Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12, 124-147
2025
Publisher: Learning Gate
DOI: 10.55214/2576-8484.v9i12.11297

A comparative analysis of ensemble learning models in credit scoring and loan default prediction

Teerath Kumar[1,2*], Sudarshan Poojary[1,3], Raja Vavekanand[3], Fida Hussain Dahri[4], Asif Ali Laghari[5]
[1]Dublin Business School, Dublin, Ireland; teerathkumar.menghwar@atu.ie (T.K.).
[2]School of Computing, Atlantic Technology University, Ireland.
[3]Datalink Research and Technology Lab, Islamkot, 69240, Pakistan.
[4]School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China.
[5]Software Collage, Shenyang Normal University China; Asiflaghari@synu.edu.cn (A.A.L.).

Abstract: This article investigates the effectiveness of three ensemble learning techniques, stacking, bagging, and boosting, in predicting credit scores and loan defaults using three distinct datasets. We compare the models based on accuracy, precision, recall, and F1 score to assess their performance in both binary and multi-class classification tasks. Among the models, stacking achieved the highest overall performance, with a multi-class credit scoring accuracy of 82%, compared to 72% for binary classification using bagging. However, bagging performed less effectively in predicting loan defaults. Boosting, while generally less effective in handling imbalanced data and complex multi-class problems, still produced acceptable results in certain scenarios. The findings suggest that stacking and bagging are particularly well-suited for credit scoring and loan default prediction, making them valuable tools for financial institutions. The study also highlights the importance of addressing class imbalance and applying feature engineering to enhance model performance. Future research should focus on improving model explainability and developing advanced techniques to handle data complexity.

Keywords: *Bagging, Boosting, Classification models, Credit scoring, Ensemble learning, Financial risk prediction, Loan default prediction, Model performance evaluation, Predictive analytics, Stacking.*

1. Introduction

Credit scoring is one of the critical components of the world economy since it forms the basis for evaluating the ability of a borrower to pay back the credit. It has a direct bearing on the flow of credit to people and firms and therefore affects personal loans, mortgages, and business financing [1]. In the past, credit scoring models have used statistical analysis on past financial information using logistic regression to evaluate the risk of extending credit. These models, although adopted in the early years of the development of financial instruments, have been criticized over time because of their inability to respond to the dynamic nature of modern financial markets [2]. Credit scoring has evolved alongside data analytics and machine learning. The growth of big data and the transformation of financial services into digital entities have signaled the need for improved models. Banks today work with large volumes of data, including standard financial data and new sources such as social media activity and transaction history [3]. This has led to the creation of more sophisticated models that are able to handle larger data sets and discover subtle relationships that simple models might miss.

Over the last few years, deep learning has become the go-to approach to overcome the drawbacks of conventional credit scoring techniques. There are other techniques, such as CNNs and RNNs, that have the capability of enhancing accuracy by establishing intricate relations in the sets [4]. However, these powerful methods have their limitations, especially in terms of interpretability of the model and fairness. Such limitations are being addressed by ensemble learning, which is a technique of using several models

History: Received: 6 October 2025; Revised: 14 November 2025; Accepted: 18 November 2025; Published: 3 December 2025
* Correspondence: teerathkumar.menghwar@atu.ie

to improve the predictability of the result and the stability of the model. It is observed that credit scoring can be enhanced in terms of accuracy and robustness, as well as fairness, by integrating deep learning with other ensemble methods like bagging, boosting, and stacking.

1.1. Problem Statement

Traditional credit risk assessment methods, such as logistic regression, struggle in today's data-rich environment, failing to incorporate non-financial data such as social media activity and purchases, which leads to inaccurate evaluations [5]. Additionally, class imbalance in credit datasets results in poor predictions, especially for high-risk borrowers. These models also lack transparency, making it difficult for regulators to ensure accountability [6]. Furthermore, many conventional models perpetuate social and economic biases, contributing to credit inequality. Deep learning integrated with ensemble learning offers a promising solution to improve accuracy, transparency, and fairness in credit scoring. Traditional credit scoring methods struggle with the complexity and volume of modern data. As data and processing techniques evolve, there is a need for new strategies to improve credit scoring models [7]. Ensemble learning, when combined with deep learning techniques like CNNs and RNNs, can enhance accuracy and handle complex data, addressing the weaknesses of traditional models [8, 9]. In addition, ensemble learning can reduce biases and improve model interpretability, ensuring fairer access to credit, especially for vulnerable groups. This research aims to enhance transparency and confidence in credit scoring models.

1.2. Research Aim and Objectives

The aim of the research is to propose and test an ensemble learning framework integrated with deep learning approaches to improve the credit scoring models' accuracy, fairness, and interpretability. Thus, this research aims to identify the shortcomings of the conventional credit scoring techniques to develop a more efficient and accurate system for evaluating borrowers.

1.2.1. Objectives

- To develop an ensemble learning framework tailored for deep learning-based credit scoring.
- To investigate techniques to mitigate class imbalance issues in credit scoring datasets to improve model performance and fairness.
- To develop methodologies to enhance the interpretability of ensemble models for credit scoring, providing insights into decision-making processes and fostering transparency and trust.
- To conduct comprehensive evaluations of the developed ensemble learning framework on benchmark credit scoring datasets.

1.3. Research Questions

1. What are the key components and design considerations for developing an ensemble learning framework specifically tailored for deep learning-based credit scoring?
2. What techniques can be effectively employed to mitigate class imbalance issues in credit scoring datasets, and how do these techniques improve model performance and fairness?
3. What methodologies can be developed to enhance the interpretability of ensemble models for credit scoring, enabling insights into decision-making processes and fostering transparency and trust?
4. How does the developed ensemble learning framework perform in terms of predictive accuracy, robustness, and computational efficiency when evaluated on benchmark credit scoring datasets?

2. LITERATURE REVIEW

Credit scoring has evolved from simple statistical methods to more complex machine learning and deep learning techniques. Due to the complexity of financial data, the limitations of conventional models

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

are increasingly evident, leading to the adoption of advanced models. This chapter provides a review of the literature on credit scoring, emphasizing traditional methods, recent advances in machine learning, deep learning techniques, and ensemble learning. Additionally, it discusses ethical issues related to credit scoring, including fairness and transparency, highlighting the challenges and opportunities associated with contemporary credit assessment methods.

2.1. Credit Scoring Models

Credit scoring is a vital process in assessing the creditworthiness of individuals and businesses. Initially, scoring models relied on statistical techniques like logistic regression, which used factors such as income, credit history, and debt to predict credit risk [10]. However, as financial risks and data complexity grew, these models became less effective. Modern techniques now incorporate diverse data sources, including social media and transaction behavior, offering better calibration but also raising concerns about data quality, bias, and explainability [11].

2.2. Traditional Credit Scoring Methods

Traditional credit scoring methods, including logistic regression and discriminant analysis, have been the backbone of credit risk evaluation. These models assess the relationship between independent variables (e.g., income, credit history) and the likelihood of default [12]. However, their linear assumptions limit their ability to model complex relationships in real-world data [13]. They also fail to incorporate unstructured data and are prone to biases, particularly in underrepresented demographics [14]. In addition, scorecards, which assign points based on credit characteristics, fail to account for variable correlations, leading to oversimplified evaluations.

2.3. Machine Learning in Credit Scoring

Machine learning (ML) techniques have emerged as more effective alternatives, handling large datasets without assuming linear relationships. Initial models like decision trees were simple to interpret but prone to overfitting and inefficiency in unseen data classification [4, 15]. Techniques such as Random Forests and Gradient Boosting Machines (GBMs) address these issues by using multiple models to improve prediction accuracy [16]. Support Vector Machines (SVMs) are also useful for high-dimensional, non-linear data [17]. While these models offer better accuracy, they lack interpretability, which is a critical issue in regulated industries such as finance [11].

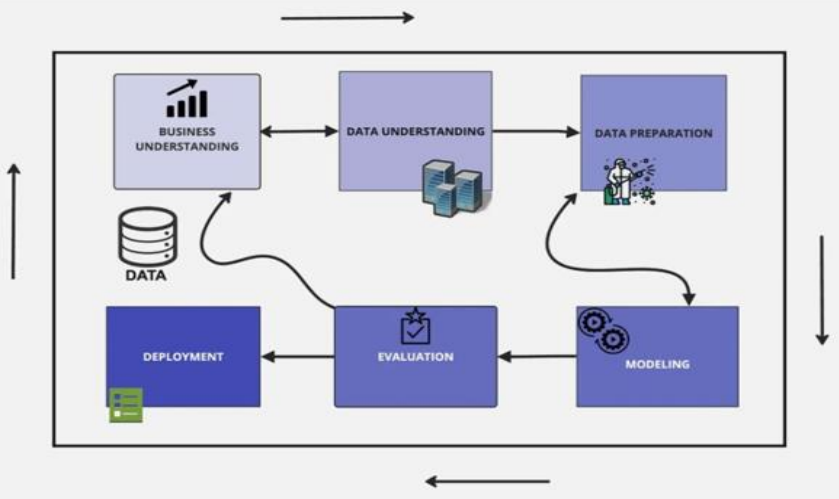

Figure 1.
Flow of credit scoring using machine learning.
Source: Dwifiani [18].

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

2.4. Ensemble Learning in Credit Scoring

Ensemble learning combines multiple models to improve prediction accuracy and reduce bias. Bagging (e.g., Random Forests), boosting (e.g., GBMs, XGBoost), and stacking are common ensemble methods used in credit scoring [19]. Bagging reduces variance and overfitting by training models on random subsets of data, while boosting focuses on correcting errors in previous models. These methods are particularly effective in handling imbalanced datasets and improving model accuracy [20]. However, ensemble methods also suffer from the "black box" problem, making it difficult to explain the decision-making process [21].

2.5. Fairness and Ethical Considerations

As credit scoring models become more advanced, issues of fairness and bias have gained importance. ML models, while more accurate, can inadvertently discriminate based on demographic factors like race, gender, or social class. To ensure fairness, fairness-aware modeling methodologies are essential. Additionally, tools like SHAP and LIME enhance model transparency by explaining how decisions are made, ensuring compliance with regulations, and fostering trust among consumers and regulators [22].

3. Methodology

This chapter outlines the research process, including the methodology, data acquisition, preprocessing, and model validation. It aims to ensure transparency, enabling replication, and addresses concerns such as fairness and data privacy to ensure the appropriate use of machine learning methods. The methodology is designed to yield accurate and reliable credit scoring results. This empirical study uses statistical analysis and machine learning to compare the performance of stacking, bagging, and boosting ensemble methods for credit risk prediction. Three Kaggle datasets are used to validate the findings, representing different borrower characteristics and loan behaviors. The performance of the methods is evaluated based on accuracy, precision, recall, F1 score, and AUC-ROC. This data-centric approach ensures the findings are grounded in real data, making the results applicable to financial risk evaluation.

3.1. Data Collection

Three credit scoring datasets are available on Kaggle, and these were used for this study. All three datasets include a number of borrower characteristics, credit history, income, loan amount, and repayment behavior, which makes all three of them appropriate for credit risk evaluation. These datasets were selected based on their size, the number of features, and the fact that they represent real-world credit scoring problems. These datasets contained missing values, outliers, and a lot of inconsistency, which needed to be addressed through preprocessing. In cases where some observations were missing, different imputation methods were used, and where there were extreme values, they were either adjusted or the values were deleted. Additionally, the data was normalized and scaled as per the requirements of the features, which is very important for applying machine learning algorithms.

3.2. Data Cleaning and Preprocessing

Data cleaning and preprocessing were essential for ensuring the quality and accuracy of the datasets. The datasets contained various borrower characteristics and financial attributes, requiring a structured approach to handle missing values, outliers, and normalization. For missing data, mean and median imputation were used for numerical fields, while the mode imputation technique was applied to categorical variables [23]. Outliers, particularly in loan amounts and income levels, were managed using the IQR rule and z-scores, retaining true variations and correcting or removing measurement errors [24]. To address the sensitivity of ensemble methods to feature scales, MinMax scaling was applied to normalize numerical data within a 0 to 1 range, ensuring balanced predictions. Categorical variables such as occupation and loan purpose were converted to a numerical format through one-hot encoding to eliminate bias and fit the machine learning algorithms [25].

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

3.3. Ensemble Learning Framework

This research applied three ensemble learning methods: stacking, bagging, and boosting, to evaluate credit risk prediction. Ensemble learning combines results from different models, improving accuracy and stability. Stacking involves training base models (logistic regression, Random Forests, and KNN) and combining their predictions through a meta-model (e.g., logistic regression) to obtain the final credit score. Cross-validation was used to prevent overfitting and ensure model credibility [26]. Bagging (Bootstrap Aggregating) reduces model variance and overfitting. Random Forests, which use multiple decision trees built on bootstrapped data samples, make independent decisions. The final prediction is determined by averaging or voting across trees, ensuring robustness against unseen data and effectively modeling non-linear patterns in credit risk data [27]. Boosting iteratively trains models to correct errors made by previous ones. AdaBoost assigns higher weights to misclassified instances, while Gradient Boosting Machines (GBMs) improve efficiency by minimizing residual errors over iterations [26].

3.4. Evaluation Metrics

The performance of the ensemble models was evaluated using accuracy, precision, recall, and F1-score, which provided insights into their effectiveness in credit risk prediction, particularly in reducing false positive rates. Although standard metrics were used, more advanced measures like AUC-ROC curves, classification reports, and confusion matrices were not included in the analysis [28]. The focus was on the ensemble models' ability to perform well on unseen data. Stacking, bagging, and boosting ensemble learning methods were successfully applied to multiple credit scoring datasets, showing better performance in terms of accuracy, stability, and generalization compared to standard models. These approaches addressed issues like class imbalance and overfitting by using diverse data and combining the strengths of different algorithms. While each method had its advantages, the study highlighted the potential of ensemble learning to revolutionize credit risk prediction.

4. Experiments and Results

The results of the three ensemble learning models: stacking, bagging, and boosting on three credit scoring datasets. The evaluation of the performance of each of the models is done using the accuracy of the model, precision, recall, and F1-score, which provides a clear picture of the models in predicting credit risk. Additionally, EDA is performed to showcase some of the main characteristics and trends of each dataset. The results should provide a basis for comparing the advantages and disadvantages of the ensemble methods and which of them can provide the most accurate and stable credit scoring results.

4.1. Exploratory Data Analysis (EDA) for Dataset 1

Dataset 1 consists of 28 fields, including CustomerID, Age, Occupation, Annual Income, and Credit Score, with variables such as Monthly Inhand Salary, Outstanding Debt, Credit Utilization Ratio, and Credit Mix being crucial for assessing credit risk. It also contains categorical variables like Payment Behaviour and Credit Score categories (Good, Standard, Poor). The bar chart of credit score distribution shows a skewed class distribution, with the Standard category being the most common (over 50,000 observations), followed by Poor (30,000) and Good (20,000). This imbalance could pose challenges for machine learning models and may require techniques like boosting to address it 2.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

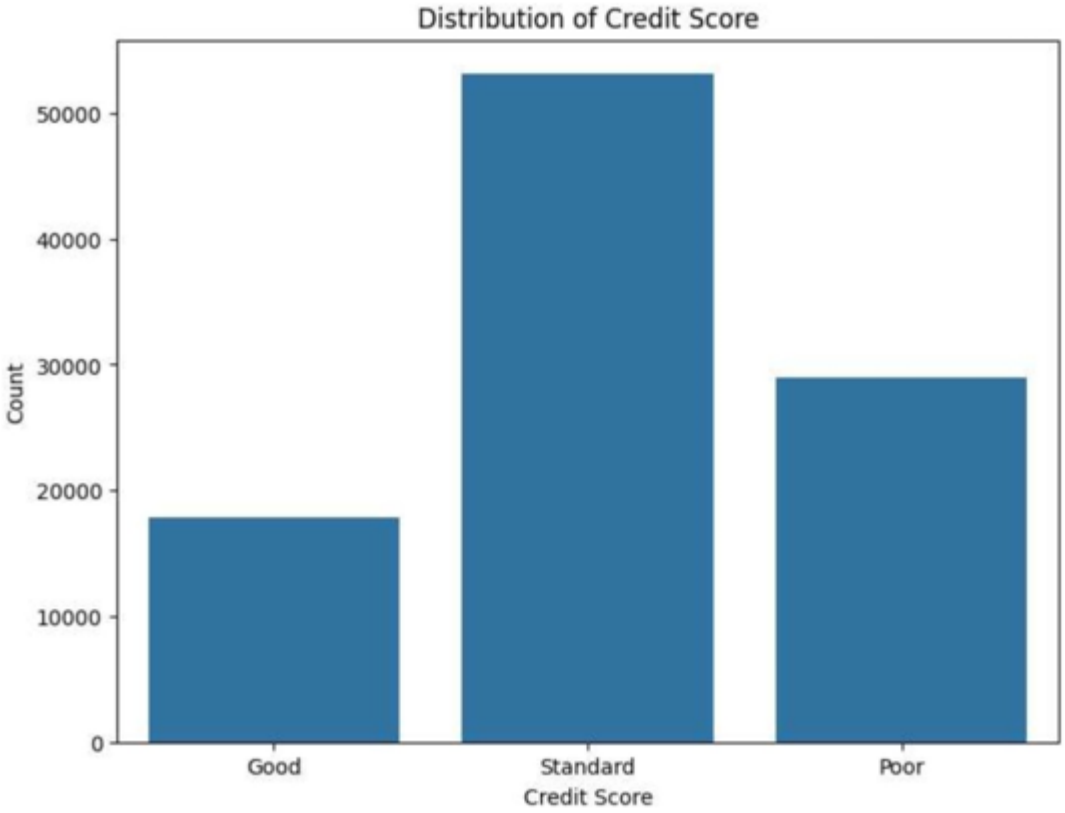

Figure 2.
Bar chart for credit score distribution.

4.2. Model Performance on Dataset 1

The performance of the three ensemble models, stacking, bagging, and boosting, on Dataset 1 is evaluated using classification reports and confusion matrices, with evaluation metrics such as precision, recall, F1-score, and accuracy as shown in Table 1. Each model is trained to predict the credit score category (Good, Poor, or Standard) for the customers in the dataset.

Table 1.
Evaluation of the stacking model for dataset 1.

	Precision	Recall	F1-Score	Support
Good	0.79%	0.78%	0.78%	3527%
Poor	0.80%	0.85%	0.83%	5874%
Standard	0.84%	0.82%	0.83%	10599%
Accuracy			0.82%	20000%
Macro avg	0.81%	0.82%	0.81%	20000%
Weighted avg	0.82%	0.82%	0.82%	20000%

The stacking model, which combines predictions from Random Forest and K-Nearest Neighbors (KNN) as base models, uses Logistic Regression as the meta-model to make the final predictions. The ensemble model achieved an overall accuracy of 82%. The precision and recall were consistently high across all three credit score categories.

The evaluation metrics for the credit score categories are as follows: For a Good Credit Score, Precision is 0.79, Recall is 0.78, and F1-Score is 0.78. For Poor Credit Score, Precision is 0.80, Recall is

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

0.85, and F1-Score is 0.83. Lastly, for the Standard Credit Score, Precision is 0.84, Recall is 0.82, and F1-Score is 0.83. See 3.

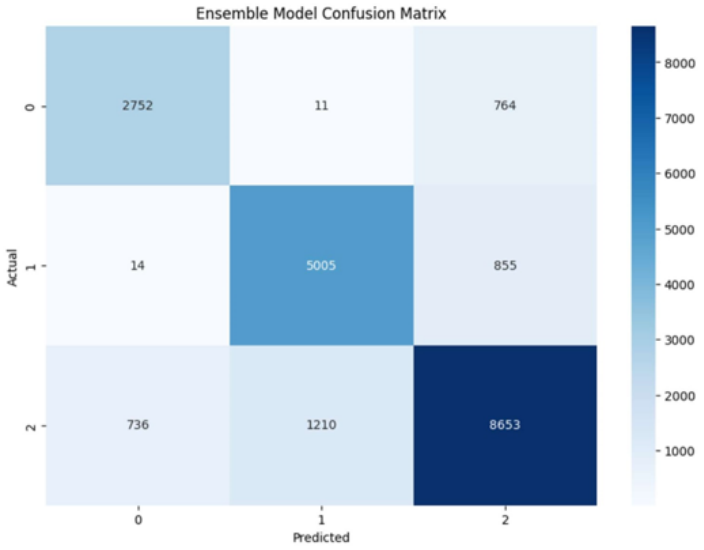

Figure 3.
Confusion matrix for stacking model(dataset-1).

The confusion matrix for the stacking model reveals that most misclassifications occurred between the Standard and Good classes. However, the model performed well in correctly classifying the Poor credit score category, showing strong recall values for this category. As Table 2 shows, the bagging model used a Decision Tree as the base estimator within the Bagging Classifier. This model achieved an overall accuracy of 80% and demonstrated robust performance in identifying customers with both good and poor credit scores.

Table 2.
Evaluation of the Bagging model for dataset-1.

	Precision	Recall	F1-Score	Support
Good	0.76%	0.78%	0.77%	3527%
Poor	0.78%	0.84%	0.81%	5874%
Standard	0.83%	0.79%	0.81%	10599%
accuracy			0.74%	20000%
macro avg	0.73%	0.73%	0.73%	20000%
Weighted avg	0.74%	0.74%	0.74%	20000%

4.2.1. Bagging Model Classification Report

The evaluation metrics for the credit score categories are as follows: For Good Credit Score, Precision is 0.76, Recall is 0.78, and F1-Score is 0.77. For Poor Credit Score, Precision is 0.78, Recall is 0.84, and F1-Score is 0.81. For Standard Credit Score, Precision is 0.83, Recall is 0.79, and F1-Score is 0.81.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

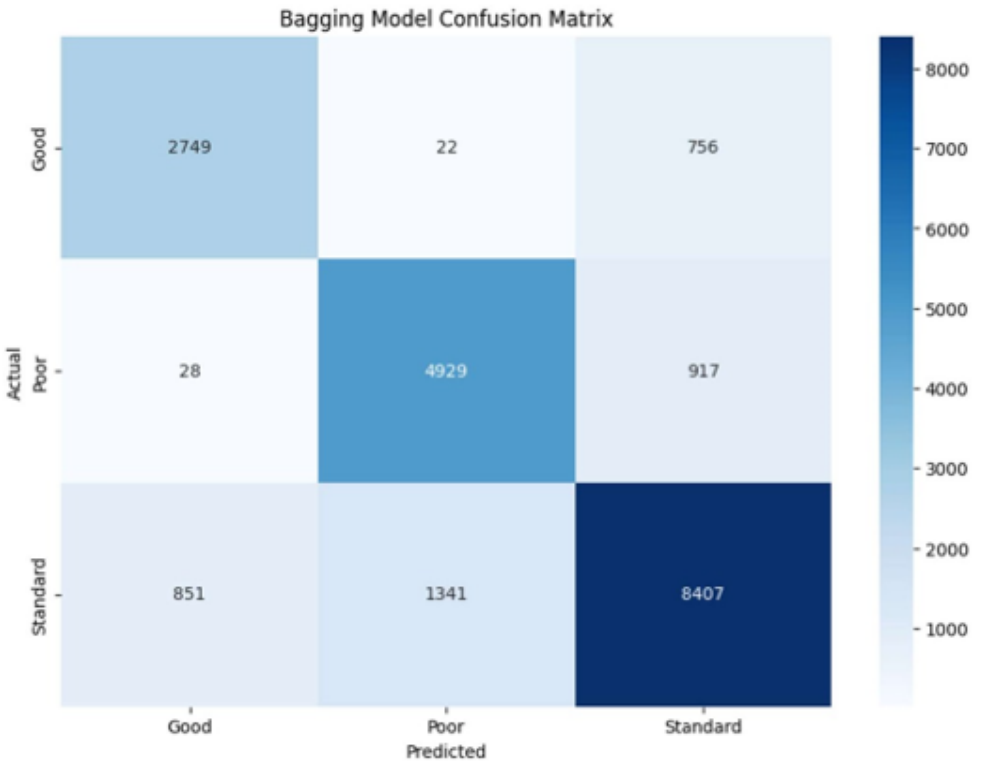

Figure 4.
Confusion matrix for bagging model (dataset -1).

The confusion matrix for the bagging model shows some misclassifications, especially between Standard and Poor classes, but overall, the model provided stable performance across all credit score categories. Boosting model, the AdaBoost model, which used Decision Trees as weak learners, achieved an overall accuracy of 74%, making it the least accurate model compared to stacking and bagging. However, boosting was effective in handling class imbalance and performed well in identifying Standard credit scores.

Table 3.
Evaluation of the Boosting model for dataset-1.

	Precision	Recall	F1-Score	Support
Good	0.70%	0.70%	0.70%	3527%
Poor	0.74%	0.73%	0.73%	5874%
Standard	0.76%	0.77%	0.77%	10599%
Accuracy			0.74%	20000%
Macro avg	0.73%	0.73%	0.73%	20000%
Weighted avg	0.74%	0.74%	0.74%	20000%

The evaluation metrics for the credit score categories, as shown in Figure 5, are as follows: for Good Credit Score, Precision is 0.70, Recall is 0.70, and F1-Score is 0.70. For Poor Credit Score, Precision is 0.74, Recall is 0.73, and F1-Score is 0.73. For the Standard Credit Score, Precision is 0.76, Recall is 0.77, and F1-Score is 0.77.

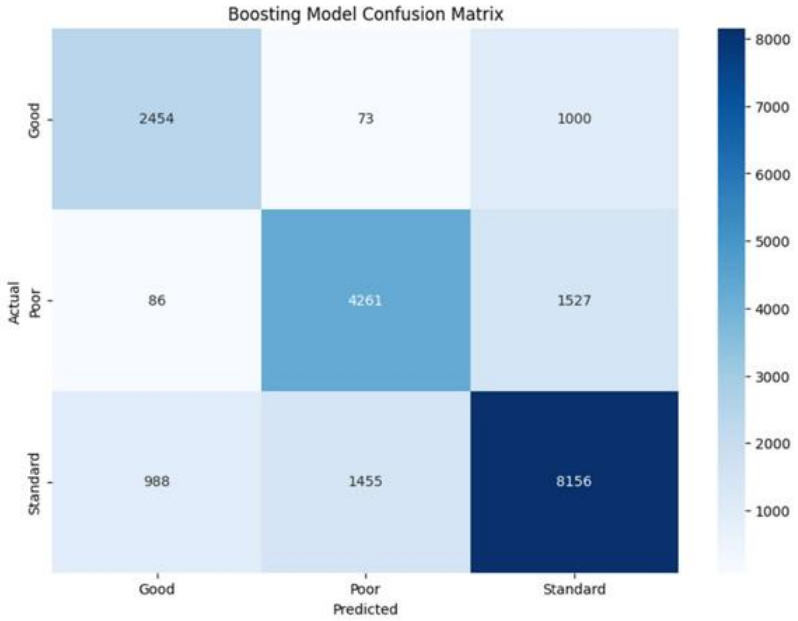

Figure 5.
Confusion matrix for boosting model (dataset-1).

The confusion matrix for boosting highlights misclassifications between the Standard and Good classes, with the highest errors occurring in the good category. Despite these challenges, the boosting model showed reasonable performance in terms of recall and precision for all classes.

Stacking emerged as the best-performing model for this dataset, achieving the highest accuracy and F1-scores across all credit score categories, while bagging and boosting showed stable but comparatively lower performance.

4.3. EDA for Dataset 2

Dataset 2 contains 1,000 observations and 28 features, including financial variables such as Income, Savings, Debt, and Credit Score. Key ratios like Savings to Income and Debt to Income are also present, offering deeper insights into the financial behaviors of individuals.

4.3.1. Key Statistics

Income ranges from €0 to €662,094, with a median of €85,090. The median savings amount is approximately €273,850, and the upper quartile exceeds €622,260. Debt varies widely, with a median of €395,095 and a maximum of €5.9 million. The average credit score is 586, spanning from 300 to 800, with most individuals falling between 500 and 700, as shown in Histogram 6.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

	INCOME	SAVINGS	DEBT	R_SAVINGS_INCOME
count	1000.000000	1.000000e+03	1.000000e+03	1000.000000
mean	121610.019000	4.131896e+05	7.907180e+05	4.063477
std	113716.699591	4.429160e+05	9.817904e+05	3.968097
min	0.000000	0.000000e+00	0.000000e+00	0.000000
25%	30450.250000	5.971975e+04	5.396675e+04	1.000000
50%	85090.000000	2.738505e+05	3.950955e+05	2.545450
75%	181217.500000	6.222600e+05	1.193230e+06	6.307100
max	662094.000000	2.911863e+06	5.968620e+06	16.111200

	R_DEBT_INCOME	R_DEBT_SAVINGS	T_CLOTHING_12	T_CLOTHING_6
count	1000.000000	1000.000000	1000.000000	1000.000000
mean	6.068449	5.867252	6822.401000	3466.320000
std	5.847878	16.788356	7486.225932	5118.942977
min	0.000000	0.000000	0.000000	0.000000
25%	1.454500	0.206200	1084.500000	319.500000
50%	4.911550	2.000000	4494.000000	1304.000000
75%	8.587475	4.509600	10148.500000	4555.500000
max	37.000600	292.842100	43255.000000	39918.000000

	R_CLOTHING	T_ENTERTAINMENT_12	...	R_EXPENDITURE_INCOME
count	1000.000000	1000.000000	...	1000.000000
mean	0.454848	14261.255000	...	0.943607
std	0.236036	12388.187688	...	0.168989
min	0.000000	0.000000	...	0.666700
25%	0.263950	4248.750000	...	0.833300
50%	0.468850	9401.000000	...	0.909100
75%	0.626300	22892.500000	...	1.000000
max	1.058300	62529.000000	...	2.000200

	R_EXPENDITURE_SAVINGS	R_EXPENDITURE_DEBT	CAT_DEBT
count	1000.000000	1000.000000	1000.000000
mean	0.913340	0.605276	0.944000
std	1.625278	1.299382	0.230037
min	0.000000	0.000000	0.000000
25%	0.158700	0.100000	1.000000
50%	0.327950	0.178600	1.000000
75%	0.833300	0.588200	1.000000
max	10.009900	10.005300	1.000000

	CAT_CREDIT_CARD	CAT_MORTGAGE	CAT_SAVINGS_ACCOUNT	CAT_DEPENDENTS
count	1000.000000	1000.000000	1000.000000	1000.00000
mean	0.236000	0.173000	0.993000	0.15000
std	0.424835	0.378437	0.083414	0.35725
min	0.000000	0.000000	0.000000	0.00000
25%	0.000000	0.000000	1.000000	0.00000
50%	0.000000	0.000000	1.000000	0.00000
75%	0.000000	0.000000	1.000000	0.00000
max	1.000000	1.000000	1.000000	1.00000

	CREDIT_SCORE	DEFAULT
count	1000.000000	1000.000000
mean	586.712000	0.284000
std	63.413882	0.451162
min	300.000000	0.000000
25%	554.750000	0.000000
50%	596.000000	0.000000
75%	630.000000	1.000000
max	800.000000	1.000000

Figure 6.
Descriptive statistics for dataset 2.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

4.3.2. Distribution of Default and Credit Scores

The distribution of default instances in the dataset reveals an imbalance, with 28.4% (284 individuals) having defaulted on their loans, and 71.6% (716 individuals) being non-defaulters, see Figure 7. This skewed dataset could impact model performance, particularly in identifying defaulters, highlighting the need for techniques like oversampling or boosting.

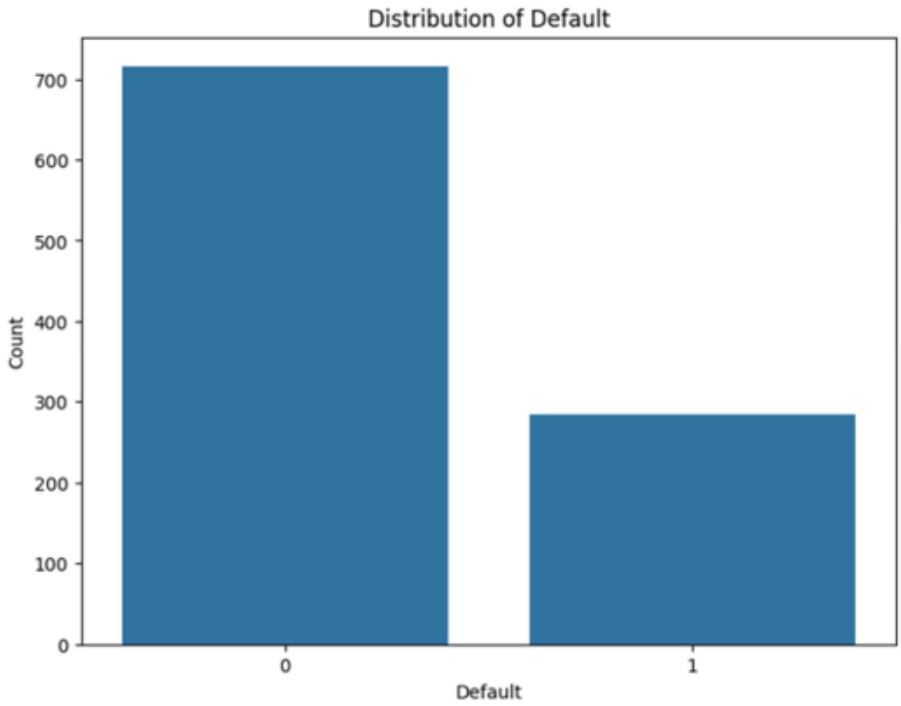

Figure 7.
Distribution of the default variable.

The histogram with the kernel density plot on the right represents the density of credit scores in the data set. The credit score is given between 300 and 800, with most of the scores falling between 500 and 700. The maximum is achieved at 600, meaning that the majority of people have average credit ratings. The distribution is slightly negatively skewed; that is, a few people have extremely low or extremely high scores. Of course, credit risk models based on credit scores will require models that can identify higher default rates for lower credit scores 8.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

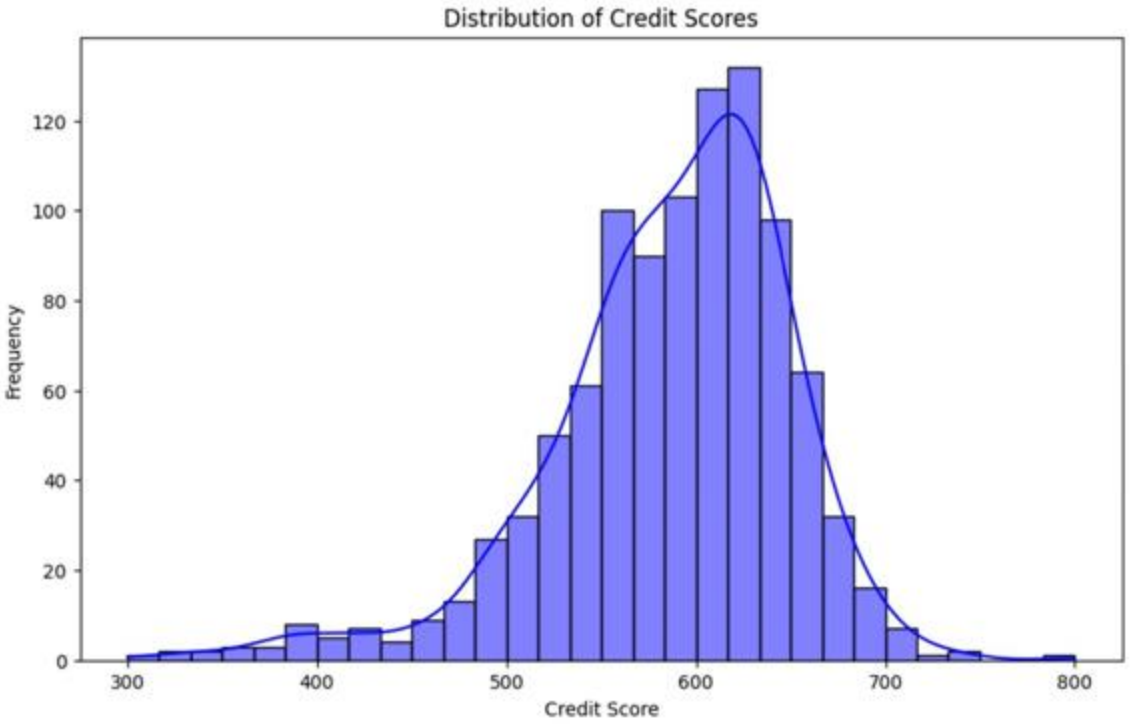

Figure 8.
Credit score distribution.

4.4. Model Performance on Dataset 2

The performance of ensemble learning models, stacking, bagging, and boosting on Dataset 2, is evaluated based on the task of predicting whether an individual defaults on their loan (Default = 1) or does not default (Default = 0). The performance metrics, including precision, recall, F1-score, and accuracy, provide insights into each model's effectiveness in handling this binary classification problem. The stacking model combines predictions from a Random Forest and K-Nearest Neighbors (KNN) model, with Logistic Regression serving as the meta-model. The model achieved an accuracy of 70%, indicating a reasonable performance for predicting defaults.

Table 4.
Evaluation of the stacking model (dataset-2).

	Precision	Recall	F1-Score	Support
Not Default (0)	0.75%	0.88%	0.81%	146%
Defaulted (1)	0.40%	0.22%	0.29%	54%
Accuracy			0.70%	200%
Macro Avg	0.58%	0.55%	0.55%	200%
Weighted Avg	0.66%	0.70%	0.67%	200%

The performance metrics for the credit default categories are as follows: For Not Default (0), precision is 0.75, recall is 0.88, and the F1-score is 0.81. For Defaulted (1), precision is 0.40, recall is 0.22, and the F1-score is 0.29, indicating lower performance in predicting defaults.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

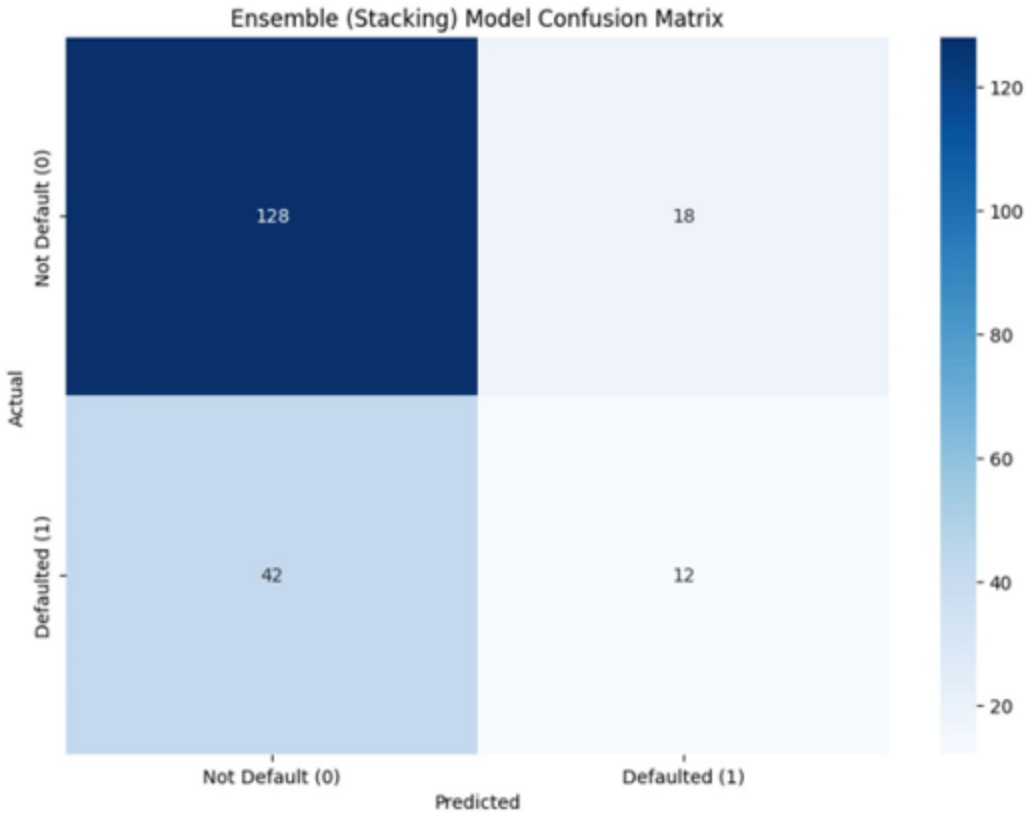

Figure 9.
Confusion matrix for stacking model (Dataset-2).

The confusion matrix (Figure 9) shows that the stacking model correctly classified 128 out of 146 instances of non-default and 12 out of 54 instances of default. However, the model struggles with correctly identifying defaulters, as shown by the lower recall for Defaulted (1). The model tends to classify more individuals as non-defaulters, leading to high precision but low recall for defaults. The Bagging Classifier, which uses multiple decision trees as base models, achieved the highest accuracy among the three models at 72%. Bagging typically reduces variance by averaging the results of many decision trees, leading to more stable predictions.

Table 5.
Evaluation of the bagging model (dataset-2).

	Precision	Recall	F1-Score	Support
Not Default (0)	0.76%	0.89%	0.82%	146%
Defaulted (1)	0.47%	0.26%	0.33%	54%
Accuracy			0.72%	200%
Macro Avg	0.62%	0.57%	0.58%	200%
Weighted Avg	0.68%	0.72%	0.69%	200%

Table 4 shows performance metrics for the credit default categories as follows: For Not Default (0), precision is 0.76, recall is 0.89, and the F1-score is 0.82. For Defaulted (1), precision is 0.47, recall is 0.26, and the F1-score is 0.33, indicating lower performance in predicting defaults.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

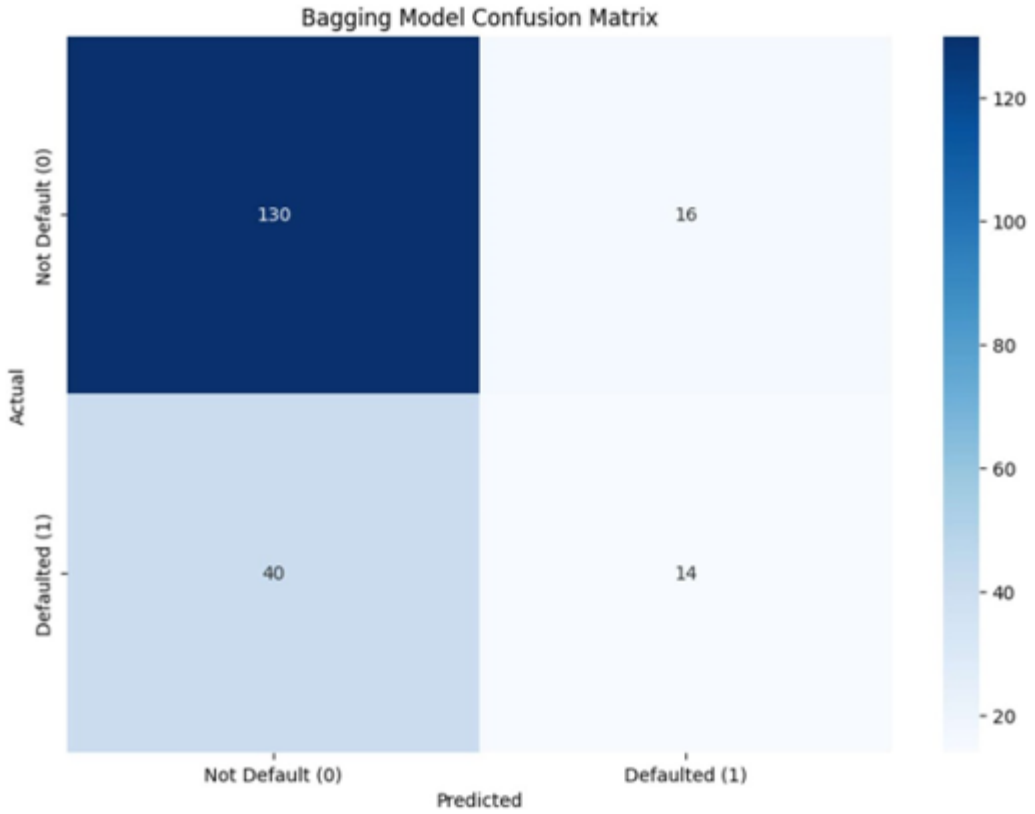

Figure 10.
Confusion matrix for bagging model (Dataset-2).

Figure 10 shows that the bagging model correctly classified 130 out of 146 non-default cases and 14 out of 54 default cases. Similar to the stacking model, the bagging model performs well for non-defaulters but struggles with identifying default cases, although it does show a slightly better recall for the Defaulted (1) class compared to stacking. The boosting model, specifically the AdaBoost model, which uses Decision Trees as weak learners, achieved an overall accuracy of 58%, making it the least accurate model. Boosting works by iteratively correcting the mistakes of previous models, but in this case, the model performed poorly in distinguishing between default and non-default cases.

Table 6.
Evaluation of the boosting model (dataset-2).

	Precision	Recall	F1-Score	Support
Not Default (0)	0.74%	0.66%	0.70%	146%
Defaulted (1)	0.29%	0.37%	0.32%	54%
Accuracy			0.58%	200%
Macro Avg	0.51%	0.51%	0.51%	200%
Weighted Avg	0.62%	0.58%	0.59%	200%

The performance metrics for the credit default categories are as follows: see Table 6. For Not Default (0), precision is 0.74, recall is 0.66, and the F1-score is 0.70. For Defaulted (1), precision is 0.29, recall is 0.37, and the F1-score is 0.32, indicating lower performance in predicting defaults.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

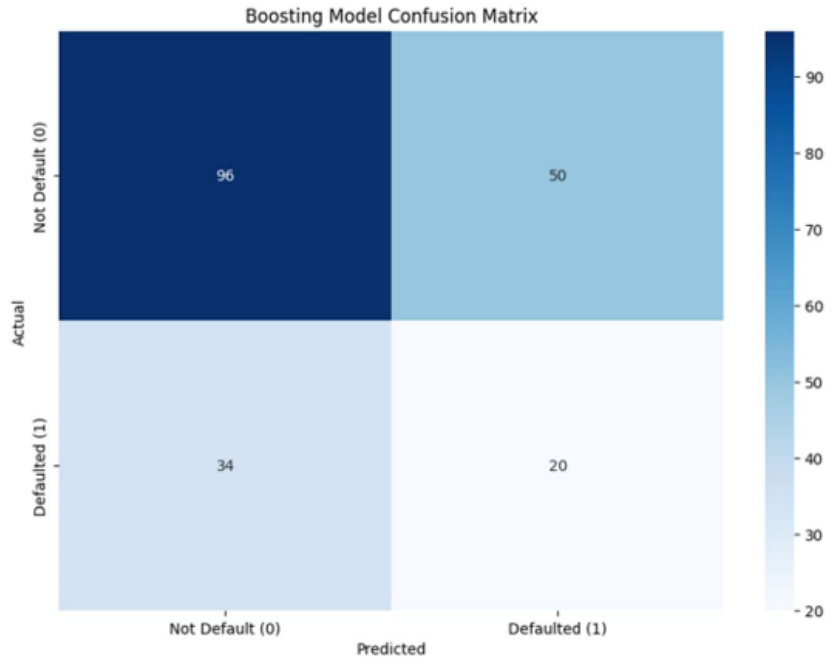

Figure 11.
Confusion matrix for boosting model (Dataset-2).

The confusion matrix in Figure 11 indicates that the model correctly classified 96 out of 146 non-default cases and 20 out of 54 default cases. The boosting model struggled the most with predicting defaults, as evidenced by its low recall and F1-score for the Defaulted (1) category.

The bagging model demonstrated the best overall performance on Dataset 2, with the highest accuracy and a better balance between precision and recall for both classes. The stacking model also performed reasonably well but had issues with recall for defaulters. The boosting model, while effective for some tasks, was the least successful in this case, particularly in identifying default instances.

4.5. Exploratory Data Analysis (EDA) for Dataset 3

Dataset 3 has 32 attributes that are aspects of the financial behavior of a customer, such as Annual Income, Number of Bank Accounts, Credit Card Usage, Number of Loans, Credit Score, etc. It involves information on customer credit and loan activities, with data exceeding 100 thousand cases. The credit score is categorized into three classes: Good, Poor, and Standard, with the class having the largest number of samples being the Standard class.

4.5.1. Key Statistics

The data for the key features are as follows: Annual income ranges from €7,000 to over €24 million, with a median of €37,578, indicating a wide income disparity across customers. The number of loans held by customers varied significantly, with a mean of 3.01 loans, although some customers had over 1,400 loans. Credit inquiries ranged from 0 to 2,597, reflecting varied levels of customer engagement with credit institutions. The number of delayed payments had an average of 30, with a maximum value of 4,397, indicating potential credit risk. Figure 12.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

	Unnamed: 0	Annual_Income	Num_Bank_Accounts	Num_Credit_Card	\
count	100000.000000	1.000000e+05	100000.000000	100000.000000	
mean	49999.500000	1.764157e+05	17.091700	22.474560	
std	28867.657797	1.429618e+06	117.404773	129.057388	
min	0.000000	7.005930e+03	0.000000	1.000000	
25%	24999.750000	1.945750e+04	3.000000	4.000000	
50%	49999.500000	3.757861e+04	6.000000	5.000000	
75%	74999.250000	7.279092e+04	7.000000	7.000000	
max	99999.000000	2.419806e+07	1798.000000	1499.000000	

	Interest_Rate	Num_of_Loan	Delay_from_due_date	\
count	100000.000000	100000.000000	100000.000000	
mean	72.466040	3.009960	21.068780	
std	466.422621	62.647879	14.860104	
min	1.000000	-100.000000	-5.000000	
25%	8.000000	1.000000	10.000000	
50%	13.000000	3.000000	18.000000	
75%	20.000000	5.000000	28.000000	
max	5797.000000	1496.000000	67.000000	

	Num_of_Delayed_Payment	Changed_Credit_Limit	Num_Credit_Inquiries	\
count	100000.000000	100000.000000	100000.000000	
mean	30.088470	10.343671	27.287480	
std	217.996071	6.725301	191.298349	
min	-3.000000	-6.490000	0.000000	
25%	9.000000	5.420000	3.000000	
50%	15.000000	9.250000	5.000000	
75%	19.000000	14.660000	9.000000	
max	4397.000000	36.970000	2597.000000	

	...	Payday Loan	Mortgage Loan	Auto Loan	Home Equity Loan	\
count	...	100000.000000	100000.000000	100000.000000	100000.000000	
mean	...	0.319440	0.313600	0.305600	0.314000	
std	...	0.466262	0.463958	0.460663	0.464119	
min	...	0.000000	0.000000	0.000000	0.000000	
25%	...	0.000000	0.000000	0.000000	0.000000	
50%	...	0.000000	0.000000	0.000000	0.000000	
75%	...	1.000000	1.000000	1.000000	1.000000	
max	...	1.000000	1.000000	1.000000	1.000000	

	High_spent_Medium_value_payments	High_spent_Small_value_payments	\
count	100000.000000	100000.000000	
mean	0.175400	0.113400	
std	0.380311	0.317083	
min	0.000000	0.000000	
25%	0.000000	0.000000	
50%	0.000000	0.000000	
75%	0.000000	0.000000	
max	1.000000	1.000000	

	Low_spent_Large_value_payments	Low_spent_Medium_value_payments	\
count	100000.000000	100000.000000	
mean	0.104250	0.138610	
std	0.305586	0.345541	
min	0.000000	0.000000	
25%	0.000000	0.000000	
50%	0.000000	0.000000	
75%	0.000000	0.000000	
max	1.000000	1.000000	

	Low_spent_Small_value_payments	Credit_Score
count	100000.000000	100000.00000
mean	0.331130	0.88830
std	0.470622	0.67512
min	0.000000	0.00000
25%	0.000000	0.00000
50%	0.000000	1.00000
75%	1.000000	1.00000
max	1.000000	2.00000

[8 rows x 32 columns]

Figure 12.
Descriptive statistics for dataset 3.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

The bar chart illustrates the distribution of credit scores in Dataset 3, showcasing three categories: 0 (Good), 1 (Standard), and 2 (Poor). The majority of the customers are categorized as Standard customers (Credit score 1), while the rest are in the Good and Poor categories. This distribution indicates a class imbalance, with the Standard category being dominant, which may affect the performance of machine learning algorithms. An imbalance between classes will necessitate methods such as oversampling or boosting to ensure that all credit score classes are well represented during model training.

4.5.2. Distribution of Credit Score

The second bar chart represents the loan portfolio of customers segregated by credit score Figure 13. Customers with a credit score of 0 (customers with poor credit scores) take more loans than customers in the standard and good categories. This pattern indicates that there is a likelihood that people with poor credit scores to have many loans, and this may be the reason why they will be performing poorly. The chart also shows that, as credit scores increase, the total loans reduce, suggesting that higher credit scores could be linked to better handling of credit. Figure 14.

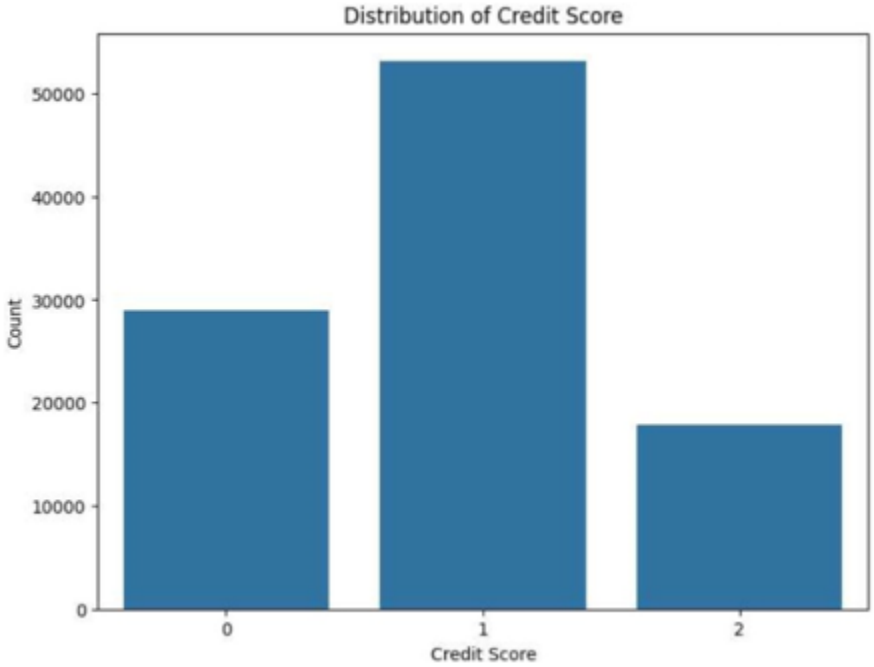

Figure 13.
Bar chart showing the distribution of Credit scores.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

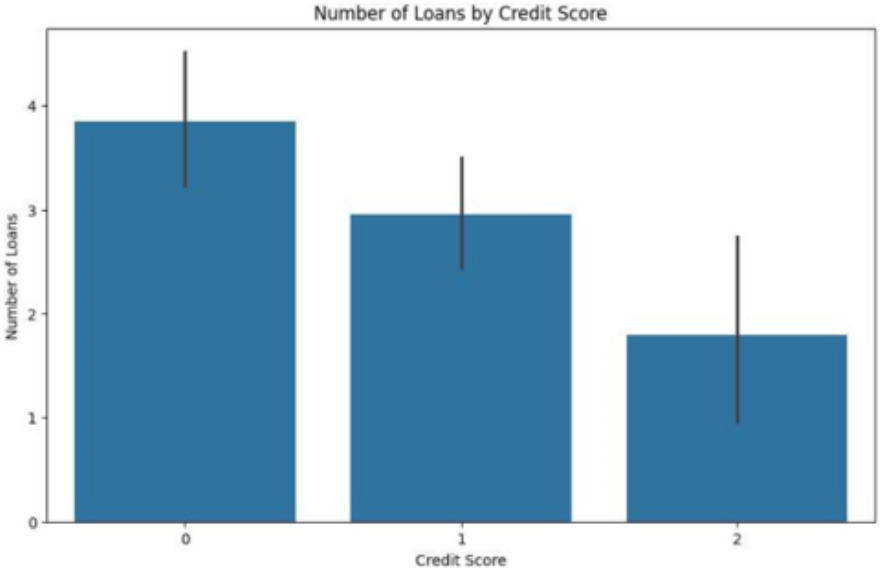

Figure 14.
Bar chart for the number of loans by credit score.

4.6. Model Performance on Dataset 3

The performance of three ensemble learning models, stacking, bagging, and boosting, was evaluated on Dataset 3 to predict credit scores across three categories: Low, Medium, and High. The stacking model combines Random Forest and K-Nearest Neighbors (KNN) as base models, with Logistic Regression as the meta-model. The overall accuracy achieved by the stacking model was 78%. The classification report and confusion matrix show relatively balanced performance across the three credit score categories.

Table 7.
Evaluation of stacking model (dataset-3).

	Precision	Recall	F1-score	Support
Low	0.78%	0.79%	0.79%	5874%
Medium	0.80%	0.81%	0.80%	10599%
High	0.73%	0.70%	0.72%	3527%
Accuracy			0.78%	20000%
Macro avg	0.77%	0.77%	0.77%	20000%
Weighted avg	0.78%	0.78%	0.78%	20000%

The performance metrics for the credit score categories are as follows: For Low Credit Score, precision is 0.78, recall is 0.79, and the F1-score is 0.79. For a Medium Credit Score, precision is 0.80, recall is 0.81, and the F1-score is 0.80. For a High Credit Score, precision is 0.73, recall is 0.70, and the F1-score is 0.72, indicating slightly lower performance compared to the other categories. Table 7.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

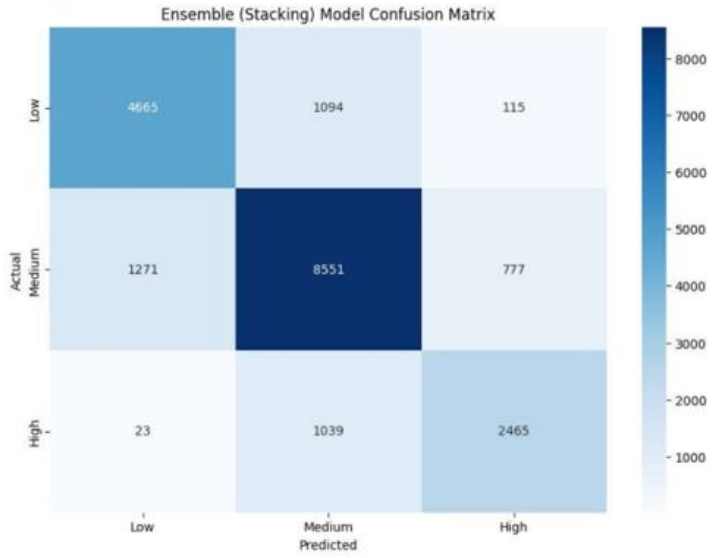

Figure 15.
Confusion matrix for stacking model (dataset-3).

The confusion matrix reveals that the stacking model performed well in classifying Low and Medium credit scores, while there were more misclassifications between Medium and High categories. There was a balanced performance seen by the stacking model, as shown in Figure 15. The Bagging Classifier with decision trees as base learners had an accuracy of about 76%. The ability to combine the forecasts from a set of decision trees and thereby achieve higher stability speaks to the effectiveness of the Bagging method for the Low and Medium credit scores.

Table 8.
Evaluation of the Bagging model (dataset-3).

	Precision	Recall	F1-score	Support
Low	0.75%	0.79%	0.77%	5874%
Medium	0.78%	0.79%	0.78%	10599%
High	0.71%	0.63%	0.67%	3527%
Accuracy			0.76%	20000%
Macro avg	0.75%	0.73%	0.74%	20000%
Weighted avg	0.76%	0.76%	0.76%	20000%

The performance metrics for the credit score categories are as follows: For Low Credit Score, precision is 0.75, recall is 0.79, and the F1-score is 0.77. For a Medium Credit Score, precision is 0.78, recall is 0.79, and the F1-score is 0.78. For a High Credit Score, precision is 0.71, recall is 0.63, and the F1-score is 0.67, indicating lower performance in the High Credit Score category, Table 8.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

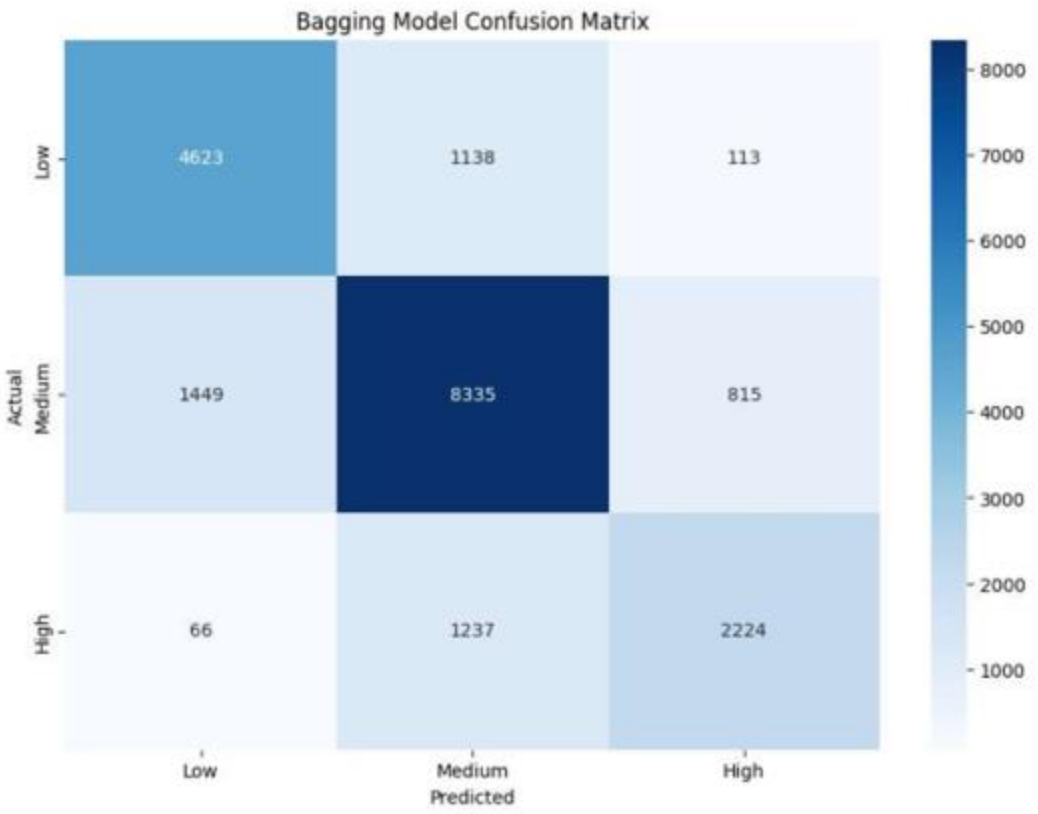

Figure 16.
Confusion matrix for bagging model (dataset-3).

From the above confusion matrix, it is evident that the category that posed a high challenge to the bagging model was the High credit score, whereby several instances were classified as Medium. However, for Low and Medium credit scores, the bagging model performed consistently and is hence considered a good model for such scores 16. The boosting model's lowest accuracy was scored by the AdaBoost model, which employs Decision Trees as weak learners, with an overall accuracy of a mere 69%. Boosting works by trying to augment performance by gradually altering the weights of misclassified instances; thus, it serves well when dealing with imbalances.

Table 9.
Evaluation of the boosting model (dataset-3).

	Precision	Recall	F1-Score	Support
Low	0.69%	0.67%	0.68%	5874%
Medium	0.72%	0.73%	0.73%	10599%
High	0.59%	0.59%	0.59%	3527%
accuracy			0.69%	20000%
macro avg	0.67%	0.67%	0.67%	20000%
weighted avg	0.69%	0.69%	0.69%	20000%

The performance metrics for the credit score categories are as follows: For Low Credit Score, precision is 0.69, recall is 0.67, and the F1-score is 0.68. For a Medium Credit Score, precision is 0.72, recall is 0.73, and the F1-score is 0.73. For a High Credit Score, precision is 0.59, recall is 0.59, and the F1-score is 0.59, indicating lower performance in the High Credit Score category. See Table 9.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

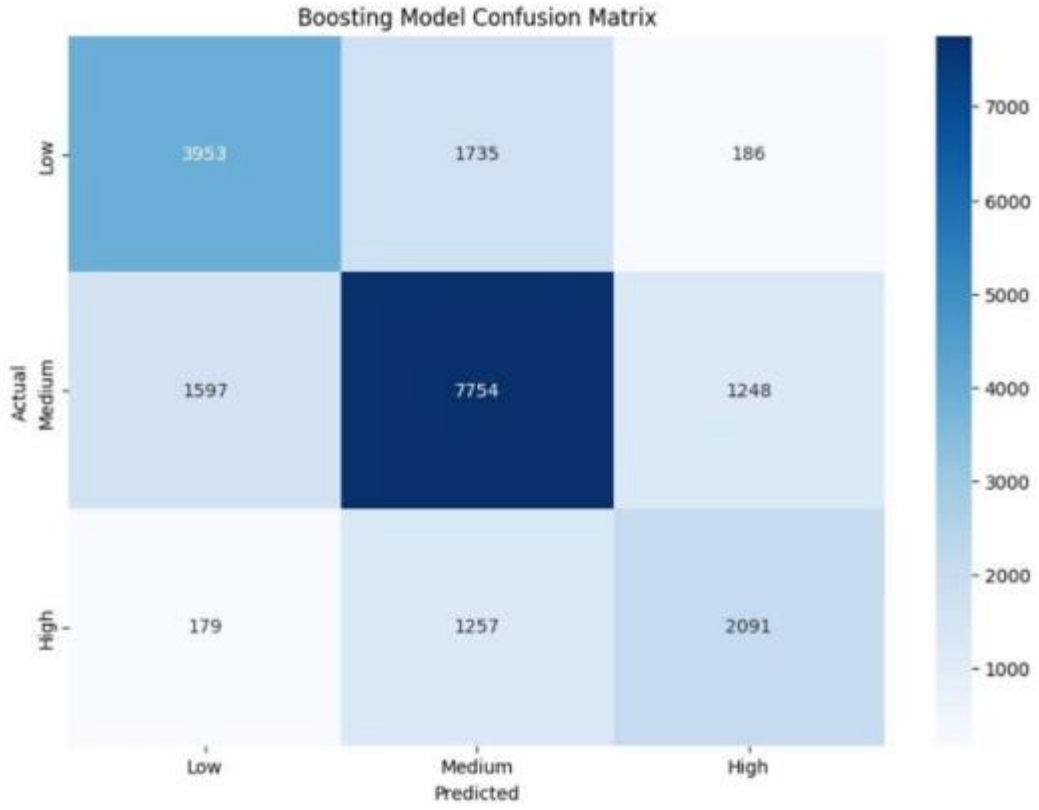

Figure 17.
Confusion matrix for boosting model (dataset-3).

From the confusion matrix in Figure 17, it is evident that the boosting model was problematic with the High credit score category, which led to a lot of confusion and misclassification. This could have been due to the lower performance of the model in the High credit scores when compared to the Low credit scores, thus achieving a lower accuracy.

The stacking model emerged as the best performer, with the highest accuracy and balanced precision-recall across all three categories. The bagging model also performed well, but struggled more with the high credit score category. The boosting model, while effective for imbalanced data, had the lowest performance, particularly in predicting high credit scores.

4.7. Comparison across Ensemble Methods

Dataset 1 stacking model proposed the highest accuracy of 82% in this dataset, indicating the efficiency of the stacking model in this context. The bagging model was performed with 80% accuracy, followed by the boosting model with only 74% accuracy. For Dataset 2, the bagging model was the best performing with an accuracy of 72%, which shows this model's ability in dealing with this dataset. The stacking model was slightly lower at 70%, while the boosting model was even lower with a score of 58%, indicating some difficulties in identifying the defaulters in this dataset. In Dataset 3, as observed, the stacking model performed well with 78%, while the bagging model was slightly behind with 76%. The boosting model had the least accuracy of 69%. In all three datasets, stacking and bagging were superior to boosting; bagging was even better in the second dataset, while stacking was the best for the first and the third datasets.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

Table 10.
Comparison of accuracy for stacking, bagging, and boosting for three datasets.

Dataset	Stacking Accuracy	Bagging Accuracy	Boosting Accuracy
Dataset 1	82%	80%	74%
Dataset 2	70%	72%	58%
Dataset 3	78%	76%	69%

5. Discussion

This section examines the performance of stacking, bagging, and boosting on three datasets for credit score and loan default prediction. Stacking performed best in multi-class classification, achieving 82% accuracy in Dataset 1, thanks to its ability to combine the strengths of Random Forest and KNN. Bagging, with 80% accuracy, was effective in stabilizing predictions in Dataset 1, while boosting lagged at 74% due to sensitivity to class imbalance. For Dataset 2 (loan default prediction), bagging was most effective at 72%, with stacking at 70%, and boosting at 58%. In Dataset 3 (multi-class classification), stacking again led with 78%, followed by bagging at 76%, while boosting struggled with 69%. Stacking and bagging consistently outperformed boosting, with stacking excelling in multi-class problems and bagging in binary classification tasks with imbalanced data. While boosting improved recall, it was prone to overfitting and less effective in complex problems. These findings suggest that stacking and bagging are suitable for real-world credit scoring applications, while boosting may not be ideal for credit risk assessment due to its overfitting tendencies. Financial institutions should consider the characteristics of their data and objectives when choosing an ensemble method.

6. Limitations and Future Work

This study has a few limitations, such as the class imbalance in Datasets 2 and 3, which affected the performance of boosting and other algorithms. Future research could address this using techniques like SMOTE or cost-sensitive learning. Additionally, the study used limited feature engineering, and more advanced techniques like PCA or RFECV could improve model accuracy by better selecting relevant features. Hyperparameter tuning, which was not applied in this study, could also enhance model performance by optimizing parameters such as the number of estimators and learning rates. Future work could focus on advanced class imbalance handling, improved feature engineering, and enhancing model interpretability using methods like SHAP or LIME to increase transparency in credit scoring systems.

7. Conclusions

This research compared the performance of three ensemble learning algorithms: stacking, bagging, and boosting, on credit score and loan default prediction datasets. Stacking excelled in multi-class classification, achieving 82% accuracy for Dataset 1 and 78% for Dataset 3, while bagging performed best in binary classification with 72% accuracy for Dataset 2. Boosting, however, showed lower performance across all datasets due to its sensitivity to class imbalance and complex structures. The study demonstrates that stacking and bagging are effective methods for credit scoring, with stacking ideal for multi-class and bagging for binary tasks. Financial institutions can benefit from adopting these models to improve accuracy, address class imbalance, and enhance risk management. Properly developed, these ensemble techniques could reduce credit risks and strengthen the financial system.

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297

Acknowledgments:

This paper is result of MSc thesis at Dublin Business School Ireland and many thanks to all co-authors, who helped me to reshape the paper and suggested experiments.

References

[1] Y. Song, Y. Wang, X. Ye, D. Wang, Y. Yin, and Y. Wang, "Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in P2P lending," *Information Sciences*, vol. 525, pp. 182-204, 2020. https://doi.org/10.1016/j.ins.2020.03.027

[2] N. Chen, B. Ribeiro, and A. Chen, "Financial credit risk assessment: A recent review," *Artificial Intelligence Review*, vol. 45, no. 1, pp. 1-23, 2016. https://doi.org/10.1007/s10462-015-9434-x

[3] H. Zhang, Y. Shi, X. Yang, and R. Zhou, "A firefly algorithm modified support vector machine for the credit risk assessment of supply chain finance," *Research in International Business and Finance*, vol. 58, p. 101482, 2021. https://doi.org/10.1016/j.ribaf.2021.101482

[4] S. Bhatore, L. Mohan, and Y. R. Reddy, "Machine learning techniques for credit risk evaluation: A systematic literature review," *Journal of Banking and Financial Technology*, vol. 4, no. 1, pp. 111-138, 2020. https://doi.org/10.1007/s42786-020-00020-3

[5] Y. Huang, L. Zhang, Z. Li, H. Qiu, T. Sun, and X. Wang, "Fintech credit risk assessment for SMEs: Evidence from China," IMF Working Paper No. 2020/193. International Monetary Fund, 2020.

[6] A. O. Scott, P. Amajuoyi, and K. B. Adeusi, "Advanced risk management solutions for mitigating credit risk in financial operations," *Magna Scientia Advanced Research and Reviews*, vol. 11, no. 1, pp. 212-223, 2024.

[7] M. Barua, T. Kumar, K. Raj, and A. M. Roy, "Comparative analysis of deep learning models for stock price prediction in the Indian marketa," *FinTech*, vol. 3, no. 4, pp. 551-568, 2024. https://doi.org/10.3390/fintech3040029

[8] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers of Computer Science*, vol. 14, no. 2, pp. 241-258, 2020. https://doi.org/https://doi.org/10.1007/s11704-019-8208-z

[9] T. Kumar, R. Brennan, A. Mileo, and M. Bendechache, "Image data augmentation approaches: A comprehensive survey and future directions," *IEEE Access*, vol. 12, pp. 187536 - 187571, 2024. https://doi.org/10.1109/ACCESS.2024.3470122

[10] P. Trivedi, "India's response to coronavirus pandemic: Nine lessons for effective public management," *The American Review of Public Administration*, vol. 50, no. 6-7, pp. 725-728, 2020.

[11] M. Bücker, G. Szepannek, A. Gosiewska, and P. Biecek, "Transparency, auditability, and explainability of machine learning models in credit scoring," *Journal of the Operational Research Society*, vol. 73, no. 1, pp. 70-90, 2022. https://doi.org/10.1080/01605682.2021.1922098

[12] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable machine learning in credit risk management," *Computational Economics*, vol. 57, no. 1, pp. 203-216, 2021. https://doi.org/10.1007/s10614-020-10042-0

[13] L. Alzubaidi *et al.*, "A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications," *Journal of Big Data*, vol. 10, p. 46, 2023. https://doi.org/10.1186/s40537-023-00727-2

[14] X. Dastile, T. Celik, and M. Potsane, "Statistical and machine learning models in credit scoring: A systematic literature survey," *Applied Soft Computing*, vol. 91, p. 106263, 2020. https://doi.org/10.1016/j.asoc.2020.106263

[15] T. Kumar, J. Park, M. S. Ali, A. S. Uddin, J. H. Ko, and S.-H. Bae, "Binary-classifiers-enabled filters for semi-supervised learning," *IEEE Access*, vol. 9, pp. 167663-167673, 2021. https://doi.org/10.1109/ACCESS.2021.3124200

[16] J. L. Breeden, "Survey of machine learning in credit risk," *Available at SSRN 3616342*, 2020. http://dx.doi.org/10.2139/ssrn.3616342

[17] P. Biecek *et al.*, "Enabling machine learning algorithms for credit scoring--explainable artificial intelligence (XAI) methods for clear understanding complex predictive models," *arXiv preprint arXiv:2104.06735*, 2021. https://doi.org/10.48550/arXiv.2104.06735

[18] A. Dwifiani, "Credit scoring flow using machine learning. Medium," 2023. https://adwifiani.medium.com/credit-scoring-flow-using-machine-learning-b08f011d0ec2

[19] D. Tripathi, A. K. Shukla, B. R. Reddy, G. S. Bopche, and D. Chandramohan, "Credit scoring models using ensemble learning and classification approaches: A comprehensive survey," *Wireless Personal Communications*, vol. 123, no. 1, pp. 785-812, 2022. https://doi.org/10.1007/s11277-021-09158-9

[20] M. S. Saeed, "Cross project software defect prediction using ensemble learning, a comprehensive review," *International Journal of Computational and Innovative Sciences*, vol. 3, no. 2, pp. 34-42, 2024.

[21] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105151, 2022. https://doi.org/10.1016/j.engappai.2022.105151

[22] M. Hejazinia *et al.*, "Fel: High capacity learning for recommendation and ranking via federated ensemble learning," *arXiv preprint arXiv:2206.03852*, 2022. https://doi.org/10.48550/arXiv.2206.03852

[23] G. Wang, J. Hao, J. Ma, and H. Jiang, "A comparative assessment of ensemble learning for credit scoring," *Expert Systems with Applications*, vol. 38, no. 1, pp. 223-230, 2011. https://doi.org/10.1016/j.eswa.2010.06.048

[24] B. Soni, "Stacking to improve model performance: A comprehensive guide on ensemble learning in Python," *Medium*, 2023.

[25] A. J. Zulfikar, M. Y. Yaakob, and R. Syah, "Application of E-glass jute hybrid laminate composite with curved shape on compressive strength of cylindrical column concrete," *Journal of Applied Engineering and Technological Science*, vol. 5, no. 1, pp. 184-196, 2023.

[26] D. Thakran, "Boosting model accuracy with ensemble learning," 2024. https://medium.com/@thakrandisharth/528boosting-model-accuracy-with-ensemble-learning-2742f360ae0

[27] J. Dey, R. Cheruku, A. Srinivas, I. Kavati, B. Vijayasree, and P. Kodali, "Enhancing ensemble models through diversity using k-means for effective diabetes classification," in *2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) (pp. 1-6). IEEE*, 2024.

[28] S. Tiwari, "Complete guide to machine learning evaluation metrics," 2024. https://medium.com/analytics-vidhya/530complete-guide-to-machine-learning-evaluation-metrics-615c2864d916

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 12: 124-147, 2025
DOI: 10.55214/2576-8484.v9i12.11297