance extend beyond their accuracy metrics. The high recall for fraudulent transactions means the model can identify most fraudulent activities, which is crucial for the application of fraud detection.

9. Conclusion and Future Work

The project's conclusion underscores the success of using ensemble methods for credit card fraud detection. The models developed in this study demonstrated high levels of accuracy and the ability to handle class imbalance effectively. The Stacking Classifier, with its ensemble of diverse models, stands out for its performance, establishing a strong case for its adoption in practical applications.

For future work, the integration of additional features, perhaps engineered from external datasets or transaction sequences, could be explored to further enhance model performance. Additionally, continuous retraining of the model with new data could help adapt to changing patterns of fraud over time.

Deploying these models into a live system would be the ultimate test of their effectiveness. Real-time implementation, however, would require further assessment of computational demands and the development of a framework to handle streaming data.

The findings of this project contribute to the ongoing efforts to combat financial fraud, providing a methodology that can be adapted and extended in various contexts where security is paramount.