Budapest University of Technology and Economics

Faculty of Electrical Engineering and Informatics

Department of Artificial Intelligence and Systems Engineering

Machine Learning Techniques for Detecting Credit Card Fraud in Highly Imbalanced Datasets

Cuadros Rivas, Alejandra Paola

KK5459

Field: Computer Engineering

Specialization: Data Science and Artificial Intelligence

Professor: Antal, Péter

2024/ Winter semester

1. Introduction

This project focuses on developing a predictive model capable of identifying fraudulent activities within a dataset characterized by a severe class imbalance, where fraudulent transactions constitute only 0.172% of the data.

2. Exploratory Data Analysis (EDA)

The dataset contains 284,807 credit card transactions, of which only 492 (~0.17%) are labelled as fraudulent. Features include 30 anonymized variables derived from PCA, as well as Time and Amount. The severe class imbalance underscores the need for tailored preprocessing and balancing techniques.

3. Data Preprocessing

The dataset was preprocessed to enhance analysis and model performance: transaction amounts were scaled using StandardScaler for uniformity, and a Time_of_day feature was created to capture temporal patterns in behaviour. Dimensionality reduction using PCA reduced the dataset from 30 to 17 features while preserving 95% of the variance, simplifying computations without significant performance loss. Finally, the data was split into 80% training and 20% testing subsets to ensure robust evaluation.

4. Balancing Techniques

We addressed the dataset's class imbalance using three techniques: under-sampling, SMOTE, and a hybrid approach. Under-sampling reduced the size of the majority class (non-fraud) to match the minority class (fraud), resulting in a balanced 50/50 distribution in training and testing datasets. While effective, this method risks losing valuable majority class data. To address this, SMOTE was applied to generate synthetic samples for the minority class, preserving feature distributions and ensuring balance in the split datasets. Lastly, a hybrid method combined partial under-sampling with SMOTE, retaining more majority class data while generating realistic synthetic samples for the minority class. This approach achieved balance with minimal data loss and consistent feature integrity.

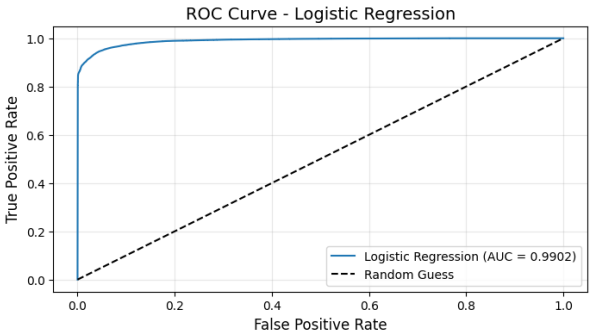
5. Unsupervised Models

We applied the Isolation Forest to detect anomalies (fraud) in the highly imbalanced dataset. The tuned model significantly reduced false positives and improved precision, achieving a balance between detecting fraud and avoiding false alarms. Despite its improvements, the model exhibited limitations in recall and overall class separation, as reflected in the low ROC-AUC score. The tuned model strikes a better balance, improving overall fraud detection reliability.

6. Supervised Models

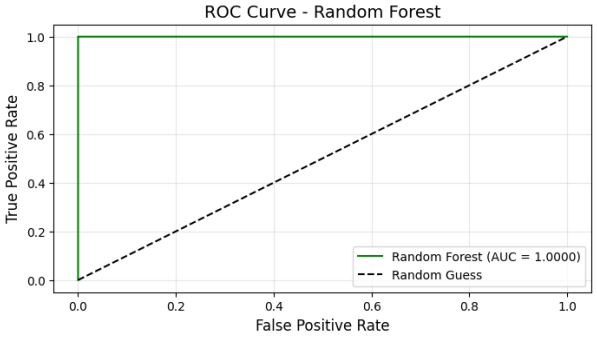
6.1. Logistic Regression

The ROC curve highlights the strong performance of the Logistic Regression model in distinguishing fraud from non-fraud transactions, with an AUC of 0.9902 indicating excellent class separation. The curve's proximity to the top-left corner demonstrates a high true positive rate with a low false positive rate,



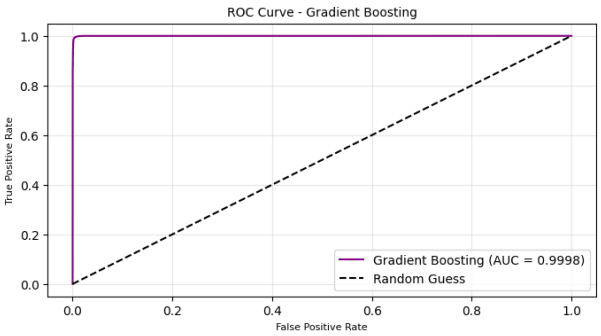
outperforming the random classifier baseline (AUC = 0.5). The confusion matrix shows 52,322 true positives, 55,360 true negatives, 1,503 false positives, and 4,541 false negatives. Precision and recall are balanced for both classes, achieving an F1-score of 95%, but the false negatives remain a concern. Overall, the model performs remarkably well on the SMOTE-balanced dataset.

6.2. Random Forest



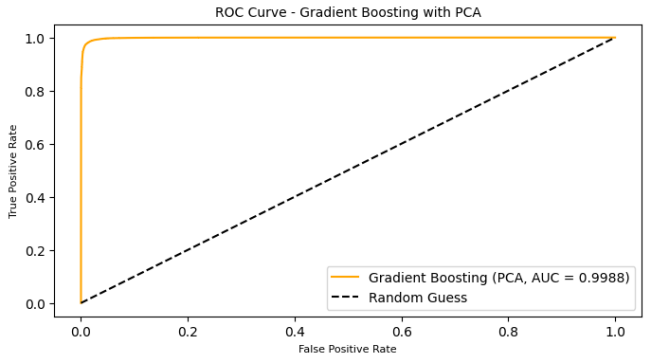
The Random Forest model demonstrates exceptional performance, achieving perfect precision, recall, F1-score, and a ROC-AUC of 1.0000 for both training and test sets. The confusion matrix shows 56,863 true positives, 56,847 true negatives, 16 false positives, and no false negatives, indicating flawless classification of fraud and non-fraud cases. Feature importance analysis highlights V14 as the most critical predictor, with V10, V17, and V4 also contributing significantly. Despite concerns about potential overfitting, identical metrics for training and test sets suggest that the model generalizes well without overfitting, showcasing its robustness and reliability in fraud detection.

6.3. Gradient Boosting



The ROC curve, with an AUC of 0.9998, highlights the model's near-perfect ability to distinguish between fraud and non-fraud transactions. Precision for fraud detection is 99%, and recall is nearly perfect at 100%, slightly outperforming non-fraud recall (99%). The Gradient Boosting model demonstrates exceptional performance, achieving 99% accuracy with a balanced F1-score of 99% for both fraud and non-fraud classes. It correctly classified 56,607 fraud cases and 56,521 non-fraud cases, with minimal errors (342 false positives and 256 false negatives). Feature importance analysis reveals that V14 is the most influential in detecting fraud, suggesting it captures key fraud patterns, while features like V4, V12, and V8 play smaller roles. Despite slightly more errors compared to Random Forest, Gradient Boosting maintains high confidence and interpretability, excelling at identifying critical fraud patterns.

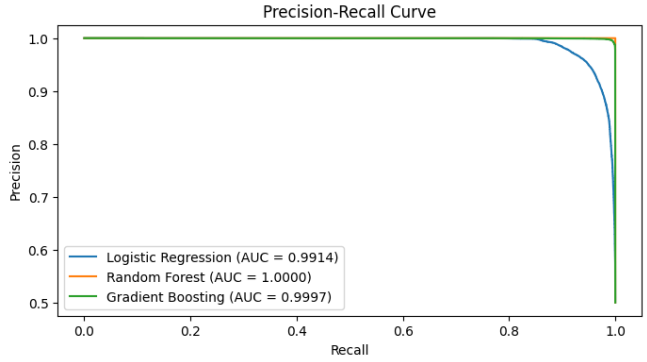
6.4 Dimensionality Reduction Impact



The PCA-reduced model successfully reduced the number of features from 30 to 17 while retaining 95% of the dataset's variance, preserving most of its information. It achieved a strong performance, with 99% precision, 97% recall, and an F1-score of 98% for fraud detection. The confusion matrix shows 55,371 true positives and 56,357 true negatives, with slightly more errors (506 false positives and 1,492 false negatives) compared to the full-featured model. The ROC-AUC score of 0.9988, though marginally lower than the full-featured model's 0.9998, still indicates excellent discriminatory power. While there is a slight drop in accuracy, the significant dimensionality reduction improves computational efficiency and reduces the risk of overfitting.

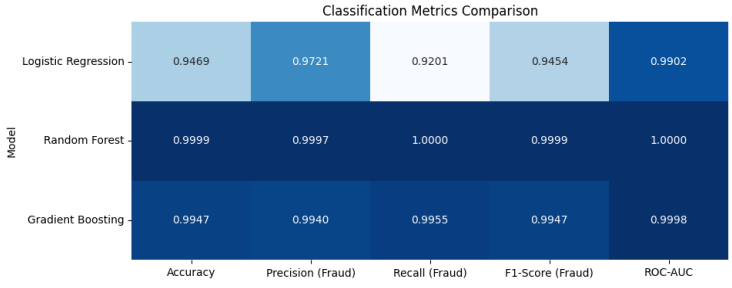
7. Model Evaluation

7.1. Precision-Recall Curve (AUPRC)



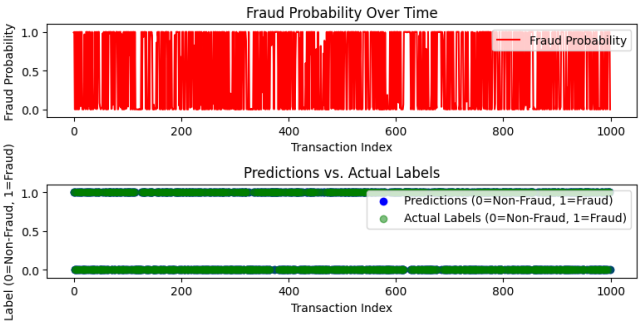
The models highlight the trade-off between precision and recall, with the Random Forest achieving near-perfect performance and an AUC of 1.0000, demonstrating exceptional ability to distinguish between fraud and non-fraud transactions. Gradient Boosting follows closely with an AUC of 0.9997, showcasing robust performance, while Logistic Regression, with an AUC of 0.9914, performs slightly lower, reflecting its limitations in handling complex decision boundaries compared to ensemble methods.

7.2. Classification Metrics



Random Forest achieves perfect performance across all metrics(accuracy, precision, recall, F1-score, and ROC-AUC)indicating potential overfitting and raising concerns about generalizability to unseen data. Gradient Boosting offers a more balanced performance, with slightly lower recall and F1-score compared to Random Forest, making it a robust yet less extreme model. Logistic Regression delivers competitive results with strong recall and F1-score but falls behind the ensemble models due to its linear nature, which limits its ability to capture complex relationships in the data.

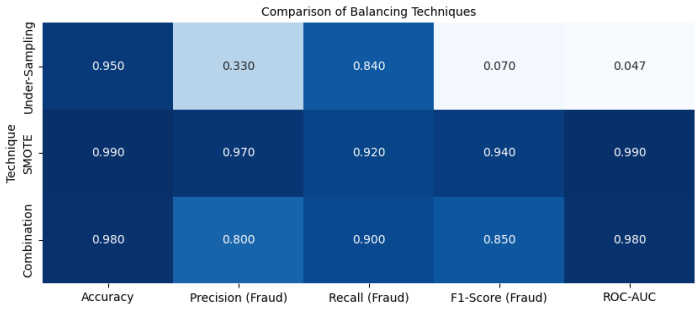
8. Real-Time Simulation



The simulation, performed on 1,000 transactions, evaluated the fraud detection system's performance. The confusion matrix revealed 509 true negatives, 487 true positives, 4 false positives, and 0 false negatives, showcasing the system's accuracy. The classification report highlighted a precision of 99% for fraudulent transactions, 100% recall, and a near-perfect F1-score of 1.00, with an overall accuracy of 99.6%. These results demonstrate the system's exceptional ability to detect fraud while maintaining high precision, with minimal errors caused by 4 false positives, reflecting its robustness and reliability.

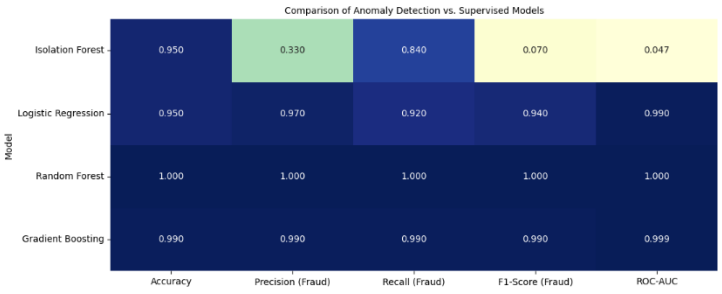
9. Comparison of Results

9.1. Balancing Techniques



Under-sampling achieves 95% accuracy but struggles with fraud detection, with low precision (33%) and an imbalanced F1-score (7%), despite a high recall (84%). Its ROC-AUC score of 0.047 reflects weak performance in distinguishing between classes. SMOTE significantly improves performance, with precision at 97% and recall at 92%, achieving an excellent F1-score and a robust ROC-AUC of 0.99. The combination method offers a balanced approach, with precision at 80%, recall at 90%, an F1-score of 85%, and a slightly lower ROC-AUC of 0.98, demonstrating solid overall performance while maintaining balance.

9.2. Unsupervised Models vs. Supervised Models



Isolation Forest achieves 95% accuracy but struggles with fraud detection, showing low precision (33%) and F1-score (7%), despite high recall (84%). Its ROC-AUC of 0.047 reflects poor performance in distinguishing fraud from non-fraud. Logistic Regression offers balanced results with high precision (97%), recall (92%), and an F1-score of 94%, supported by a robust ROC-AUC of 0.99. Random Forest delivers perfect metrics across all categories, but this suggests potential overfitting. Gradient Boosting performs nearly as well, with slightly lower precision and recall compared to Random Forest, maintaining strong performance with an ROC-AUC of 0.999. These results suggest that SMOTE for balancing and Gradient Boosting for modelling may provide the best trade-off between performance and reliability.

10. Conclusion

- SMOTE proved to be the most effective balancing technique, significantly enhancing model performance metrics such as precision, recall, F1-score, and ROC-AUC. It retained critical information from the majority class while effectively augmenting the minority class.
- Under-sampling, although simple, resulted in severe information loss, leading to poor performance metrics and limited applicability.
- Combination techniques offered a balanced approach but fell short of SMOTE in critical metrics like precision and F1-score, making them less effective overall.
- Gradient Boosting demonstrated robust performance across all metrics when paired with SMOTE, achieving near-perfect precision, recall, and F1-score while maintaining strong generalization and avoiding overfitting.
- Random Forest also delivered exceptional results but showed signs of overfitting, as evidenced by its perfect metrics in both training and testing phases.
- Logistic Regression provided a strong baseline performance but lagged behind tree-based models in precision and recall, limiting its effectiveness for complex fraud detection tasks.
- Anomaly detection methods like Isolation Forest, despite their conceptual appeal, underperformed compared to supervised models in precision and overall effectiveness.
- PCA effectively reduced feature dimensionality without significant performance loss, highlighting the potential for further simplifying the dataset without sacrificing predictive power.
- SHAP analysis identified key features like V14, V4, and V12 as the most critical contributors to fraud detection, enhancing the model's interpretability and transparency.
- LIME offered local explanations for individual predictions, providing valuable insights into why specific transactions were classified as fraud or non-fraud.
- The implementation of a real-time fraud detection pipeline demonstrated the feasibility of deploying these models in live environments, achieving near-perfect accuracy during simulations.
- Future work includes exploring dynamic fraud detection models, cost-sensitive active learning, ensemble optimization, expanded explainability methods, and real-world testing to enhance adaptability, precision, and scalability of the system.

11. Bibliography

[1] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *Ann. Statist. 29,* vol. 5, pp. 1189 - 1232, October 2001.

[2] K. W. Bowyer, N. V. Chawla, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research,* vol. 16, p. 321–357, June 2002.

[3] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *Plos ONE,* pp. 1-21, 4 March 2015.

[4] M. B. A. McDermott, L. H. Hansen , H. Zhang, G. Angelotti and J. Gallifant, "A Closer Look at AUROC and AUPRC under Class Imbalance," 2024.