CREDIT CARD FRAUD DETECTION AND CREDIT RISK ANALYSIS: MACHINE LEARNING APPROACHES

ANKITA ANAND * SIDDHARTHA P. CHAKRABARTY †

Abstract

This study explores advancements in techniques for detection of credit card fraud detection and credit risk analysis using machine learning approaches. Two distinct datasets were analyzed: a credit card fraud dataset, addressing extreme class imbalance, and the Australian credit dataset for credit risk prediction. Methods like SMOTE and SMOTE-ENN were employed to tackle data imbalance, while ensemble techniques such as Bagging, Boosting, and LSTM classifiers enhanced model accuracy. Results demonstrated notable improvements, with the LSTM ensemble achieving a PRAUC of 0.8269 for fraud detection and Boosting achieving 87.5% accuracy for credit risk prediction. The findings emphasize the importance of robust data pre-processing and ensemble learning in financial data analytics.

Keywords: Fraud Detection; Credit Risk Analysis; SMOTE; Ensemble Learning

JEL Classification: C63

1 INTRODUCTION

As the banking sector has diversified from traditional commercial banking to investment banking, the credit risk exposure of the financial institutions, including banks has become more diversified amd complex. In traditional commercial banking, which predominately involved acceptance of deposits and disbursement of loans, the credit risk exposure was limited to default on loans. But with the investment bank facilitated development of derivative markets many more forms of credit risk exposure (such as counterparty credit risk) has emerged [1]. In this work, we will focus on two critical challenges faced by financial institutions, as a part of commercial banking activities, namely, credit card fraud and credit risk analysis. While the former is characterized by unauthorized transactions on credit cards, the latter emerges from default on loans and its likelihood.

Traditional approaches to fraud detection and credit risk management, such as rule-based systems and risk scoring mechanisms, have limitations. Rule-based systems, while being transparent and easy to implement, is often inadequate in adapting to the dynamic and evolving tactics of perpetrators, resulting in high false positive rates and as well as inefficiencies [2]. Similarly, conventional statistical models used for credit risk analysis is not always adept at handling complex and large-scale datasets, thereby limiting their predictive accuracy.

*Mehta Family School of Data Science and Artificial Intelligence, Indian Institute of Technology Guwahati, Guwahati-781039, India, e-mail: ankita.anand@iitg.ac.in

†Mehta Family School of Data Science and Artificial Intelligence and Department of Mathematics, Indian Institute of Technology Guwahati, Guwahati-781039, India, e-mail: pratim@iitg.ac.in, Phone: +91-361-2582606

Accordingly, Machine Learning (ML) has emerged as an alternative tool to address these challenges, via its ability to handle big datasets, identify patters, thereby leading to efficient predictive algorithms. In the specific context of our work, for fraud detection, ML algorithms can analyze transaction data to detect anomalies, leading to significant reduction of false positives while at the same time enhancing security [2, 3]. Further, in the case of credit risk analysis, ML models can be adopted to make an assessment of borrower profiles and also predict default risks (with greater precision) via the incorporation of sophisticated algorithms and feature engineering techniques.

This study aims to advance solutions approaches for both the problems, namely credit card fraud detection and credit risk analysis, by adoption of novel ML techniques and ensemble methods. For credit card fraud detection, we address the inherent class imbalance problem by employing data balancing techniques such as SMOTE (Synthetic Minority Oversampling Technique) and SMOTE-ENN (SMOTE with Edited Nearest Neighbor), as well as ensemble learning strategies like EasyEnsemble and BalanceCascade [4]. For credit risk prediction, we focus on optimizing model performance using feature selection methods, including Information Gain (IG), and advanced ensemble models like Boosting and Stacked Classifiers [5].

The results demonstrate significant improvements in performance metrics across both tasks. For fraud detection, ensemble models like the Long Short-Term Memory (LSTM) ensemble achieved high Precision-Recall Area Under the Curve (PRAUC) scores, showcasing their ability to effectively handle imbalanced datasets. In credit risk prediction, ensemble methods such as Boosting delivered superior accuracy, highlighting their potential in real-world financial applications. This work underscores the importance of integrating data pre-processing techniques with advanced ML models to enhance predictive capabilities in financial systems.

Various techniques, ranging from rule-based systems to advanced ML methods, have been developed to address these challenges. Bhatla et al. [2] provided an early foundation in fraud detection by outlining methods such as rule-based systems, risk scoring, and Address Verification Service (AVS). These traditional approaches were augmented by advancements in technologies like Europay, Mastercard, and Visa (EMV) and biometric authentication, which introduced higher security standards for transaction verification.

In the context of fraud detection with the ability to analyze complex patterns in large datasets, ML has emerged as a predominant driver. Varmedja et al. [3] evaluated algorithms such as logistic regression, Random Forest (RF), and Multilayer Perceptron (MLP), highlighting RF as particularly effective due to its high precision and recall. To address class imbalance, the SMOTE method was employed, which improved upon the performance of models, via the generation of synthetic samples for the minority class. Mqadi et al. [6] demonstrated the efficacy of SMOTE in enhancing the performance of algorithms like Decision Trees and RF, further solidifying its role as a key oversampling method.

Undersampling techniques have also been explored in order to manage imbalanced datasets. Liu et al. [4] proposed EasyEnsemble and BalanceCascade, which iteratively undersample majority data and combine classifier predictions, thereby improving the detection rates. Similarly, ensemble learning approaches have shown significant promise. Esenogho et al. [7] developed a neural network ensemble combining LSTM with AdaBoost and SMOTE-ENN resampling, achieving remarkable sensitivity and specificity metrics. For credit risk analysis, Pandey et al. [8] demonstrated that Extreme Learning Machines (ELM) outperformed traditional models like

Bayesian Networks and Support Vector Machines (SVM), achieving 96.33% accuracy on a German credit dataset. Ensemble techniques such as Bagging and Boosting further enhanced model robustness and stability.

Finally, stacked classifiers have been proposed as robust solutions for complex datasets. Ileberi et al. [5] combined algorithms such as RF, Gradient Boosting, and XGBoost with feature selection using IG, achieving an AUC of 0.870. This approach exemplifies the potential of aggregating predictions from multiple base learners to improve accuracy, making it an amenable choice for imbalanced and challenging datasets.

2 METHODOLOGY

In this section, we present the details about the datasets and. techniques to be adopted for data pre-processing, in addition to the description of modeling approaches and methods of feature selection, which will be employed in the course of the study carried out in this paper.

2.1 DATASETS

For the purpose of the study carried out in this paper, we have made use of two datasets, on which we have applied the proposed methodologies.

The first dataset pertains to credit card fraud and is available at "`https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud`". This dataset (which is highly imbalanced) contains $284,807$ transactions, of which only 492 were labeled as fraudulent. The dataset includes anonymized features, which were transformed using the Principal Component Analysis (PCA), along with additional attributes, such as transaction time and amount involved. The dataset highlights the challenge of detecting in a real-world scenario where fraudulent cases are relatively rare.

The second dataset is the Australian credit dataset, which is available at "`https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/australian/australian.dat`". This dataset maintained at the University of California Irvine (UCI) Machine Learning Repository, is balanced containing 690 instances with 14 attributes. These attributes cover a range of borrower characteristics, such as age, employment status, and financial history, all of which aids in the prediction of credit risk.

2.2 CREDIT CARD FRAUD DETECTION

In this sub-section the steps used for credit card fraud detection is outlined in the form of Algorithm 1.

(1) The data pre-processing involved the normalizing the numerical feature `Amount`, by making use of the standard scaling, thereby ensuring consistent feature contributions. The stratified splitting is done for the train, test and validation split. Stratified splitting means that during the generation of a training/validation dataset split, attempt will be made to keep the the same percentages of classes in each split.

(2) The SMOTE and SMOTE-ENN [7] were applied in order to balance the dataset, via generation of synthetic samples for the minority class and the removal of noisy examples from the majority class.

(3) The model training was done by implementing RF, Decision Tree, Naive Bayes (NB), SVM, and MLP for

the baseline comparisons. Further, we developed an LSTM ensemble [7] to exploit the sequential patterns in transactional data, thereby improving the robustness against imbalance.

(4) The evaluation was carried out using metrics such as PRAUC, F1-score, and accuracy was computed, with a focus on PRAUC, in order to evaluate the performance based on the imbalanced dataset [9].

Algorithm 1 Credit Card Fraud Detection Methodology

Input: Credit Card Fraud Dataset.
Step 1: Data pre-processing
Removal of duplicates and feature scaling of `amount`.
Usage of stratified splitting for train-test split.
Step 2: Combating imbalance in training dataset
Methods used:
(i) Random oversampling imbalanced dataset.
(ii) SMOTE [6].
(iii) Combining SMOTE with Tomek links [10].
(iv) SMOTE-ENN [7].
Output: Balanced dataset.
Step 3: Models used for training on pre-processed data
(A) RF Model with:
(i) Trained **RF** model on unbalanced pre-processed training data for baseline comparison.
(ii) RF with random oversampling.
(iii) RF with SMOTE oversampling.
(iv) RF with SMOTE oversampling and Tomek links under sampling, for removing samples of data from the majority class which has the lowest Euclidean distance with the minority class data.
(v) RF with SMOTE oversampling and ENN [7].
(B) Ensemble Models used:
(i) Easy Ensemble with AdaBoost [4].
(ii) Balance cascade with AdaBoost [4].
(iii) LSTM ensemble classifier [7].

2.3 Credit Risk Prediction

In this sub-section the steps used for credit risk protection is described in the form of Algorithm 2.

(1) In order to carry out data pre-processing, feature selection was conducted by making used of IG, so as to identify and retain the most predictive attributes, thereby reducing the feature space for improved efficiency.

(2) The model training step involved the evaluation of classifiers, including RF, Gradient Boosting, and XGBoost. A stacked classifier was built by leveraging RF, Gradient Boosting, and XGBoost thereby improving the predictive accuracy[5] and also handling the complex interactions between the features.

(3) In the evaluation stage, the Accuracy, F1-score, and AUC metrics were computed, in order to compare the individual models and the stacked classifier, with a specific focus on the imbalanced datasets.

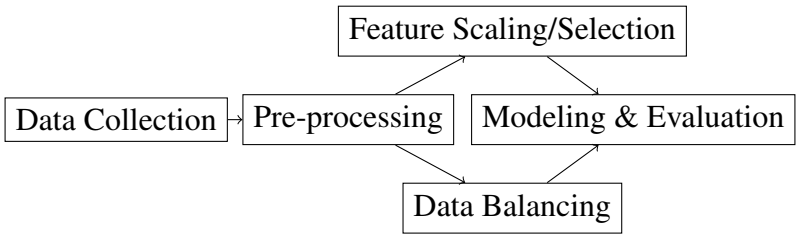
Algorithm 2 Credit Risk Prediction Methodology

Input: Australian Credit Dataset. This is a balanced dataset.
(A) Bagging and Boosting
Step 1: Data pre-processing
Scaling the features and doing train-test split.
Step 2: ML Models used for training
Used fine-tuned classifiers by setting hyperparameters [8].
Classifiers used:
(i) NB
(ii) Decision Tree
(iii) K- Nearest Neighbour (KNN)
(iv) SVM
(v) MLP
Ensemble models for Australian dataset:
(i) Bagging
(ii) Boosting
(B) Stacking
Used stacking with RF, Gradient Boosting, and XGBoost classifiers, along with IG based feature selection [5].
(i) Enhanced feature selection with IG feature selection, with the threshold being set at 30%.
(ii) Enhanced stacked model with hyperparameter tuning and class balancing
(iii) RF, Gradient Boost and XGBoost used as base estimators
(iv) Used RF as final estimator and defined stacked model using Scikit-learn library.

2.4 Summary of Methodology

For the problem on fraud detection, the SMOTE-ENN combined with RF and LSTM ensemble model demonstrated improved detection rates in case of rare fraudulent transactions. For credit risk prediction problem, advanced ensemble methods, particularly the stacked classifier, outperformed the baseline models, thereby confirming the effectiveness of feature selection and ensemble learning in the handling of structured datasets. For the implementation, the ML models that were adopted, included the NB, Decision Trees, SVM, RF and LSTM Ensemble. Finally, in order to enhance model efficiency and interpretability, feature selection was employed using IG. This method evaluated the contribution of each feature to the target variable, whereby only the most predictive features were retained. The reduction of feature space, led to the models achieving improved computational efficiency without compromising the accuracy levels. The IG Feature Selection (IG-FS) involved three key steps,

namely, calculation of IG, application of IG threshold, and the creation of a ranked subset, with selected features for the purpose of training the model. A schematic representation of the methodology adopted in this work is presented below.



3 Results

3.1 Results on Credit Card Fraudulent Transaction Detection

We begin with the presentation of the results pertaining to the problem of fraud detection in credit card transactions, from three perspectives as follows.

(1) Results of models on imbalanced dataset: The initial RF model, when applied to the imbalanced dataset (without the adoption of any sampling techniques), achieved interesting performance metrics. The model analysis resulted in an accuracy level of 0.9996, indicating that it succeeded in classifying the majority of transactions correctly. The F1 Score, which balances precision and recall, was 0.8549, reflecting a not so strong performance for both the classes. It may be noted that the majority of the samples belonged to one class. However, the PRAUC, which is particularly useful in scenarios of class imbalance, was 0.8011. This showed that since almost 99.8% of the data belonged to the majority class, which represents the non-fraudulent transactions. Therefore accuracy is an ineffective metric to judge the model, and there was still scope for improvement in terms of capturing the minority class, in a more effective manner. This result serves as a baseline for comparing the effectiveness of various advanced sampling techniques. RF was used as the ML model in baseline and after data balancing, because it outperformed Logistic Regression (LR), NB, and MLP in terms of performance, as demonstrated in [3]. The results pertaining to the RF model based analysis are presented in Table 1.

Metric	**Score**
Recall	0.7676
Precision	0.9646
F1 Score	0.8549
Accuracy	0.9996
PRAUC	0.8011

Table 1: RF Model on Initial Dataset without any Sampling

(2) Results of models on over-sampled dataset: The results from evaluation of various oversampling techniques, as presented in Table 2 reveals the following. The SMOTE oversampling achieved the highest PRAUC of

0.8267, surpassing the baseline by 0.0256. This improvement indicates that SMOTE, which generates synthetic samples for the minority class, effectively enhanced the model's ability to capture fraudulent transactions. The approach via SMOTE plus ENN also showed strong performance with a PRAUC of 0.8262, an improvement of 0.0251 over the baseline. This technique combines SMOTE with ENN to refine the dataset by removing noisy examples, thus further improving the performance. In summary, both the SMOTE oversampling and SMOTE plus ENN methods improved the PRAUC as compared to the baseline, thereby highlighting their effectiveness in addressing the class imbalance.

Random Forest with	**Recall**	**Precision**	**F1**	**Accuracy**	**PRAUC**
Random Oversampling	0.7324	0.9541	0.8287	0.9995	0.8253
SMOTE Oversampling	0.7394	0.9211	0.8203	0.9995	0.8267
SMOTE + Tomek	0.7464	0.9298	0.8281	0.9994	0.8173
SMOTE + ENN	0.7465	0.9217	0.8249	0.9995	0.8262

Table 2: Random Forest with Various Sampling Techniques

(3) Ensemble model metrics analysis: The analysis of the performance of ensemble models is detailed in Table 3. Among these models, the LSTMEnsemble achieved a PRAUC of 0.8269, surpassing the baseline PRAUC and showing a notable improvement. This model also achieved a high accuracy of 0.9995, indicating its effectiveness in correctly classifying both fraudulent and non-fraudulent transactions. In contrast, the BalanceCascade and EasyEnsemble methods had very low PRAUC scores of 0.0008, in spite of having high recall rates. These models showed limited success in distinguishing between the classes, which can possibly be attributed to their inability to effectively handle the class imbalance. Overall, the LSTMEnsemble model demonstrated superior performance in terms of PRAUC and accuracy, making it the most effective among the ensemble techniques evaluated.

Ensemble Model	**Recall**	**Precision**	**F1**	**Accuracy**	**PRAUC**
BalanceCascade	0.8803	0.0413	0.0788	0.9657	0.0008
EasyEnsemble	0.8803	0.0776	0.1427	0.9824	0.0008
LSTMEnsemble	0.7535	0.9224	0.8295	0.9995	0.8269

Table 3: Ensemble Models with Different Techniques

3.2 Results on Credit Risk Prediction

This section examines the results obtained from implementing ML models on the Australian credit dataset. The focus is on individual models, ensemble techniques, and their fine-tuned versions to compare their performance with baseline results from Pandey et al.[8] and Emmanuel et al.[5].

(1) Individual model performance: The individual models were fine-tuned to improve accuracy on the Australian dataset, with their results presented in Table 4. The NB achieved an accuracy improvement, outperforming the

baseline result from Pandey et al.[8], with a final accuracy of 0.8309. Other models, such as Decision Tree and MLP, showed comparable performance but did not fully reach the accuracy levels reported by Pandey et al[8]. SVM performed exceptionally well, closely matching Pandey et al.'s baseline with an accuracy of 0.8600. These results highlight the improvements achieved through fine-tuning, although some models remained sensitive to the data's class balance and feature distributions.

Model	**Fine-tuned Accuracy**	**Accuracy [8]**
NB	0.8309	0.7826
Decision Tree	0.8420	0.9072
KNN	0.8240	0.8910
SVM	0.8600	0.8594
MLP	0.8150	0.8695

Table 4: Comparison of Fine-tuned and Results for Individual Models on the Australian Dataset [8]

(2) Ensemble model performance: Ensemble methods like Bagging and Boosting improve model accuracy by combining the predictions of multiple base models. Bagging (Bootstrap Aggregating) involved the creation of multiple subsets of the dataset by randomly sampling with replacement and then training separate models on each subset. The final prediction is made by averaging (for regression) or voting (for classification) across all models. This technique reduces variance and helps prevent overfitting. Boosting is an iterative technique that builds a sequence of models, where each new model corrects errors made by previous ones. Each instance in the dataset is given a weight, and misclassified instances receive higher weights in subsequent models, so that the ensemble focuses more on difficult cases. The final prediction is based on the combined output of all models. Boosting improves model performance by reducing both bias and variance. In addition to individual classifiers, ensemble techniques such as Bagging and Boosting were evaluated on the Australian dataset to further enhance predictive power. As shown in Table 5, Boosting achieved the highest fine-tuned accuracy of 0.8750, outperforming the individual models and demonstrating the effectiveness of ensemble methods for credit risk prediction. Bagging also provided a notable improvement, with an accuracy of 0.8650. These results confirm the utility of ensemble models in increasing classification stability and accuracy on moderately balanced datasets like the Australian credit dataset.

Ensemble Model	**Fine-tuned Accuracy**
Bagging	0.8650
Boosting	0.8750

Table 5: Fine-tuned Ensemble Model Results for Australian Dataset

(3) Stacked model and feature selection performance: The stacked model in this study [5] uses IG Feature Selection (IG-FS) to reduce the feature space by ranking and retaining the most predictive features above a set

threshold. Three base models, namely, RF, Gradient Boosting, and XGBoost are fine-tuned and stacked with a final RF estimator, incorporating balanced class weights to handle class imbalance. The model details are as follows:

(a) Base Estimators:

- RF: `n_estimators=200`, `max_depth=10`, `class_weight='balanced'`, `random_state=42`
- Gradient Boosting (GB): `n_estimators=200`, `learning_rate=0.05`, `max_depth=5`, `random_state=42`
- XGBoost (XGB): `n_estimators=200`, `learning_rate=0.05`, `max_depth=5`, `scale_pos_weight=2`, `random_state=42`

(b) Final Estimator: Random Forest with `n_estimators=200`, `max_depth=10`, `class_weight='balanced'`, `random_state=42`

A comparison with the results given in Table 6, the stacked model performed well with an accuracy of 0.8454 and an F1-score of 0.8470, closely matching the results of Emmanuel et al. [5]. Although AUC was slightly lower than Emmanuel's reported 0.9340, this model showed strong precision and recall, especially for the minority class, underscoring the effectiveness of stacking and IG-FS for credit risk prediction.

Metric	**Model Results**	**Paper Results [5]**
Accuracy	0.8454	0.8623
F1-Score	0.8470	0.8458
AUC	0.8532	0.9340
Precision (0)	0.92	Not provided
Recall (0)	0.82	Not provided
Precision (1)	0.76	Not provided
Recall (1)	0.89	Not provided

Table 6: Comparison of Model and Paper Results for the Australian Dataset [5]

4 Conclusion and Future Directions

This study examined credit card fraud detection and credit risk prediction using ML models on imbalanced and moderately balanced datasets, respectively. For the credit card fraud detection problem, RF combined with various oversampling techniques such as SMOTE and SMOTE + ENN improved the PRAUC and F1 scores. The LSTMEnsemble model outperformed other ensembles in handling the class imbalance, achieving a high PRAUC and demonstrating effectiveness in accurately distinguishing between fraudulent and legitimate transactions. For credit risk prediction prediction, fine-tuned individual models and ensemble techniques were applied to the Australian dataset to enhance predictive accuracy. Boosting proved to be the most effective, achieving an accuracy of

0.8750, with SVM and other individual models also showing strong performance. Ensemble techniques demonstrated their value in improving stability and accuracy, with model results closely matching those reported in related studies. This analysis highlights the role of model choice, feature selection, and ensemble methods in producing robust credit risk predictions.

In summary, the findings of this study highlights the effectiveness of ensemble learning and advanced preprocessing techniques. In fraud detection, the LSTM ensemble's ability to capture temporal patterns was pivotal, while SMOTE-ENN effectively addressed the issue of data imbalance. For credit risk prediction, Boosting algorithms demonstrated their strength in reducing bias and variance. On the other hand, challenges included computational intensity, particularly for deep learning models like LSTMs, and the risk of overfitting in oversampled datasets. Future research should explore optimization techniques to mitigate these issues. Future work could focus on real-time deployment, incorporating additional behavioral features, and exploring hybrid models that combine ML with rule-based systems. Additionally, continual model retraining with updated datasets would ensure adaptability to evolving financial threats.

REFERENCES

[1] J. C. Hull, *Risk management and financial institutions,+ Web Site*, vol. 733. John Wiley & Sons, 2012.

[2] T. P. Bhatla, V. C. Prabhu, and A. Dua, "Understanding credit card frauds," in *Rule-Based Fraud Detection Systems*, 2003.

[3] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, "Credit card fraud detection - machine learning methods," 03 2019.

[4] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 2, pp. 539–550, 2009.

[5] E. Ileberi, Y. Sun, and Z. Wang, "A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method," *Journal of Big Data*, vol. 11, 02 2024.

[6] N. Mqadi, N. Naicker, and T. Adeliyi, "A SMOTE based oversampling data-point approach to solving the credit card data imbalance problem in financial fraud detection," *International Journal of Computing and Digital Systems*, vol. 10, pp. 277–286, 02 2021.

[7] E. Esenogho, I. D. Mienye, T. G. Swart, K. Aruleba, and G. Obaido, "A neural network ensemble with feature engineering for improved credit card fraud detection," *IEEE Access*, vol. 10, pp. 16400–16407, 2022.

[8] T. Pandey, A. Jagadev, S. Mohapatra, and S. Dehuri, "Credit risk analysis using machine learning classifiers," in *Credit risk analysis using machine learning classifiers*, pp. 1850–1854, 08 2017.

[9] 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*, vol. 10, no. 3, p. e0118432, 2015.

[10] R. A. A. Viadinugroho, “Imbalanced classification in python: SMOTE-Tomek Links method combining SMOTE with Tomek Links for imbalanced classification in Python,” 2023.