Akusta, A. (2026). "Enhanced Loan Approval Prediction Using a Cascaded Machine Learning Framework", Eskişehir Osmangazi Üniversitesi İİBF Dergisi, 21(1), 258 – 278.
https://doi.org/10.17153/oguiibf.1596734

Başvuru/Submission: 05.12.2024 Kabul/Acceptance: 16.02.2025 *Araştırma Makalesi/Research Article*

Enhanced Loan Approval Prediction Using a Cascaded Machine Learning Framework

Ahmet AKUSTA[1]

Çok Katmanlı Makine Öğrenimi Çerçevesi ile Gelişmiş Kredi Onay Tahmini	Enhanced Loan Approval Prediction Using a Cascaded Machine Learning Framework
Öz	**Abstract**
Bu çalışma, kredi onayı tahmini için geleneksel süreçlerin doğasında bulunan verimsizlikleri ele alan çok aşamalı kademeli bir makine öğrenimi çerçevesi sunmaktadır. Çerçeve, her aşamada tahminleri iyileştirmek için Gradient Boosting, Destek Vektör Makinesi ve XGBoost'u birleştirmektedir. Bu yaklaşım, çerçevenin karmaşık ilişkileri ve dengesiz verileri ele almak için her tekniğin güçlü yönlerinden yararlanmasını sağlar. Çerçeve, Gradient Boosting'in ve ev sahipliği gibi geleneksel özelliklerin tahmin doğruluğunu artırmadaki kritik rolünü gösteren kapsamlı bir finansal veri kümesi üzerinde test edilmiştir. Araştırma, çerçevenin kredi riski değerlendirmesini artırma, temerrütleri azaltma ve karar vermeyi kolaylaştırma kapasitesini göstermekte ve böylece finansal kurumlara operasyonel verimlilik ve finansal istikrar için sağlam bir araç sağlamaktadır.	This study introduces a multi-stage cascaded machine learning framework for loan approval prediction, which addresses the inefficiencies inherent in traditional processes. The framework combines Gradient Boosting, Support Vector Machine, and XGBoost to refine predictions at each stage. This approach allows the framework to leverage the strengths of each technique in order to handle complex relationships and imbalanced data. The framework was tested on a comprehensive financial dataset, demonstrating the critical role of Gradient Boosting and traditional features like home ownership in improving predictive accuracy. The research illustrates the framework's capacity to augment credit risk assessment, curtail defaults, and facilitate decision-making, providing financial institutions with a robust instrument for operational efficiency and financial stability.
Anahtar Kelimeler: Kredi Onay Tahmini, Basamaklı Makine Öğrenimi, Kredi Risk Değerlendirmesi, Finansal Karar Alma, Risk Yönetimi	**Keywords:** Loan Approval Prediction, Cascaded Machine Learning, Credit Risk Assessment, Financial Decision-Making, Risk Management
JEL Kodları: G21, C53, C45	**JEL Codes:** G21, C53, C45

Statement of Research and Publication Ethics	This study has been prepared in accordance with the principles of scientific research and publication ethics.
Authors' Contributions to the Article	The entire study was prepared by the author.
Conflict of Interest Statement	There is no conflict of interest arising from this study for the author or any third parties.

[1] Assoc. Prof. Dr., Konya Technical University, Rectorate, ahmetakusta@hotmail.com

1. Introduction

The loan approval process is critical in the banking industry, significantly impacting profitability and risk management (Hussain et al., 2024). As the volume of loan applications continues to rise, financial institutions face challenges in efficiently processing these applications while maintaining accuracy and compliance (Deepa et al., 2024). Ineffective assessments may lead to the approval of high-risk loans or the rejection of creditworthy applicants, potentially resulting in financial losses and decreased customer satisfaction (Garg et al., 2024). Therefore, there is a pressing need for more efficient, automated systems to enhance the loan approval process.

Machine learning has emerged as a promising solution to address these challenges in the loan approval domain. Its ability to handle large datasets, capture complex, non-linear relationships, and potentially provide faster, more accurate decisions offers significant advantages over traditional statistical methods (Sarkar et al., 2024). Machine learning models can analyze vast amounts of applicant data to identify patterns indicative of creditworthiness, thereby improving decision-making efficiency and possibly minimizing risks associated with lending (Alessi & Savona, 2021). Banks may reduce the likelihood of defaults by more accurately identifying high-risk applicants and enhancing their overall risk management strategies (Yemmanuru et al., 2024).

Several studies have applied various machine learning techniques to predict loan approval outcomes. Algorithms such as Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, and ensemble methods like Extreme Gradient Boosting (XGBoost) have been utilized to assess loan applications and predict default risks (Deepa et al., 2024; Hussain et al., 2024; Sarkar et al., 2024). While these models have demonstrated improvements over traditional methods, they also face limitations that may affect their practical applicability. For instance, some models may suffer from overfitting, reducing generalizability to new data (Cheng et al., 2012). Moreover, single-model approaches may need help to handle imbalanced datasets, which are common in financial contexts where the number of approved loans significantly exceeds the number of defaults (Lenka et al., 2021). This imbalance can lead to biased models that perform poorly on minority classes, such as defaulted loans.

Ensemble learning techniques have been proposed to address these limitations by combining multiple models and improving predictive performance in loan approval prediction (Pandey et al., 2013). Ensemble methods, such as bagging and boosting, can enhance accuracy by leveraging the strengths of individual models and mitigating their weaknesses (Alagic et al., 2024). For instance, integrating outputs from different classifiers may refine predictions and lead to more robust and reliable results in credit risk assessment (García et al., 2019). Such approaches can help manage data complexity and imbalanced datasets by aggregating the outputs of multiple models specializing in different aspects of the prediction task (Mistry & Mandal, 2024). However, despite these advancements, challenges remain in effectively integrating diverse models and optimizing ensemble configurations to address prediction inconsistencies and improve performance (Liu & Li, 2024).

The primary objective of this study is to develop an enhanced loan approval prediction model utilizing a multi-stage cascaded machine learning framework. This approach aims to improve predictive accuracy by sequentially combining multiple classifiers, thereby leveraging their strengths. By refining predictions at each stage, the framework seeks to enhance the robustness of the model and provide more accurate assessments of loan applications. This study aims to contribute to the existing literature by proposing applying cascading ensemble methods in the context of loan approval prediction, aiming to address challenges associated with complex financial data and imbalanced datasets. Specifically, the research investigates whether a cascaded framework can outperform traditional single-model and ensemble approaches regarding predictive accuracy and robustness. The study aims to fill a gap in the literature regarding applying multi-stage cascaded models in financial risk assessment, which has yet to be extensively explored in previous research.

The practical implications of this research may be significant for financial institutions. An improved loan approval prediction model may enhance risk management by more effectively identifying high-risk applicants, reducing the incidence of loan defaults, and minimizing financial losses (Zhang et al., 2023). Moreover, by automating and refining the loan approval process, institutions might achieve faster decision-making, improving customer experience and operational efficiency (Hussain et al., 2024). On a broader scale, adopting advanced machine learning systems in financial decision-making could contribute to a more stable and efficient financial industry, where data-driven insights support better outcomes for institutions and customers (Alessi and Savona, 2021). This study's findings will offer practical insights into how cascaded machine learning frameworks can be implemented to achieve these benefits.

2. Literature Review

The integration of machine learning (ML) into financial decision-making, particularly in loan approval prediction, has been a significant area of focus in recent years. As the volume of loan applications increases, financial institutions are turning to advanced algorithms to streamline the decision-making process, enhance risk management, and reduce operational inefficiencies. This literature review synthesizes the current state of knowledge, emphasizing advancements, challenges, and emerging trends in applying machine learning to predict loan approvals and manage credit risks.

Machine learning to automate loan approval processes has shown promising results in improving accuracy and efficiency. Gupta et al., (2020) highlighted the benefits of utilizing ML models to reduce manual efforts by leveraging historical data for predicting loan approval outcomes. Their study demonstrated that ML techniques could effectively identify reliable loan applicants, thereby minimizing the risks associated with manual evaluation.

Similarly, Sarisa et al., (2023) reported that the Random Forest algorithm achieved an accuracy of 83.45% in predicting loan approvals, indicating its potential utility in automating and enhancing loan processing systems. In another study, Nagaraj et al. (2023) demonstrated that XGBoost, a gradient-boosting algorithm, excelled in identifying risky applicants, offering higher accuracy rates than traditional statistical methods. These studies collectively suggest the effectiveness of machine learning algorithms in enhancing the precision of loan approval predictions.

Ensemble learning approaches, which combine predictions from multiple models, have consistently demonstrated their effectiveness in financial applications. Meshref (2020) highlighted the superior performance of ensemble methods, such as Bagging and Boosting, in loan approval prediction, achieving an accuracy rate of 83.97%. Aleksandrova and Armianova (2022) emphasized that cost-sensitive ensemble techniques, such as XGBoost with weighted thresholds, are particularly effective in addressing the misclassification of high-risk applicants, thereby optimizing both accuracy and profit.

The potential of cascading classifiers in financial prediction remains an area of exploration. Oliveira et al., (2005) introduced cascading classifiers for handwritten digit recognition, demonstrating their computational efficiency and ability to refine predictions in sequential stages. However, their application to financial data, including loan approvals, still needs to be fully validated. This gap indicates an opportunity for future research to adapt cascading frameworks to handle the unique challenges of financial datasets.

Financial datasets are often characterized by class imbalances, where default or rejection cases are significantly underrepresented. This imbalance poses challenges for standard ML algorithms, which may favor the majority class. Shingi (2020) addressed this issue by integrating SMOTE (Synthetic Minority Over-sampling Technique) into federated learning systems, effectively balancing training data and improving prediction performance. Similarly, Aleksandrova and Armianova (2022) incorporated cost-sensitive techniques to mitigate the impact of imbalances, showing improved classification outcomes in loan approval models. These approaches highlight the critical role of data preprocessing and model adaptation in enhancing the performance of ML models on skewed datasets.

While addressing class imbalances is critical for improving fairness and accuracy, another important dimension in financial risk modeling involves managing uncertainty and evolving data contexts. Some literature has also emphasized the need to incorporate uncertainty modeling and probabilistic reasoning into financial risk assessment and loan approval predictions. Unlike traditional deterministic models, these approaches account for financial data streams' intricate, evolving nature. For example, Borchani et al., (2015) introduced a dynamic Bayesian modeling framework capable of adapting to time-varying features and customer populations, thereby providing more robust and context-sensitive credit risk predictions. Additional advancements in handling uncertainty include adopting fuzzy logic-based rule systems, as Antonelli et al. (2017) demonstrated, which leveraged type-2 fuzzy rule-based classifiers to accommodate imprecise inputs and latent variables. These studies highlight that incorporating uncertainty-aware strategies can deliver more resilient and adaptive risk evaluation tools, essential for navigating today's dynamic financial landscapes.

Interpretability is a crucial aspect of machine learning applications in finance, as stakeholders require transparent and understandable models. Zhang et al. (2023) demonstrated using Shapley Additive Explanations (SHAP) values to explain feature contributions in loan default predictions, offering actionable insights into applicant behavior. Their profit-driven prediction model improved accuracy and provided valuable information for lenders to identify potential defaulters. These studies underscore the need for explainable ML systems in high-stakes financial decision-making.

In addition to model interpretability, leveraging diverse data types has emerged as a significant area of innovation, expanding the scope of traditional financial datasets to include unconventional and unstructured information. Beyond traditional financial metrics, researchers are increasingly exploring data types like supply chain information, textual disclosures, and qualitative signals to complement standard credit records. Luo et al., (2022) developed a credit risk evaluation model incorporating supply chain attributes, enabling lenders to form a more holistic view of a borrower's financial health. In a related vein, Li and Dai, (2024) leveraged Natural Language Processing (NLP) methods to process and interpret large volumes of unstructured data, such as financial statements and market news, thereby uncovering subtle risk indicators often overlooked by classic numeric models. Zhou et al.,

(2024) further reinforced this trend, demonstrating that analyzing extensive sets of raw financial data can identify influential features without requiring a priori assumptions.

Hybrid models, which combine ML techniques with domain-specific tools, are being recognized as powerful approaches for loan approval and risk assessment. Pimcharee and Surinta (2022) explored integrating feature selection methods, such as Chi-square and information gain, with multi-layer perceptrons, achieving notable accuracy improvements in personal loan prediction. Their findings indicated that feature selection enhances the model's ability to identify critical factors influencing loan approval decisions.

Abakarim et al., (2018) proposed a real-time binary classification framework based on deep neural networks, outperforming traditional models in speed and precision. Their deep learning approach enabled instantaneous predictions, aligning with the increasing demand for faster and more reliable financial services. Such advancements demonstrate the growing potential of hybrid frameworks in addressing the complexities of loan approval processes.

While cascading classifiers offer another avenue for hybrid approaches, their application in financial datasets still needs to be improved. Oliveira et al., (2005) primarily focused on non-financial contexts. Future studies could explore how cascading classifier techniques might be adapted to improve computational efficiency and predictive accuracy in loan approval tasks.

Federated learning and reinforcement learning are considered among the most promising trends in financial applications. Oualid et al., (2023) demonstrated the effectiveness of federated learning in maintaining data privacy while enabling collaborative model training across institutions. This approach is particularly relevant in financial contexts where sensitive customer data must be protected.

Santos and Lima (2024) highlighted integrating reinforcement learning with combinatorial optimization in stock trading. While their work focuses on financial markets, similar techniques could inspire innovations in dynamic loan approval systems, potentially enhancing real-time decision-making processes.

Loan prediction systems benefit from task-specific adaptations of ML algorithms. Pimcharee and Surinta (2022) emphasized the importance of feature selection techniques in identifying critical factors for personal loan approval. Utilizing methods like Chi-square and information gain improved the relevance and accuracy of predictive models.

Additionally, real-time decision-making is gaining traction in loan approvals. Abakarim et al. (2018) developed a deep neural network-based system capable of making instantaneous predictions, significantly improving upon traditional methods that rely on batch processing. Such innovations align with the increasing demand for prompt and efficient financial services.

Despite these advancements, there are still challenges that remain. Zhang et al. (2023) noted that many ML models focus on maximizing accuracy without considering profit-oriented metrics, which are critical in financial applications. Incorporating profitability measures into model optimization could enhance their practical utility and align predictive models with business objectives.

While ensemble methods show promise, their computational complexity can limit scalability, especially in large-scale financial institutions. The application of cascading classifiers, while promising, requires further research to adapt them to financial datasets. Oliveira et al. (2005) demonstrated their utility in reducing computational overhead, but they must address their suitability for imbalanced or high-dimensional financial data. Future studies could explore hybrid models that integrate cascading approaches with ensemble methods to address these limitations.

3. Methodology

The methodology follows a structured approach, beginning with selecting the dataset and variables, followed by data preprocessing. Using python programming language, the data is then processed through a sequential machine learning framework consisting of three stages: Gradient Boosting Classifier, Support Vector Machine, and XGBoost Meta-Model. Each stage refines the predictions from the previous one, aiming to enhance model accuracy. The final step involves evaluating the performance of the models using appropriate metrics, aiming to ensure the effectiveness of the proposed framework.

3.1. Dataset and Variables

The dataset employed in this study was retrieved from Kaggle and is available under the Apache 2.0 license (Gupta, 2024). It consists of various financial and demographic features related to loan applicants. The dependent variable is loan status, a binary indicator where 0 represents a loan not approved and 1 represents a loan approved. Independent variables consist of demographic attributes and financial metrics. The variables used in the study are summarized in Table 1.

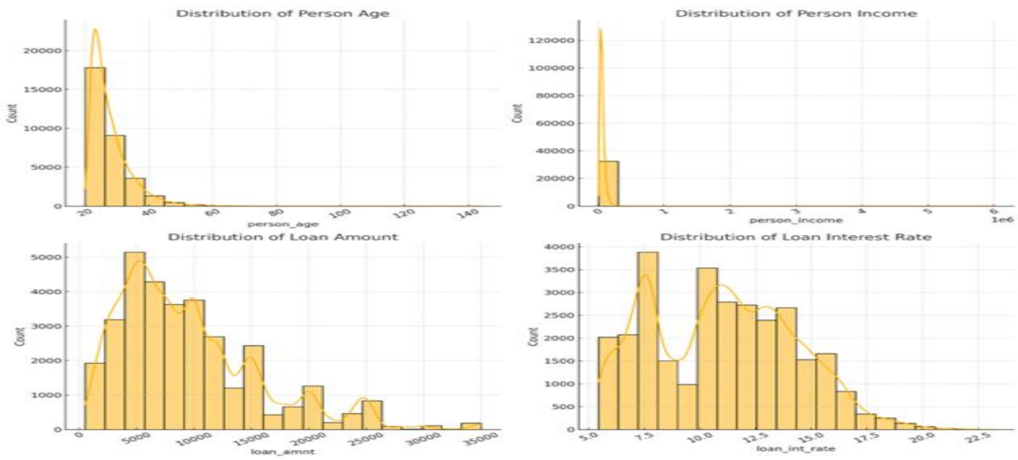
Table 1: Variables in the Dataset

Variable	**Description**	**Data Type**
Person Age	Age of the applicant	Numeric
Person Income	Annual income of the applicant	Numeric
Home Ownership Status	Ownership status of the applicant's residence	Categorical
Loan Amount	Amount of the loan applied for	Numeric
Loan Interest Rate	Interest rate of the loan	Numeric
Loan Grade	Grade assigned to the loan application	Categorical
Loan Intent	Purpose for which the loan is intended	Categorical
Employment Length	Number of years the applicant has been employed	Numeric
Credit Default History	Whether the applicant has defaulted previously	Categorical
Credit History Length	Length of the applicant's credit history	Numeric
Loan Percent Income	Loan amount as a percentage of the applicant's income	Numeric

3.2. Exploratory Data Analysis

Exploratory data analysis was conducted to understand the underlying patterns and relationships within the dataset. Numerical features such as a Person's Age and Income were explored for distribution and potential skewness. The age distribution of loan applicants appeared right-skewed, with a higher concentration of individuals between 20 and 40 years old and fewer applicants above 60. The income distribution also showed a right-skewed pattern, with most applicants earning less than $100,000 annually. A few applicants had significantly higher incomes, which presented as outliers.

Figure 1: Distribution of Person Age and Loan Amount



As shown in Figure 1, Loan amounts ranged from $5,000 to $35,000, with a concentration of around $10,000 to $15,000. Interest rates commonly fell between 7% and 12%, with higher rates being less frequent. The skewness in features like income suggested that data transformations, such as logarithmic scaling, might be considered to normalize the distributions for modeling purposes.

Categorical variables were analyzed to gain insights into the characteristics of the applicants. Most applicants either rented or had a mortgage, with fewer owning their homes outright. Loan intents were mainly for personal use and debt consolidation, with other purposes like education and home improvement being less common. Loan Grades, ranging from A to E, showed that most loans were within grades A to C. Regarding credit history, a substantial number of applicants had no history of default. In contrast, a smaller group had previous defaults.

Figure 2: Home Ownership Status, Loan Intent, and Loan Grade Distribution

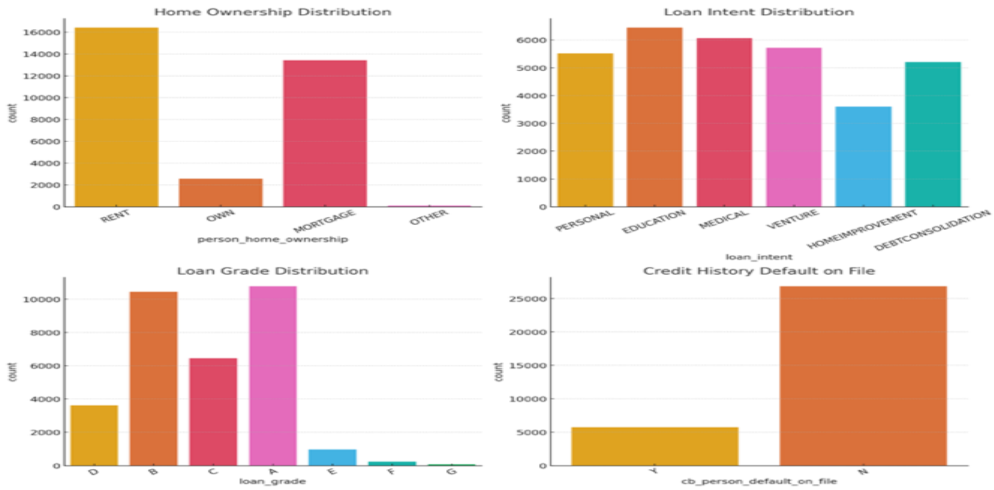
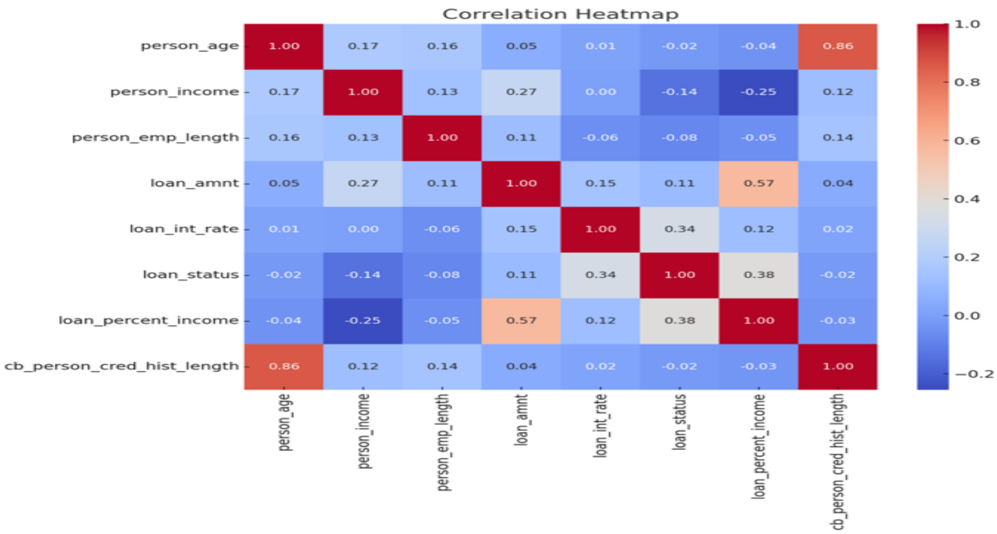


Figure 2 presents the distribution of critical loan-related attributes, offering insights into the characteristics of the borrower population. The "Home Ownership Distribution" indicates that the majority of loan applicants fall under the "MORTGAGE" category, followed closely by those who "RENT," while a smaller proportion identify as "OWN." The "OTHER" category is almost negligible. This distribution suggests that most applicants either have ongoing financial obligations related to housing or rely on rented accommodations, potentially influencing their financial capacity and credit risk. Similarly, the "Loan Intent Distribution" demonstrates that the primary purposes for borrowing are "DEBT CONSOLIDATION," followed by "HOME IMPROVEMENT" and "PERSONAL" loans. This prevalence of debt consolidation loans indicates that many borrowers may be focused on restructuring or managing existing financial liabilities.

Further, the "Loan Grade Distribution" in Figure 2 shows that loans are predominantly assigned high grades, such as "A" and "B," which implies that the majority of applicants are assessed as having relatively favorable credit profiles. In contrast, lower grades, such as "F" and "G," appear infrequent, suggesting that loans to applicants with poor creditworthiness are rare. Finally, the "Credit History Status on File" chart illustrates that most borrowers have an established credit history, with only a tiny proportion lacking this critical financial documentation. This finding emphasizes the importance of credit history in the lending process, potentially reflecting its role in mitigating risk and guiding loan approval decisions.

As shown in Figure 3, a correlation analysis was performed to identify relationships between numerical features. A positive correlation was found between a Person's Age and Credit History Length, indicating that older applicants tend to have longer credit histories. Most other features exhibited weak to moderate correlations.

Figure 3: Correlation Heatmap



Potential data transformations and feature engineering steps were considered based on the exploratory data analysis findings. Addressing skewness in the data and appropriately encoding categorical variables were identified as important steps for improving model performance.

3.3. Data Preprocessing

Handling missing values was an essential step in preparing the data for modeling. The median serves as a robust measure of central tendency, demonstrating resilience to outliers and mitigating distortions in skewed distributions where the mean may be disproportionately influenced by extreme values. Additionally, median imputation helps preserve the original data distribution, reducing the risk of artificial shifts that could bias statistical analyses or machine learning models (Pavithrakannan et al, 2021). Missing values in numerical variables were filled with the median of each respective column, which helps to mitigate the impact of outliers and maintain the central tendency of the data.

Mode imputation stabilizes records with multiple missing values and remains effective even when attributes have weak associations (Nishanth & Ravi, 2016). Furthermore, it maintains the integrity of categorical variables by avoiding the introduction of non-existent categories while offering a straightforward and computationally efficient baseline (Nugroho et al., 2023). Missing values in categorical variables were filled with mode, preserving the most common category in the data.

Categorical variables such as Home Ownership Status, Loan Intent, Loan Grade, and Credit Default History were encoded using label encoding. This method assigns a unique integer to each category within a feature, converting them into a numerical format suitable for machine learning algorithms (Pedregosa et al., 2011). Label encoding was chosen to avoid increasing the dimensionality of the dataset, which can occur with one-hot encoding, especially when the number of categories is relatively small.

The dataset was split into training and test sets, with 80% of the data used to train the machine learning models and 20% reserved for evaluating the performance of unseen data. This split helps ensure the evaluation metrics reflect the models' generalization ability.

All numerical features were standardized using the StandardScaler, which scales the data with a mean of zero and a standard deviation of one.

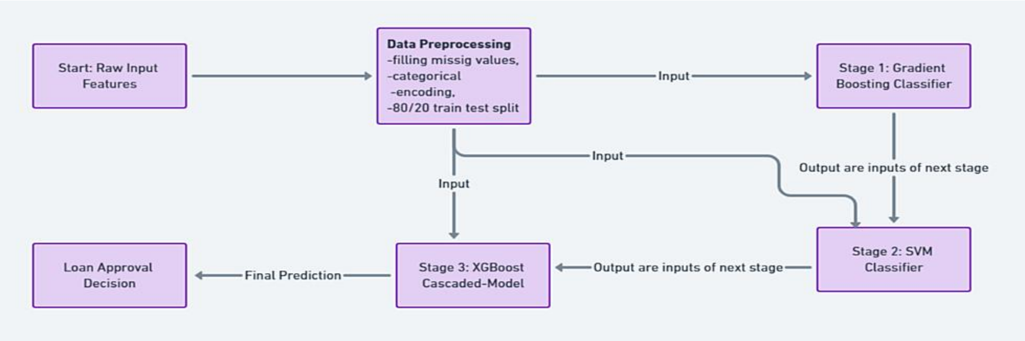
3.4. Cascaded Machine Learning Framework

Cascaded machine learning frameworks aim to improve predictive performance by sequentially integrating multiple models, with the output of one model serving as additional input features for the next (Wolpert, 1992). This approach uses the strengths of different algorithms to correct errors from previous stages and improve overall accuracy. By combining models in this way, cascaded frameworks can capture complex patterns in the data that may be missed by a single model, thus reducing generalisation error and improving predictive capabilities (Breiman, 2001).

This methodology is similar to stacked generalisation, where higher-level models are trained to correct the biases of lower-level models (Wolpert, 1992). By using the predictions of earlier models as inputs to later ones, the framework potentially addresses the limitations of individual algorithms and increases the robustness of the overall model. This strategy is consistent with resource efficiency in cascading use, where outputs from one process become inputs to another, maximising overall efficiency (Rehberger & Hiete, 2020).

Implementing a cascading framework allows for integrating different machine learning techniques, each contributing uniquely to predictive performance. This multistep approach can improve accuracy in complex prediction tasks by leveraging the diverse strengths of different algorithms (Chen & Guestrin, 2016). Furthermore, cascaded frameworks are adaptable and can be tailored to specific problems, making them a versatile tool in machine learning applications (Ke et al., 2017).

Figure 4: Cascaded Machine Learning Framework Workflow



The study implemented a three-stage cascaded machine learning framework to improve predictive performance potentially, as shown in Figure 4. In the first stage, a gradient-boosting classifier was used to sequentially build an ensemble of weak learners, with each successive model attempting to correct the errors of the previous one. The predicted credit approval probabilities from this model, called gradient boosting predictions, were used as additional features in the following stages.

The second stage used a support vector machine. The original data set, augmented with the gradient boosting predictions, was used to find the optimal hyperplane separating the classes. The predictions from the Support Vector Machine (SVM) model were included as features in the final stage.

In the third and final stage, an XGBoost classifier was used as a metamodel. This model integrated all the features from the previous stages, including the gradient boosting and SVM predictions, to make the final credit approval predictions.

3.5. Model Structures

The proposed multi-stage cascaded machine learning framework aims to enhance loan approval prediction by leveraging three classifiers sequentially: Gradient Boosting, SVM, and XGBoost Meta-Model. Each stage uses the predictions from the preceding stage as additional features to improve the robustness and reliability of the final predictions. Cross-validation techniques were integrated into the evaluation process to ensure the models generalize well to unseen data. This enables a thorough assessment of each model's performance across multiple data splits.

The first stage employs a Gradient Boosting Classifier, which captures complex non-linear patterns and interactions within the data. This model is particularly effective for feature selection and reducing bias due to its iterative nature. The Gradient Boosting model was trained with default hyperparameters, including a learning rate of 0.1 and a maximum depth of 3. These parameters aim to balance model complexity and training efficiency. The model outputs probabilities of loan approval, which are subsequently used as features in the second stage. While the Gradient Boosting model demonstrated intense precision, it showed limitations in recall, particularly in identifying minority class instances.

In the second stage, an SVM classifier was trained using the original features and the prediction probabilities from the Gradient Boosting model. Including these predictions enhances the SVM's ability to define more precise decision boundaries. The SVM model utilized a radial basis function (RBF) kernel, suitable for non-linear data distributions.

The model was configured with a regularization parameter 1.0 and a scale gamma value, which adjusts the influence of individual data points. Despite high precision, the SVM struggled with recall, indicating it was overly conservative in identifying positive cases. This stage, however, potentially contributed valuable prediction probabilities that were important inputs for the XGBoost meta-model.

The final stage integrates predictions from previous classifiers through an XGBoost meta-model, combining refined outputs to produce the final classification. This stage capitalizes on XGBoost's ability to handle structured data efficiently and its robustness to overfitting.

The meta-model was trained with a maximum depth of 6 and 100 estimators, optimizing the binary classification objective. By incorporating predictions from both Gradient Boosting and SVM models as features, the meta-model achieved significant improvements in precision and recall. This stage effectively addressed the limitations of the earlier models, particularly by enhancing recall.

Cross-validation was implemented across all stages of the cascaded framework using StratifiedKFold with five splits. This method ensures that each fold preserves the original distribution of the target variable, mitigating issues related to class imbalance (Pedregosa et al., 2011). By evaluating the models on multiple training and testing splits, cross-validation provides a robust assessment of their performance and reduces the risk of overfitting to a single data partition. The consistent evaluation across folds confirmed the reliability of the cascaded framework, with the XGBoost meta-model demonstrating the most balanced performance in terms of precision and recall across all splits. This validation technique underscores the model's ability to generalize effectively to unseen data, making it a suitable approach for real-world applications in financial decision-making.

3.6. Results

The performance of the models was carefully evaluated using established classification metrics, including accuracy, precision, recall, and F1-score. These metrics were chosen for their ability to provide a balanced and comprehensive assessment of model performance. Accuracy represents the proportion of correct predictions across all samples, offering a general view of model efficacy. Precision, which measures the proportion of correctly predicted positive cases among all predicted positives, provides insights into the model's ability to avoid false positives. Recall evaluates the proportion of actual positive cases that were correctly identified, reflecting the model's sensitivity to actual cases. The F1-score, calculated as the harmonic mean of precision and recall, is particularly valuable in capturing the trade-offs between these two metrics, especially in imbalanced datasets.

Table 2: Classification Report (No Cross Validation)

Model	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	87.60%	80.08%	58.69%	67.73%
Support Vector Machine	80.34%	81.54%	14.67%	24.87%
XGBoost Meta-Model	92.83%	93.43%	72.80%	81.84%

Table 2 presents the results of the models without cross-validation. The Gradient Boosting model achieved an accuracy of 87.60%, with a precision of 80.08% and a recall of 58.69%. These metrics highlight the model's ability to identify patterns effectively, although its relatively lower recall indicates limitations in correctly identifying all positive cases, as reflected in its F1-score of 67.73%. The Support Vector Machine (SVM) model demonstrated lower performance, achieving an accuracy of 80.34% and a precision of 81.54%. However, its recall of 14.67% and an F1-score of 24.87% revealed significant challenges in identifying positive cases. In contrast, the XGBoost meta-model achieved the highest performance, with an accuracy of 92.83%, precision of 93.43%, recall of 72.80%, and an F1-score of 81.84%.

Figure 5: Confusion Matrix for Classification (No Cross Validation)

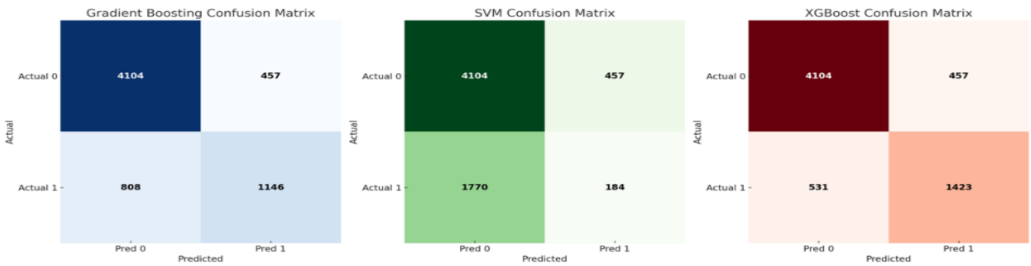


Figure 5 provides further insight into the classification performance without cross-validation. The Gradient Boosting model produced 597 false negatives (FN) against 848 true positives (TP), reflecting its recall limitations. The SVM model exhibited a stark imbalance, with 1233 FN and only 212 TP, corroborating its exceptionally low recall (14.67%). In contrast, the XGBoost meta-model demonstrated a more balanced profile, with 393 FN and 1052 TP, aligning with its higher recall (72.80%) and F1-score (81.84%). These results suggest that the XGBoost model was able to integrate information from the previous models effectively, addressing their weaknesses and enhancing overall predictive performance. These results suggest that the XGBoost model was able to integrate information from the previous models effectively, addressing their weaknesses and enhancing overall predictive performance.

Table 3: Classification Report (With Cross Validation)

Model	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	0.8791	0.8066	0.5868	0.6793
Support Vector Machine	0.7985	0.8619	0.0946	0.1671
XGBoost Meta-Model	0.9310	0.9418	0.7287	0.8216

The models demonstrated similar trends when evaluated under cross-validation, as shown in Table 3. The Gradient Boosting model achieved a slightly improved accuracy of 87.91%, with a precision of 80.66% and a recall of 58.68%, leading to an F1-score of 67.93%. While this model performed well in identifying general patterns, its recall suggests it still struggled to identify some positive cases. The SVM model, under cross-validation, demonstrated an accuracy of 79.85% and an increased precision of 86.19%. However, its recall remained low at 9.46%, resulting in an F1-score of 16.71%, reinforcing the model's challenges in handling positive cases independently. The XGBoost meta-model, again, outperformed the other models, achieving an accuracy of 93.10%, precision of 94.18%, recall of 72.87%, and an F1-score of 82.16%.

Figure 6: Confusion Matrix for Classification (With Cross Validation)

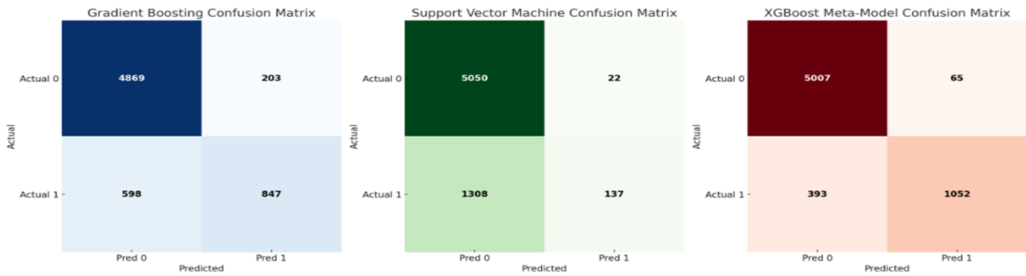


Figure 6 illustrates the cross-validated confusion matrices. The Gradient Boosting model showed minimal deviation from its non-validated performance, with 598 FN and 847 TP, consistent with its stable recall (58.68%). The SVM model's challenges intensified under cross-validation, as evidenced by 1308 FN and only 137 TP, which aligns with its further reduced recall (9.46%). The XGBoost meta-model maintained robust performance, with 393 FN and 1052 TP, mirroring its results in Figure 4. This consistency underscores its reliability across validation methods and its capacity to generalize effectively. These metrics highlight the meta-model's ability to synthesize predictions from the earlier models, leveraging their complementary strengths to deliver a more robust and balanced performance.

The progression of model performance across the stages demonstrates the effectiveness of the cascaded modeling approach. In the initial stage, the Gradient Boosting model laid a strong foundation by capturing general trends in the data, but its recall limitations highlighted the need for further refinement. While exhibiting intense precision, the SVM model struggled with recall, suggesting that it was more suited for enhancing specificity rather than sensitivity. The XGBoost meta-model, as the final stage, capitalized on the insights provided by the preceding models, achieving significant improvements in accuracy, precision, and recall. The resulting F1-score of 82.16% demonstrates a balanced performance, indicating the meta-model's ability to classify positive and negative cases effectively. As detailed in Tables 2 and 3, these findings underscore the importance of integrating diverse predictive approaches in a cascaded framework to optimize overall model performance. This approach suggests the potential of meta-modeling to address the inherent limitations of individual models while potentially leveraging their strengths.

3.7. Feature Importance

Analyzing the importance of features in the XGBoost meta-model provides insights into which variables most significantly influenced the final predictions. The importance scores are presented in Table 4.

Table 4: Feature Importance in XGBoost Meta-Model

Feature	Importance Score
Gradient Boosting Predictions	0.4267
Home Ownership Status	0.1487
Loan Grade	0.0715
Person Income	0.0656
Loan Intent	0.0644
Loan Percent Income	0.0556
Loan Interest Rate	0.0307
Employment Length	0.0297
Person Age	0.0242
SVM Predictions	0.0242
Loan Amount	0.0211
Credit History Length	0.0209
Credit Default History	0.0166

The Gradient Boosting Predictions had the highest importance score of 0.4267, indicating that this feature contributed most significantly to the XGBoost meta-model's decisions. This suggests that the information encapsulated in the Gradient Boosting model's predictions was highly valuable for the final prediction task.

Home Ownership Status and Loan Grade were the most important features, with importance scores of 0.1487 and 0.0715, respectively. These features are traditionally significant in loan approval decisions, reflecting an applicant's stability and creditworthiness.

The SVM Predictions had a lower importance score of 0.0242, comparable to features like Person Age and Loan Amount. While the SVM Predictions contributed to the model, their impact was less pronounced than the Gradient Boosting Predictions.

4. Discussion

The progression in accuracy observed in the cascaded model demonstrates the cumulative benefits achieved by integrating predictions from multiple machine learning models, each of which contributes its unique strengths and addresses the weaknesses of the others. The Gradient Boosting model, which serves as the foundation for subsequent models, demonstrated strong performance, with an accuracy of 87.60%. However, the relatively lower recall indicates potential difficulties in identifying specific positive cases. This could be attributed to an inherent bias towards the majority class, a well-documented issue in imbalanced datasets (Shingi, 2020). The model's robustness is consistent with the findings of Zhang et al. (2023), who highlight the effectiveness of Gradient Boosting in capturing complex patterns, particularly in loan default prediction tasks.

In contrast, the standalone Support Vector Machine (SVM) exhibited weaker performance, achieving an accuracy of 80.34% and a recall of only 14.67%. While Pimcharee and Surinta (2022) found SVM effective in personal loan datasets without extensive feature selection or preprocessing, its performance can vary depending on dataset complexity and non-linear relationships. In this study, the SVM's poor recall suggests it struggled to generalize adequately, potentially due to imbalanced data exacerbating the model's focus on the majority class (Aleksandrova & Armianova, 2022). This aligns with observations in the literature that kernel transformations and tailored feature engineering are often necessary for SVMs to handle non-linear patterns effectively.

The XGBoost meta-model brought substantial improvement, achieving an accuracy of 92.83% and a recall of 72.80%. This improvement reflects the model's ability to mitigate the limitations of the prior models, particularly by incorporating predictions from the Gradient Boosting model, which emerged as features of high importance. Nagaraj et al. (2023) and Zhang et al. (2023) demonstrated XGBoost's capability to refine predictions and improve recall in financial risk prediction tasks. These findings underscore the value of leveraging robust base models, such as Gradient Boosting, to feed into ensemble frameworks.

The relatively lower importance of SVM predictions in the meta-model suggests that while SVM contributed to diversity, its impact was less significant than Gradient Boosting. This aligns with Meshref (2020), who demonstrated that even classifiers with lower individual performance can provide valuable diversity to an ensemble model, enhancing overall robustness. The cascading framework utilized diverse inputs, which is in line with the observations of Oliveira et. al. (2005), who highlighted the potential computational efficiency and prediction refinement obtained through sequential model outputs.

Features such as Home Ownership Status and Loan Grade played critical roles in the final model, with high-importance scores reflecting their relevance as indicators of financial stability and creditworthiness. These findings align with Gupta et al. (2020), who emphasized the significance of traditional financial risk indicators, and (Ding, 2023), who advocated for integrating domain-specific knowledge into machine learning models to improve predictions. Incorporating these features demonstrates the cascading model's ability to combine advanced machine-learning techniques with domain knowledge.

Despite the progress made, specific challenges related to the models' performance and generalization capabilities persist. The Gradient Boosting model's initial limitations in recall suggest the potential for further optimization, such as hyperparameter tuning or advanced preprocessing techniques, to enhance its ability to identify positive cases. While the XGBoost meta-model effectively reduced false negatives, its reliance on robust base models underscores the critical role of strong foundational classifiers in cascading frameworks.

The study also points out the trends in financial machine learning, where ensemble and hybrid approaches have shown potential to outperform standalone models in certain contexts (Aleksandrova and Armianova, 2022; Zhang et al., 2023). Future research could explore incorporating interpretability frameworks, such as SHAP values, to enhance the interpretability of predictions, as recommended by Zhang et al. (2023). Additionally, advanced techniques like federated learning, as demonstrated by Oualid et al. (2023), have the potential to enhance privacy and robustness in loan prediction systems.

In conclusion, the multi-stage cascaded machine learning framework demonstrated the potential to improve loan approval prediction accuracy. By leveraging the strengths of different algorithms, the approach may offer a viable solution to the challenges associated with complex financial data and imbalanced datasets. Further research involving diverse datasets, interpretability techniques, and privacy-preserving methods could enhance the applicability of this framework in real-world financial contexts.

5. Conclusion

This study investigated the effectiveness of a multi-stage cascaded machine learning framework for predicting loan approvals, addressing challenges inherent in traditional methods. By sequentially integrating Gradient Boosting, Support Vector Machine (SVM), and

XGBoost models, we observed a notable improvement in predictive accuracy, with the XGBoost meta-model reaching approximately 92.83% accuracy. This advancement suggests that the framework may effectively capture complex patterns within financial data, mainly benefiting from the high importance of Gradient Boosting predictions.

Our findings indicate that ensemble methods can enhance credit risk assessment by leveraging advanced algorithms and traditional financial features like Home Ownership Status and Loan Grade. These results have practical implications for financial institutions, potentially leading to better risk management, reduced default rates, and more efficient loan processing.

The study may have limitations as the dataset might not fully represent the complexities and general trends of real-world financial data, which could result in challenges such as data imbalance and limited model interpretability. Future research should focus on applying this framework to more diverse and representative datasets, enhancing interpretability through explainable AI techniques, and exploring privacy-preserving methods like federated learning.

In conclusion, the multi-stage cascaded machine learning framework presents a promising approach to improving loan approval predictions. Addressing the complexities associated with financial data analysis and imbalanced datasets is a crucial step that can contribute to advancing predictive modeling in finance, but it may also pose challenges that need to be carefully managed. Ongoing refinement and validation of this framework could impact financial risk assessment practices and operational efficiencies within the industry.

References

Abakarim, Y., Lahby, M., & Attioui, A. (2018). Towards an efficient real-time approach to loan credit approval using deep learning. *9th International Symposium on Signal, Image, Video and Communications, ISIVC 2018 - Proceedings,* 306–313. https://doi.org/10.1109/ISIVC.2018.8709173

Alagic, A., Zivic, N., Kadusic, E., Hamzic, D., Hadzajlic, N., Dizdarevic, M., & Selmanovic, E. (2024). Machine learning for an enhanced credit risk analysis: A comparative study of loan approval prediction models integrating mental health data. *Machine Learning and Knowledge Extraction, 6*(1), 53–77. https://doi.org/10.3390/make6010004

Aleksandrova, Y., & Armianova, M. (2022). Evaluation of cost-sensitive machine learning methods for default credit prediction. *International Conference Automatics and Informatics, ICAI 2022 - Proceedings,* 89–94. https://doi.org/10.1109/ICAI55857.2022.9960023

Alessi, L., & Savona, R. (2021). Machine learning for financial stability. In *Data science for economics and finance: Methodologies and applications* (pp. 65–87). https://doi.org/10.1007/978-3-030-66891-4_4

Antonelli, M., Bernardo, D., Hagras, H., & Marcelloni, F. (2017). Multiobjective evolutionary optimization of type-2 fuzzy rule-based systems for financial data classification. IEEE Transactions on Fuzzy Systems, 25(2), 249–264. https://doi.org/10.1109/TFUZZ.2016.2578341

Borchani, H., Martínez, A. M., Masegosa, A. R., Langseth, H., Nielsen, T. D., Salmerón, A., Fernández, A., Madsen, A. L., & Sáez, R. (2015). Dynamic Bayesian modeling for risk prediction in credit operations. Frontiers in Artificial Intelligence and Applications, 278, 17–26. https://doi.org/10.3233/978-1-61499-589-0-17

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13–17 August 2016,* 785–794. https://doi.org/10.1145/2939672.2939785

Cheng, C., Xu, W., & Wang, J. (2012). A comparison of ensemble methods in financial market prediction. *Proceedings of the 2012 5th International Joint Conference on Computational Sciences and Optimization, CSO 2012,* 755–759. https://doi.org/10.1109/CSO.2012.171

Deepa, M., Pal, S., & Ghusey, V. P. (2024). Monetary loan eligibility prediction using logistic regression algorithm. *2nd International Conference on Emerging Trends in Information Technology and Engineering, Ic-ETITE 2024.* https://doi.org/10.1109/ic-ETITE58242.2024.10493584

Ding, Z. (2023). Construction and exploration of a financial risk control model based on machine learning. *2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques, EASCT 2023.* https://doi.org/10.1109/EASCT59475.2023.10393547

García, V., Marqués, A. I., & Sánchez, J. S. (2019). Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction. *Information Fusion, 47,* 88–101. https://doi.org/10.1016/j.inffus.2018.07.004

Garg, D., Shelke, N. A., Kitukale, G., & Mehlawat, N. (2024). Leveraging financial data and risk management in banking sector using machine learning. *2024 IEEE 9th International Conference for Convergence in Technology, I2CT 2024.* https://doi.org/10.1109/I2CT61223.2024.10544336

Gupta, A., Pant, V., Kumar, S., & Bansal, P. K. (2020). Bank loan prediction system using machine learning. *Proceedings of the 2020 9th International Conference on System Modeling and Advancement in Research Trends, SMART 2020,* 423–426. https://doi.org/10.1109/SMART50582.2020.9336801

Gupta, H. (2024). *Credit Risk Dataset*, Version 1. Retrieved October 5, 2024, from https://www.kaggle.com/datasets/harshg97/credit-risk-dataset/versions/1

Hussain, M. Z., Ejaz, S., Batool, E., Hasan, M. Z., Mustafa, M., Khalid, A., Hussain, U., Khan, Z., Javaid, A., Ashraf, M. F., Awan, R., & Yaqub, M. A. (2024). Bank loan prediction system using machine learning models. *2024 IEEE 9th International Conference for Convergence in Technology, I2CT 2024.* https://doi.org/10.1109/I2CT61223.2024.10543786

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 2017-December,* 3147–3155. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85038215940&partnerID=40&md5=e7f6611aa423276ab95bd73aa45ad93e

Lenka, S. R., Pant, M., Barik, R. K., Patra, S. S., & Dubey, H. (2021). Investigation into the efficacy of various machine learning techniques for mitigation in credit card fraud detection. *Advances in Intelligent Systems and Computing, 1176,* 255–264. https://doi.org/10.1007/978-981-15-5788-0_24

Li, T., & Dai, X. (2024). Financial risk prediction and management using machine learning and natural language processing. International Journal of Advanced Computer Science and Applications, 15(6), 211–219. https://doi.org/10.14569/IJACSA.2024.0150623

Liu, Z., & Li, D. (2024). Research of Dempster-Shafer's theory and ensemble classifier financial risk early warning model based on Benford's law. *Computational Economics.* https://doi.org/10.1007/s10614-024-10679-1

Luo, S., Xing, M., & Zhao, J. (2022). Construction of artificial intelligence application model for supply chain financial risk assessment. *Scientific Programming*, 2022. https://doi.org/10.1155/2022/4194576

Meshref, H. (2020). Predicting loan approval of bank direct marketing data using ensemble machine learning algorithms. *International Journal of Circuits, Systems and Signal Processing, 14,* 914–922. https://doi.org/10.46300/9106.2020.14.117

Mistry, K. A. J., & Mandal, B. (2024). Unified deep ensemble architecture for multiple classification tasks. *Lecture Notes in Networks and Systems, 1065 LNNS,* 544–557. https://doi.org/10.1007/978-3-031-66329-1_35

Nagaraj, P., Nikhil, K., Sai Ram Santosh Babu, K. V. S., Reddy, D. H. T., Sekar, R. R., & Rajkumar, T. D. (2023). Loan prediction analysis using innumerable machine learning algorithms. *2023 International Conference on Data Science, Agents and Artificial Intelligence, ICDSAAI 2023.* https://doi.org/10.1109/ICDSAAI59313.2023.10452622

Nishanth, K. J., & Ravi, V. (2016). Probabilistic neural network based categorical data imputation. *Neurocomputing, 218*, 17-25. https://doi.org/10.1016/j.neucom.2016.08.044

Nugroho, H., Utama, N. P., & Surendro, K. (2023). Smoothing target encoding and class center-based firefly algorithm for handling missing values in categorical variable. *Journal of Big Data*, 10(1), 10. https://doi.org/10.1186/s40537-022-00679-z

Oliveira, L. S., Britto Jr., A. S., & Sabourin, R. (2005). Improving cascading classifiers with particle swarm optimization. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 2005,* 570–574. https://doi.org/10.1109/ICDAR.2005.138

Oualid, A., Maleh, Y., & Moumoun, L. (2023). Federated learning techniques applied to credit risk management: A systematic literature review. *EDPACS, 68*(1), 42–56. https://doi.org/10.1080/07366981.2023.2241647

Pandey, T. N., Jagadev, A. K., Choudhury, D., & Dehuri, S. (2013). Machine learning-based classifiers ensemble for credit risk assessment. *International Journal of Electronic Finance, 7*(3–4), 227–249. https://doi.org/10.1504/IJEF.2013.058604

Pavithrakannan, R., Fenn, N. B., Raman, S., Kalyanaraman, V., Murugananthan, V. K., & Janarthanan, J. (2021). Imputation analysis of central tendencies for classification. *2021 IEEE International IOT,* *Electronics and Mechatronics Conference (IEMTRONICS)*, 1-7. https://doi.org/10.1109/IEMTRONICS52119.2021.9422507

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pimcharee, K., & Surinta, O. (2022). Data mining approaches in personal loan approval. *Engineering Access, 8*(1), 15–21. https://doi.org/10.14456/mijet.2022.2

Rehberger, M., & Hiete, M. (2020). Allocation of environmental impacts in circular and cascade use of resources: Incentive-driven allocation as a prerequisite for cascade persistence. *Sustainability (Switzerland), 12*(11). https://doi.org/10.3390/su12114366

Santos, G. D., & Lima, K. R. (2024). Impact of combinatorial optimization on reinforcement learning for stock trading in financial markets. *ACM International Conference Proceeding Series.* https://doi.org/10.1145/3658271.3658282

Sarisa, H. K., Khurana, V., Koti, V. C., & Garg, N. (2023). Loan prediction using machine learning. *2023 International Conference on Advances in Computation, Communication and Information Technology, ICAICCIT 2023,* 194–198. https://doi.org/10.1109/ICAICCIT60255.2023.10466049

Sarkar, T., Rakhra, M., Sharma, V., & Singh, A. (2024). An empirical comparison of machine learning techniques for bank loan approval prediction. *Proceedings of International Conference on Communication, Computer Sciences and Engineering, IC3SE 2024,* 137–143. https://doi.org/10.1109/IC3SE62002.2024.10593355

Shingi, G. (2020). A federated learning-based approach for loan defaults prediction. *IEEE International Conference on Data Mining Workshops, ICDMW, 2020-November,* 362–368. https://doi.org/10.1109/ICDMW51313.2020.00057

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Yemmanuru, P. K., Yeboah, J., & Nti, I. K. (2024). Customer credit risk: Application and evaluation of machine learning and deep learning models. *2024 IEEE 3rd International Conference on Computing and Machine Intelligence, ICMI 2024 - Proceedings.* https://doi.org/10.1109/ICMI60790.2024.10585896

Zhang, L., Wang, J., & Liu, Z. (2023). What should lenders be more concerned about? Developing a profit-driven loan default prediction model. *Expert Systems with Applications, 213.* https://doi.org/10.1016/j.eswa.2022.118938

Zhou, Y., Xiao, Z., Gao, R., & Wang, C. (2024). Using data-driven methods to detect financial statement fraud in the real scenario. *International Journal of Accounting Information Systems, 54*. https://doi.org/10.1016/j.accinf.2024.100693

Extended Summary

Enhanced Loan Approval Prediction Using a Cascaded Machine Learning Framework

Loan approval prediction plays a crucial role in financial risk management, significantly influencing operational efficiency and institutional profitability. Traditional methods often face challenges in managing the complexities of financial data and addressing class imbalances, such as the overrepresentation of approved loans compared to defaults. This study introduces a multistage cascaded machine learning framework to address these issues using a novel approach. By integrating Gradient Boosting, Support Vector Machine (SVM), and XGBoost sequentially, the framework leverages the unique strengths of each model to refine predictions and improve overall performance.

Machine learning has revolutionized financial decision making by providing the ability to process large datasets, identify nonlinear patterns, and make accurate predictions. However, single model approaches frequently encounter issues such as overfitting and poor generalization, mainly when applied to imbalanced datasets. Although ensemble methods, such as boosting and bagging, have improved predictive accuracy, they remain limited in handling data complexity and class disparities. The proposed framework builds on these methods, employing a cascading structure to optimize model performance and address their inherent weaknesses.

The study utilized an open-source financial dataset from Kaggle, which includes loan applicants' demographic and financial attributes. The dependent variable was binary, indicating whether a loan application was approved or denied. In contrast, the independent variables included factors such as applicant age, income, credit history length, and loan-specific details. Exploratory data analysis revealed important insights, such as right-skewed income distributions and correlations between applicant age and credit history length, which informed the preprocessing strategy.

To address data quality issues, missing values were imputed using the median for numerical features and the mode for categorical features. Categorical variables, including home ownership, loan grade, and loan intent, were label encoded to preserve their ordinal relationships without increasing dataset dimensionality. Numerical features were standardized to ensure compatibility with algorithms sensitive to feature scaling, such as SVM. The dataset was split into training and testing subsets, with 80% allocated for training and 20% for testing.

The proposed framework consists of three sequential stages, with each model contributing uniquely to overall predictive performance. In the first stage, Gradient Boosting was used to capture intricate data relationships. This model iteratively improved predictions by focusing on errors from prior iterations, achieving an accuracy of 87.60%, precision of 80.08%, and recall of 58.69%. While the model demonstrated strong precision, its limited recall indicated a bias toward the majority class.

In the second stage, SVM was employed to refine decision boundaries using the original features and predictions from the Gradient Boosting model. This stage demonstrated a precision of 81.54%, reflecting its ability to avoid false positives. However, its recall was notably low at 14.67%, indicating difficulty identifying minority class instances. Despite these challenges, SVM contributed valuable predictions later synthesized by the final stage.

The third stage used XGBoost as a metamodel to integrate predictions from the previous stages, combining the strengths of Gradient Boosting and SVM. This stage achieved superior performance, with an accuracy of 92.83%, precision of 93.43%, recall of 72.80%, and an F1score of 81.84%. By addressing the limitations of earlier models, XGBoost provided a balanced and robust output, demonstrating the effectiveness of the cascaded framework.

The performance of the cascaded framework was assessed using key classification metrics, including accuracy, precision, recall, and F1score. Without cross validation, the Gradient Boosting model struggled with recall, achieving 58.69%, despite a reasonable accuracy of 87.60%. SVM showed high precision at 81.54% but had poor recall at 14.67%. The XGBoost metamodel outperformed both, achieving 92.83% accuracy, 93.43% precision, and 72.80% recall.

Under cross validation, similar trends emerged. Gradient Boosting achieved an accuracy of 87.91% and a recall of 58.68%, while SVM maintained high precision at 86.19% but struggled with recall, achieving only 9.46%. The XGBoost metamodel consistently excelled, with 93.10% accuracy, 94.18% precision, 72.87% recall, and a F1score of 82.16%.

Analyzing the importance of features in the XGBoost metamodel provided valuable insights. Predictions from the Gradient Boosting model significantly influenced the final decisions, showcasing its strength in identifying complex patterns. Traditional financial features, such as home ownership status and loan grade, also emerged as critical predictors, emphasizing the importance of incorporating domain-specific knowledge into machine learning models. Predictions from the SVM model contributed less substantially, highlighting its auxiliary role in refining decision boundaries within the cascaded framework.

This study offers several practical contributions to financial institutions. The proposed framework enhances credit risk assessment by improving loan approval predictions, reducing default rates, and increasing operational efficiency. Its ability to handle class imbalances ensures a balanced and reliable predictive model. The research also advances ensemble learning by demonstrating how the sequential integration of diverse models can address the limitations of standalone algorithms and enhance predictive accuracy. Moreover, by combining advanced machine learning techniques with traditional financial features, the framework underscores the critical role of interpretability and domain knowledge in high-stakes financial decision-making.

Despite its promising results, the study may have some limitations, as the dataset might not fully capture the complexities and dynamics of real-world financial ecosystems, potentially affecting the framework's broader applicability. Additionally, the computational requirements of the cascading framework could pose challenges for smaller institutions with limited resources.

Future research should explore testing the framework on diverse and representative datasets to assess its generalizability. Enhancing recall through advanced preprocessing techniques, such as targeted imputation strategies and anomaly detection, could improve its performance on imbalanced data. Incorporating interpretability tools like SHAP values could increase transparency and stakeholder trust by providing clear insights into prediction processes. Federated learning offers another promising direction, enabling institutions to collaborate on improving models without compromising data privacy. Finally, combining cascading frameworks with explainable AI methods could balance performance and interpretability, fostering accuracy and user confidence.

The multistage cascaded machine learning framework significantly advances loan approval prediction. By leveraging the strengths of Gradient Boosting, SVM, and XGBoost in a sequential structure, the framework addresses longstanding challenges in financial data analysis, such as complexity and class imbalances. Its superior performance compared to traditional models highlights its potential as a robust tool for financial institutions.