



MACHINE LEARNING FOR CONSUMER LOAN DEFAULT PREDICTION IN RETAIL BANKING

Đorđe Pucar^{1*},
[0000-0001-7445-0758]

Dragan Soleša²
[0000-0003-1185-7969]

¹Faculty of Applied
Management,
Economics and Finance,
University Business Academy in
Novi Sad,
Belgrade, Serbia

²Faculty of Economics and
Engineering Management,
University Business Academy in
Novi Sad,
Novi Sad, Serbia

Abstract:

Consumer loan defaults represent consistent source of cost pressures within retail banks, eroding capital reserves and necessitating provisioning that prevents further loaning. While machine learning is now widely adopted in credit risk modelling, empirical studies report mixed evidence regarding the extent of its performance improvements over Logistic Regression. These differences may partly reflect variation in datasets, pre-processing methods, and evaluation metrics. In this paper, a comparison of Logistic Regression, Random Forest, and Gradient Boosting using a publicly available consumer lending dataset containing 32,575 rows and eleven variables is conducted. Evaluation was carried out using an 80/20 holdout validation and additionally five-fold cross-validation. Gradient Boosting outperformed both alternatives across all evaluation metrics, with ensemble methods demonstrating higher recall for borrowers than the Logistic Regression baseline. Loan grade, interest rate, and the ratio of the loan to income accounted for prediction in both ensemble classifiers. The recall gap between Gradient Boosting and Logistic Regression carries direct implications for the design of automated credit approval systems.

Keywords:

credit risk assessment; probability of default; ensemble classifiers; credit scoring; class imbalance.

JEL Classification:

E51, G21, G32

INTRODUCTION

The retail loan decision-making process is important in both directions. Lending to borrowers who are unable to repay their loans can result in losses, increased requirements, and a buildup of non-performing loans. The banking sector generated comparable challenges during the 2008–2009 financial crisis. Striking the right balance requires reliable default prediction, and the Basel II and Basel III frameworks formalized this expectation by requiring banks to maintain internally validated probability-of-default models under the Internal Ratings-Based approach (Basel Committee on Banking Supervision, 2006).

Article info:

Received: May 11, 2026

Correction: September 01, 2026

Accepted: September 09, 2026

*E-mail: djordje@mef.edu.rs



Logistic Regression has been the mainstay of consumer credit risk scoring models for decades, prized for interpretability and compliance purposes over predictive performance. Altman (1968) established that balance-sheet ratios could discriminate between failing and solvent firms with reasonable accuracy, and this paradigm carried into consumer lending via scorecard development (Thomas, 2000; Siddiqi, 2012). The drawback of logistic regression is clear: it forces the decision boundary to be linear in feature space; however, as borrower heterogeneity increases and non-linear interactions arise, a linear boundary becomes increasingly inadequate. Default probability does not depend on a weighted combination of income, loan amount, and credit history, but on a more complicated interaction among these factors.

Random Forest (Breiman, 2001) and Gradient Boosting (Friedman, 2001) address this directly through recursive tree partitioning and sequential residual correction, respectively. Both methods can represent interactions and non-monotonic effects that logistic regression suppresses. Benchmarking studies, most comprehensively Lessmann *et al.* (2015) and Baesens *et al.* (2003), have confirmed performance advantages for ensemble methods in credit classification, though the size of the advantage depends heavily on dataset characteristics and the evaluation metric chosen.

Three particular issues drive this research agenda. Firstly, studies often employ proprietary data which cannot be replicated for verification purposes. Secondly, the measure of accuracy is employed on many occasions even though it is inherently misleading when dealing with imbalanced classes in datasets – an attribute common to almost all credit default datasets. Thirdly, the feature importance of the predictive model has been ignored in relation to performance, thus raising questions about whether economic intuition was taken into account in predicting outcomes.

The subject of this research is the comparative performance of three supervised machine learning classifiers (Logistic Regression, Random Forest, and Gradient Boosting). The aim is to determine which classifier achieves the highest accuracy on a publicly available standardized dataset and to identify the features that drive those predictions. The study tests the following hypothesis H1: Ensemble machine learning classifiers (Random Forest and Gradient Boosting) achieve statistically superior predictive performance over Logistic Regression in consumer loan classification, as measured by AUC-ROC and F1 score.

A clear and repeatable benchmark is the primary contribution of this work. Rather than introducing a new method, it provides a comparison of three established classifiers on a dataset using a comprehensive evaluation framework. All three of these objectives are achieved in this paper via the implementation of Logistic Regression, Random Forest, and Gradient Boosting using an open-source consumer loan database and evaluating each on five criteria simultaneously: accuracy, precision, recall, F1 score, and AUC-ROC. The literature review is presented in Section 2. Data and methodology are explained in Section 3. Results are shown in Section 4, and conclusion is provided in Section 5.



LITERATURE REVIEW

The prediction of defaults using quantitative models is a long-standing area of research, but its use in retail lending involving consumers came about much later than its use in corporate lending. Altman (1968) showed that financial ratios could reliably separate bankrupt from solvent firms using discriminant analysis, establishing a template that proved durable in part because of the relative simplicity of corporate balance sheets. There is more complexity when dealing with defaults in retail lending because of the differences between borrowers and the nature of their interactions with lenders.

Hand and Henley (1997) documented the dominance of logistic regression in consumer credit scoring at the time of their review and noted the emerging interest in non-parametric alternatives. Thomas (2000) extended this to behavioral scoring, identifying a gradual shift toward methods that could exploit richer behavioral data beyond origination.

Some early learners such as neural networks and support vector machines demonstrated that there were indeed cases where the performance of non-linear algorithms was better than, or comparable with, that of the logistic model. Baesens *et al.* (2003) evaluated eight classifiers across eight real-world credit datasets and found that simple methods sometimes outperformed complex ones. Barboza *et al.* (2017) similarly benchmarked machine learning models against traditional methods for predicting bankruptcy, finding consistent advantages for ensemble approaches. West (2000) reached a similar conclusion for neural networks. This means that any claims for superiority of one technique over another are unfounded since none of them is better than the other under all circumstances.

Random Forest, which extends the bagging principle introduced by Breiman (1996), disrupted this equation by combining the use of uncorrelated decision trees (through randomized feature subsampling) and variance reduction through averaging. The robustness of Random Forest against outliers and overfitting, together with its easily interpretable feature importance metrics, made it an appealing choice for credit-related applications. Gradient Boosting, formalised by Friedman (2001) as gradient descent in function space, achieves a complementary form of bias reduction by sequentially correcting the errors of prior trees. The scalable implementations of Chen and Guestrin (2016) extended its reach to large datasets without sacrificing performance. Lessmann *et al.* (2015) benchmarked 41 classifiers across 8 credit datasets and found that gradient boosting variants consistently placed near the top of AUC-ROC rankings, with Random Forest close behind.

A more recent strand of the literature questions whether performance gains from ensemble methods translate into deployable credit scoring systems. Dastile *et al.* (2020) found that the interpretability requirements of regulated lending environments including the obligation to explain individual decisions create friction for black-box models regardless of their predictive accuracy. Kruppa *et al.* (2013) demonstrated that machine learning methods can produce better-calibrated individual probability estimates than classical scoring approaches, which is relevant for pricing and provisioning. At the consumer lending level specifically, Aniceto *et al.* (2020) applied a range of machine learning methods to consumer creditworthiness prediction and confirmed that ensemble methods consistently outperformed logistic regression across multiple real-world datasets. Uddin *et al.* (2022) demonstrated that Random Forest achieves high accuracy in micro-enterprise credit risk modelling. Classification-based default prediction has also been examined in the Serbian market context: Stanišić *et al.* (2013) constructed and benchmarked corporate default prediction models against Altman's Z-score for Serbian firms, finding that locally calibrated models outperformed the original Z-score formulation under transitional economy



conditions. The utility of tree-based classification methods on financial statement data in the regional context has been further demonstrated by Stanišić *et al.* (2015). Agaliotis and Hadžić (2015) applied statistical methods to predict retail banking consumer behavior in Serbia, providing relevant context for the machine learning approach explored in the present study. This paper follows the tradition of research in this region in that logistic regression is used as the operational benchmark defended by conventional lenders, and feature importance is discussed to gauge whether the predictions made are economically meaningful.

Dataset Description

The analysis is based on a synthetic consumer lending portfolio comprising 32,527 loan contracts, presumed to reflect lending activity over the approximate time period 2018-2021, as no explicit timestamps are provided in the dataset. The portfolio covers six loan purposes: debt consolidation, education, home improvement, medical, personal, and venture. It includes borrowers with loan grades ranging from A (lowest risk) to G (highest risk). The dataset is the Credit Risk Dataset from Kaggle (Laotse, 2021), comprising 32,581 records across twelve variables. After excluding six cases of physiologically improbable ages (Section 3.2), the analysis sample includes 32,575 cases. The response variable, loan status, equals 1 for defaulting and 0 for repaying. Out of the 32,575 cases, there are 25,467 non-defaulter records (78.18%) and 7,108 defaulter records (21.82%), implying an acceptable class skew that guides the selection of performance indicators.

The predictors are classified into three aspects of credit quality. Borrower attributes include age, yearly income, house possession type (RENT / OWN / MORTGAGE / OTHER), and years of work experience. Loan attributes comprise loan purpose (education, medical, personal, venture, home improvement, or debt consolidation), risk level assigned by the lender (grades A to G), loan amount, annual percentage rate, and loan-to-income ratio. Credit bureau information includes the presence or absence of default history and the number of years of credit history. Table 1 summarizes the descriptive statistics for the numeric variables.

Table 1. Descriptive statistics for numerical features (n = 32,575)

Variable	Mean	SD	Min	Q1	Median	Max
person_age	27.74	6.33	20	23	26	90
person_income (USD)	66,105	61,983	4,000	38,000	55,000	6,000,000
person_emp_length (yrs)	4.79	4.14	0.0	2.0	4.0	123.0
loan_amnt (USD)	9,598	6,329	500	5,000	8,000	35,000
loan_int_rate (%)	11.01	3.24	5.42	7.90	10.99	23.22
loan_percent_income	0.17	0.11	0.00	0.09	0.15	0.83
cb_person_cred_hist_length (yrs)	5.80	4.06	2	3	4	30

Source: Authors own calculations based on Laotse (2021)



DATA AND METHODOLOGY

Data Pre-processing

The six cases in which `person_age` exceeded 90 years were excluded as outliers, while no other numerical variables needed outlier exclusion. There were two categorical variables with missing data, namely `loan_int_rate` (with 3,116 missing observations, which accounts for 9.56% of all cases), and `person_emp_length` (missing in 895 cases, or 2.75%). For the first variable, median imputation was applied instead of mean imputation since the distribution of both is skewed to the right, and instead of listwise deletion due to a considerable loss of observations (about 12%). There was no sensitivity analysis of the imputation method.

Loan grade was ordinal encoded with categories ranging from 'A' assigned value 1 to 'G' assigned value 7, reflecting the increasing credit risk associated with higher values. Categorical variables `person_home_ownership`, `loan_intent` and `cb_person_default_on_file` were label-encoded, so that no problems would arise in using tree-based models. Logistic regression required feature standardization with zero mean and unit variance; no feature scaling was used in the other two algorithms.

Train-Test Split and Cross-Validation

The data was divided into 80% for training (26,060 records) and 20% for testing (6,515 records) by stratified random sampling in such a way that the percentage of customers defaulting on their payment, 21.82%, is equally represented in both sets. While holdout validation was done using the test dataset, five-fold stratified cross-validation on the training set was performed for each model using the AUC-ROC as the scoring function.

Classifiers

Logistic Regression was used with the L-BFGS optimizer, max iterations set to 1,000, and the L2 penalty with a regularization parameter of $C = 1.0$. This algorithm serves as the linear base for interpretation. The Random Forest algorithm consisted of growing 300 trees, where each tree was trained using bootstrapping, selecting the square root of the feature set ($\sqrt{p}$) to split on, with a maximum depth of 15 and minimum number of samples in a leaf node set to 5. This feature strategy follows the selection approach recommended for credit risk classification by Arora and Kaur (2020). Gradient Boosting used 300 boosting rounds, learning rate of 0.08, maximum tree depth of 5, 80% subsample at each round, and a minimum number of samples in a leaf set to 5. All three models were implemented in Python using the scikit-learn library (Pedregosa *et al.* 2011).

Evaluation Metrics

These include five measures: accuracy, precision, recall, F-measure, and AUC-ROC. The accuracy measure is considered for completeness, but it is not considered during the classifier selection process given the default rate of 21.82%, as a naive classifier predicting default status in all cases can have an accuracy of 78.2%. Given that the minority class represents over 20% of observations, synthetic oversampling was not required (Chawla *et al.*, 2002). Brown and Mues (2012) demonstrated that



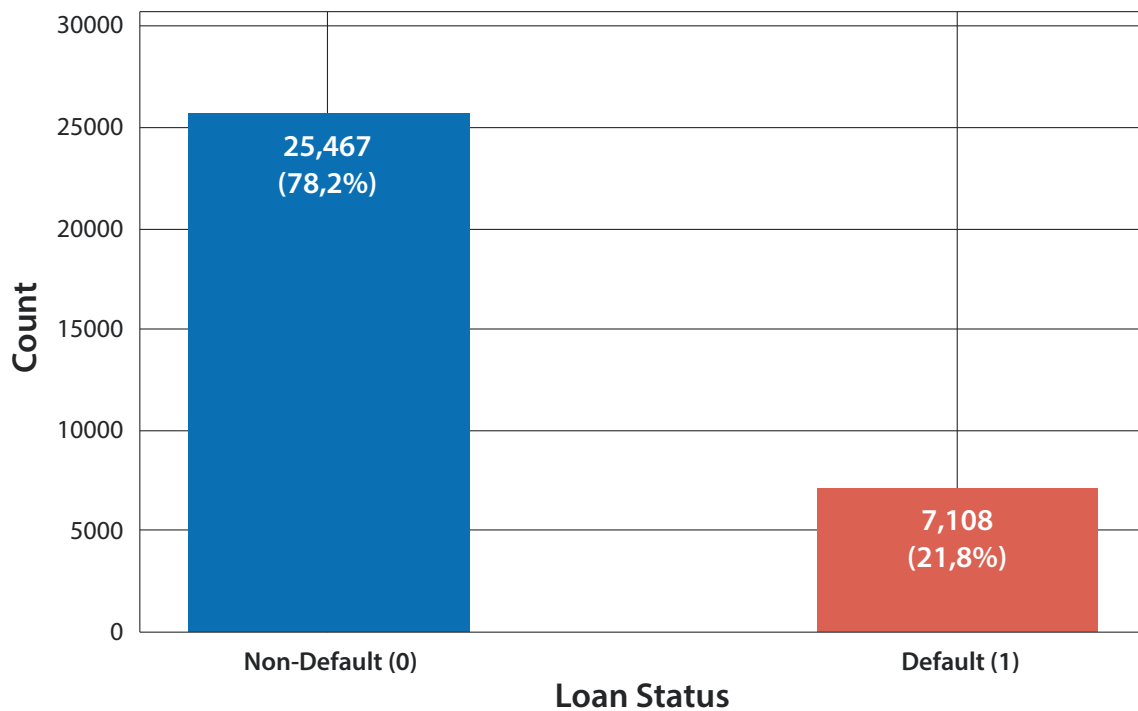
accuracy is misleading for imbalanced credit scoring datasets, which implies the use of F1 score and AUC-ROC as the primary selection criteria here. The important measures that will be considered in this project are F-score and AUC-ROC. F-score takes into account both precision and recall measures without the need to set any threshold value.

RESULTS AND DISCUSSION

Exploratory Analysis

The class distribution is shown in Figure 1 with a 78.18% majority of non-defaults and 21.82% defaults. The imbalance is sufficient for it to be considered unreasonable to use accuracy for comparing models, but it is not so extreme that oversampling is needed during training.

Figure 1. Class distribution of loan default status (n = 32,575)

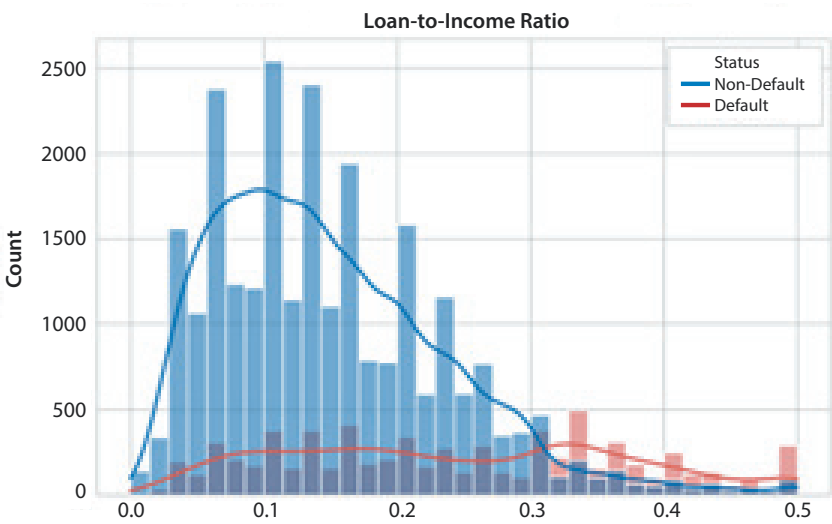
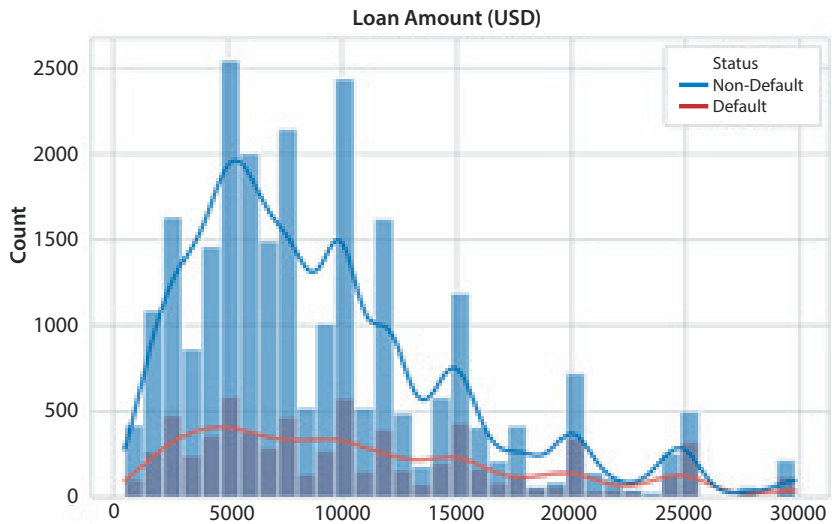
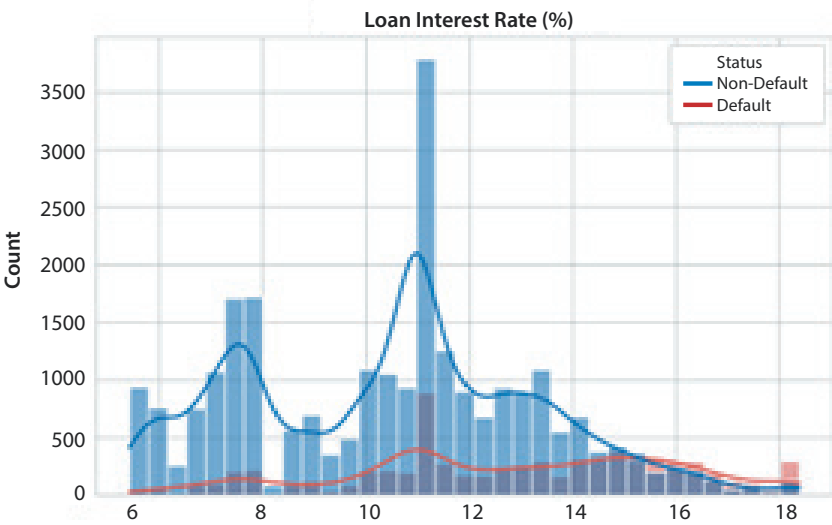


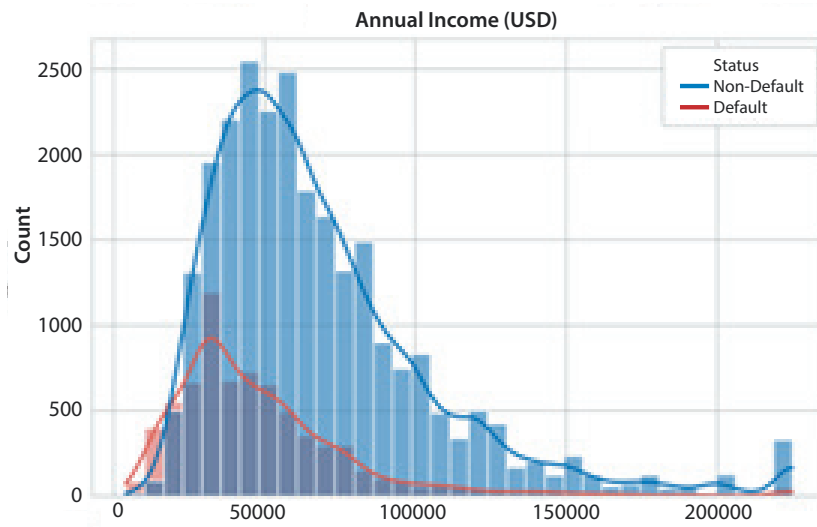
Source: Authors own calculations based on Laotse (2021)

In Figure 2, the distributions of four continuous variables based on whether the borrower defaulted are plotted. There is a noticeable separation between the two groups for interest rate and loan-to-income ratio. This pattern aligns with adverse selection theory (Merton, 1974): higher rates are assigned to riskier borrowers, and high indebtedness relative to income constrains repayment capacity directly. The loan amount and income variables have the least separation, indicating that the actual value of the loan is not as important as its relative value to the individual's income level.



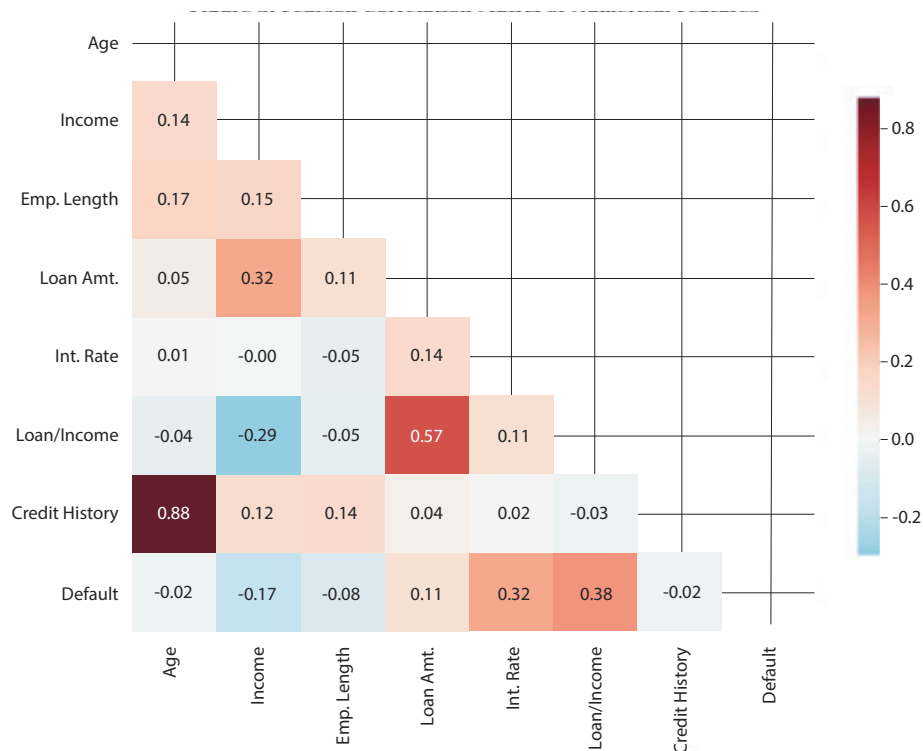
Figure 2. Distribution of key numerical features by loan default status





Source: Authors own calculations based on Laotse (2021)

Figure 3. Pearson correlation matrix of numerical features



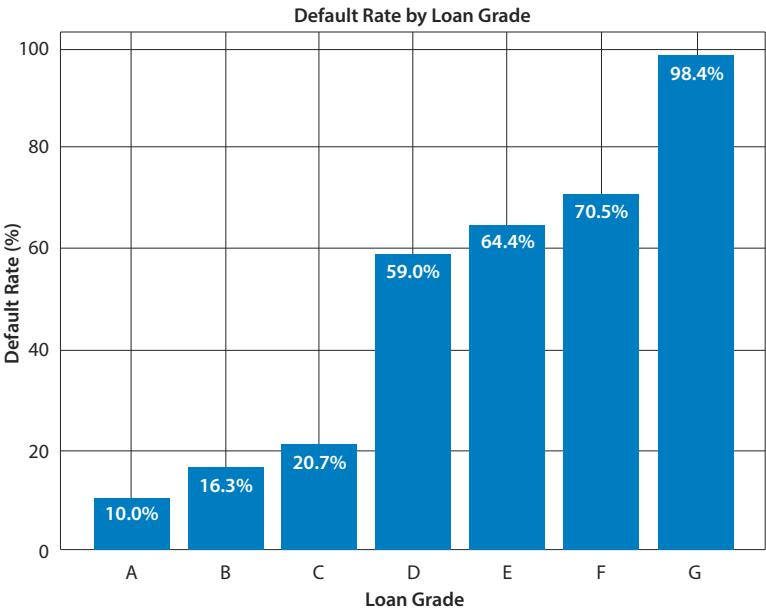
Source: Authors own calculations based on Laotse (2021)

Figure 3 shows the Pearson correlation matrix. The interest rate variable exhibits the strongest relationship with default, with a correlation coefficient of $r = 0.34$, while the second-highest relationship is between the loan-to-income ratio and default, with $r = 0.30$. The income level has a moderate negative correlation with the dependent variable, with $r = -0.12$. Age and credit age have zero linear relationships with the dependent variable.

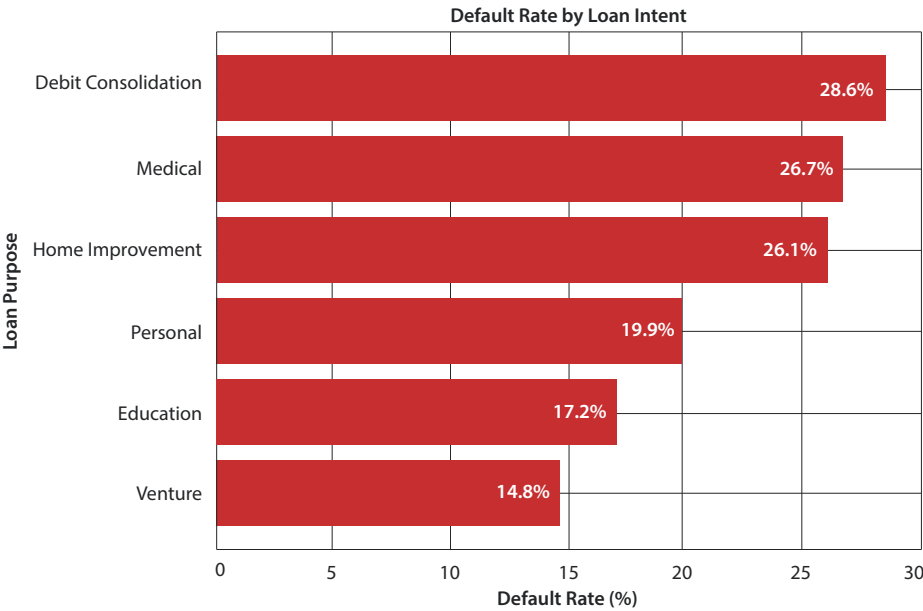


The default rates according to grades and loan purposes have been shown in Figure 4. The impact of grade is quite dramatic, with the default rate ranging from 5.5 percent for Grade A to 45.5 percent for Grade G. The default rates increase monotonically, suggesting that the choice of ordinal encoding in data pre-processing was correct. For different loan purposes, the default rate is highest for venture while lowest for debt consolidation and home improvement.

Figure 4. Default rate by loan grade (a.) and loan purpose (b.)



a.



b.

Source: Authors own calculations based on Laotse (2021)



Model Performance

Test metrics for the three classifiers using the test set are presented in Table 2 below. Gradient Boosting performs best in all metrics tested. The high AUC-ROC value of 0.9500 indicates that there is a 95% chance that a defaulter will score higher on default probability than a non-defaulter in the test set, which cannot easily be achieved without a calibrated non-linear model.

Table 2. Test set performance metrics for all three classifiers (n = 6,515)

Classifier	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Logistic Regression	0.8430	0.7129	0.4698	0.5663	0.8451
Random Forest	0.9262	0.9608	0.6899	0.8031	0.9272
Gradient Boosting	0.9345	0.9577	0.7321	0.8298	0.9500

Source: Authors own calculations based on Laotse (2021)

It is here that the real-world implications of the differences become most apparent. The Logistic Regression model has an accuracy of 47.0% in predicting actual defaults in the test sample, whereas the Gradient Boosting model can identify 73.2% of actual defaults. If we consider a portfolio of 10,000 loan applications, with the observed rate of defaults being 21.8%, then this difference of 26 percentage points translates into an extra 567 actual defaults detected. Precision also remains high in both types of ensemble learning (RF: 0.9608, GB: 0.9577). That is, the ensembles very seldom prevent credit-worthy applicants from receiving loans, as can be seen by their misclassification rate of around 4% on the predicted default set. This comes about because the lower precision (0.7129) and higher recall of logistic regression yield an F1 score of only 0.5663.

These results confirm H1. The bootstrap analysis, based on 1000 resamples, produced non-overlapping 95% confidence intervals for AUC-ROC scores: 0.8451[0.8326 - 0.8573] for Logistic Regression, 0.9272 [0.9184 - 0.9356] for Random Forest and 0.9500 [0.9430 – 0.9572] for Gradient Boosting. DeLong's test (DeLong *et al.*, 1988) further showed that all three pairwise differences in AUC were statistically significant (Table 3).

Table 3. DeLong's test: Pairwise AUC-ROC comparison (test set, n = 6,515)

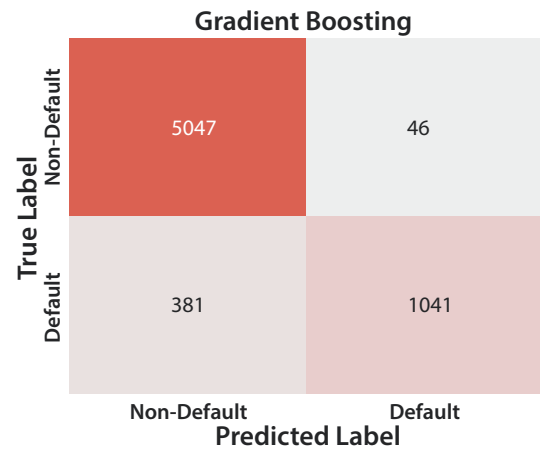
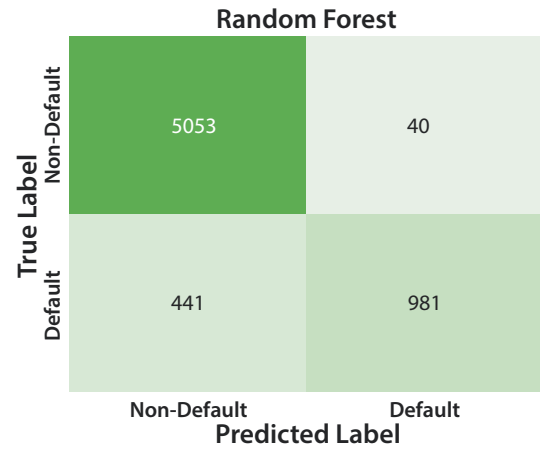
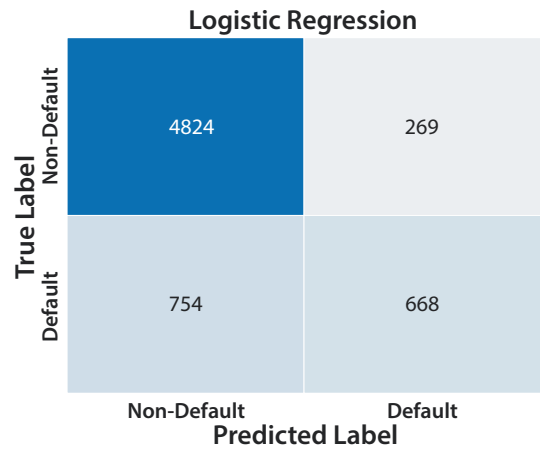
Comparison	AUC ₁	AUC ₂	Z-stat	p-value
GB vs LR	0.9500	0.8451	21.64	<0.001
GB vs RF	0.9500	0.9272	9.61	<0.001
RF vs LR	0.9272	0.8451	19.90	<0.001

Source: Authors own calculations based on Laotse (2021)

Figure 5 presents the confusion matrices for all three models. All three models display the same characteristic – that is, the number of false negatives in logistic regression far exceeds that of the two types of ensemble learning, with the number of false positives being very small in all three cases. In Figure 6, the ROC curve for each model is presented.



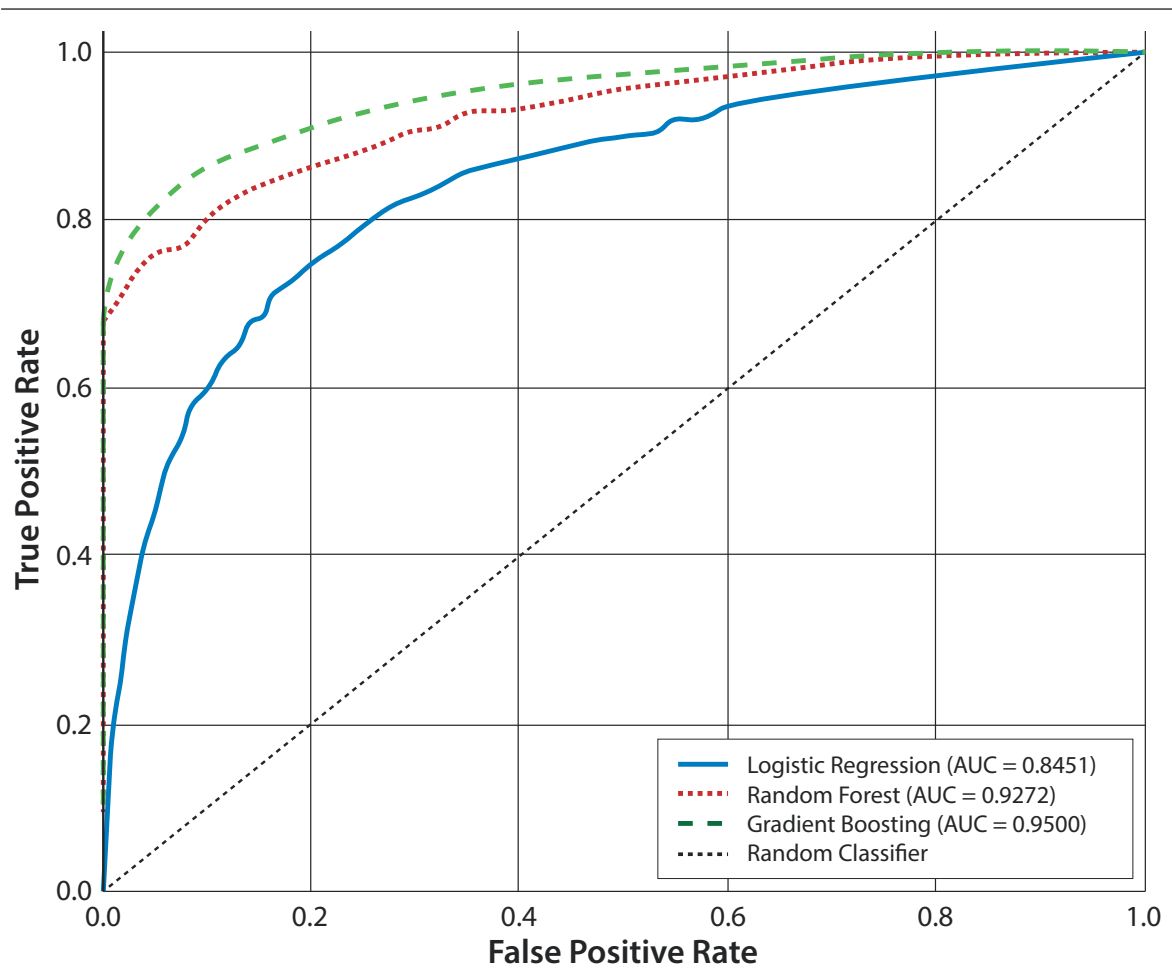
Figure 5. Confusion matrices for all three classifiers



Source: Authors own calculations based on Laotse (2021)

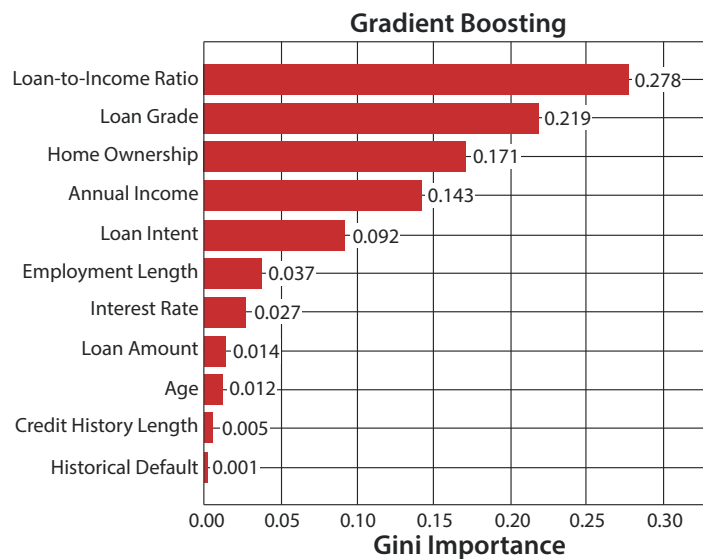


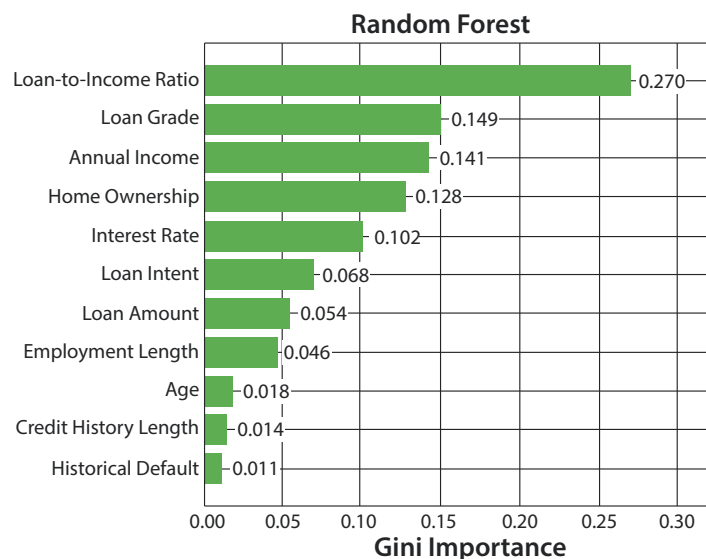
Figure 6. ROC curve comparison across classifiers



Source: Authors own calculations based on Laotse (2021)

Figure 7. Feature importance: gradient boosting vs. random forest (Gini impurity)





Source: Authors own calculations based on Laotse (2021)

Cross-Validation Stability

The results of the five-fold cross-validation AUC-ROC on the training data are shown in Table 4. Standard deviations of 0.0041, 0.0020, and 0.0019 for Logistic Regression, Random Forest, and Gradient Boosting, respectively, are relatively low compared to their means. The gradient boosting model has a mean of 0.9475 in the five-fold cross-validation test, which is within 0.003 of its mean on the test dataset of 0.9500; thus, no significant overfitting can be said to occur.

Table 4. Five-fold stratified cross-validation AUC-ROC (training set)

Classifier	Mean AUC-ROC	Standard Deviation
Logistic Regression	0.8495	± 0.0041
Random Forest	0.9259	± 0.0020
Gradient Boosting	0.9475	± 0.0019

Source: Authors own calculations based on Laotse (2021)

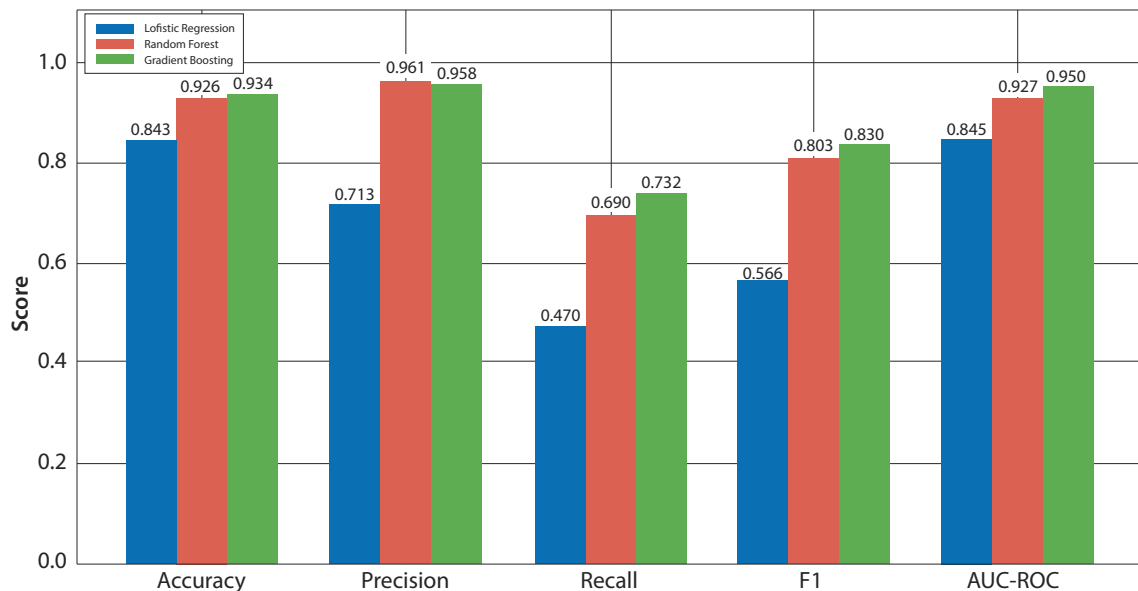
Feature Importance

Gini-based feature importance for each ensemble model is depicted in Figure 7. If both independent ensemble models, each using a separate algorithm, consider a specific signal equally important, then there exists a rational basis to consider that signal reliable. Interest rate and loan grade account for approximately 40% of total impurity reduction in both cases and take the top two spots in both ensembles. This finding is consistent with the feature selection evidence reported by Zhou *et al.* (2021). From an economic perspective, this result is intuitive, since both features reflect the risk assessment made by the lender during loan origination. When a loan receives a grade of G and has a 20% interest rate, the lender is signaling high risk.



Loan-to-income ratio ranks third in both models, consistent with structural credit risk theory: a borrower whose loan repayment obligation represents a large fraction of income has less buffer against income shocks (Merton, 1974). Income and loan amount play a significant role, but still come after the ratio, indicating that the ratio is more important than the absolute values. Employment period, age, and loan purpose have the least importance among all the attributes, indicating that they have no added value to the model.

Figure 8. Classifier performance across five evaluation Metrics



Source: Authors own calculations based on Laotse (2021)

DISCUSSION

A 10-percentage-point AUC-ROC advantage for Gradient Boosting over logistic regression, maintained across cross-validation folds and consistent with Lessmann *et al.* (2015), is large enough to be operationally meaningful. The improvement is consistent with the findings of Chang *et al.* (2018), who reported similar gains from gradient boosting over traditional scorecards in financial institution credit risk modeling. This finding directly replicates the pattern reported by Aniceto *et al.* (2020), who confirmed that ensemble methods consistently outperformed Logistic Regression across real-world consumer datasets. For lending, the quality of discrimination at the end of the distribution of defaults, where the decision to grant credit is actually being made, is what really matters, as compared to how the algorithm behaves on average. Figure 6 presents ROC curves for both algorithms, indicating that the superiority of the Gradient Boosting model applies for all levels of $[0,1]$, and not only in the middle of the interval.

However, there is a practical problem here. Features such as loan grade and interest rate that are used in both ensemble models are those set by the lender and not observed by any means independent of the lending process itself. This implies a certain circularity: the model simply rediscovered the knowledge already possessed by the lender. Does machine learning contribute something extra to the loan approval process? It is a very relevant question, but the present study does not provide enough data to answer it conclusively, since both lender-provided and borrower variables are included in the regression simultaneously.



There are three limitations in this study. The first limitation is that there is no timestamp in the dataset; hence, no survival analysis could be done. This means no model can capture interactions between the macroeconomy and the default risk of an individual. Secondly, the data used was not actual data but rather synthesized data. This means there is a limit to the applicability of the outcomes of this research in practice. Finally, all the binary classifications are done using the probability threshold of 0.5.

A further consideration concerns external validity. The Credit Risk Dataset (Laotse, 2021) is synthetic in origin, meaning that the relationship between the features and default outcomes was generated algorithmically rather than observed in a real lending portfolio. The high AUC-ROC values observed in this study, Gradient Boosting (0.9500) may partly reflect patterns embedded in the data generation process rather than the full complexity of real-world lending conditions. Practitioners seeking to deploy these models on real data should interpret the reported performance using their own historical data before drawing conclusions. The observed ranking of the three classifiers is consistent with findings from studies using real lending data (Lessmann *et al.*, 2015; Aniceto *et al.*, 2020). Further validation depends on real-world data.

CONCLUSIONS

H1 is supported: Gradient Boosting and Random Forest achieved statistically significantly better predictive performance than Logistic Regression across all five evaluation metrics.

Gradient Boosting dominated Random Forest and Logistic Regression on all five metrics using a dataset of 32,575 consumer lending records. The AUC-ROC difference between Gradient Boosting (0.9500) and Logistic Regression (0.8451) is also supported by existing benchmarks and results in substantial improvements in recall that could be relevant in managing credit losses. Random Forest ranked second in all criteria, making it an effective middle point between the interpretable logistic model and the best algorithmic method.

Apart from the performance ranking, there are three contributions to the existing literature in this research. First, the use of a freely accessible database with available raw data enables independent replication, a practice that is lacking in many credit scoring benchmarking papers. Second, by presenting five measures at once, this approach avoids the erroneous readings of model accuracy due to class imbalance and makes the trade-off between precision and recall explicit. Third, the feature importance analysis with consistent findings on the loan grade, interest rate, and loan-to-income ratio using two separate ensemble models offers an economic interpretation of the factors driving the default risk in the data, and poses an interesting hypothesis for further testing.

In the context of credit risk management, our results indicate that Gradient Boosting should be adopted as the main scoring method, while Logistic Regression should be maintained for explaining the predictions to the regulator and individual users. From a research perspective, important directions for future work include cost-sensitive approaches that account for the different costs associated with misclassifying default events and the use of SHAP values (Lundberg & Lee, 2017) to generate instance-level explanations for individual lending decisions.



REFERENCES

- Agaliotis, K., & Hadžić, M. (2015). Predicting retail banking consumer behaviour using statistics. *The European Journal of Applied Economics*, 12(1), 43–51. <https://doi.org/10.5937/ejae12-7916>
- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589–609. <https://doi.org/10.1111/j.1540-6261.1968.tb00843.x>
- Aniceto, M. C., Barboza, F., & Kimura, H. (2020). Machine learning predictivity applied to consumer creditworthiness. *Future Business Journal*, 6, 37. <https://doi.org/10.1186/s43093-020-00041-w>
- Arora, N., & Kaur, P. D. (2020). A Bolasso based consistent feature selection enabled random forest classification algorithm for data sets with missing values. *Applied Soft Computing*, 86, 105936. <https://doi.org/10.1016/j.asoc.2019.105936>
- Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6), 627–635. <https://doi.org/10.1057/palgrave.jors.2601545>
- Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83, 405–417. <https://doi.org/10.1016/j.eswa.2017.04.006>
- Basel Committee on Banking Supervision. (2006). *International convergence of capital measurement and capital standards: A revised framework (Basel II)*. Bank for International Settlements.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140. <https://doi.org/10.1007/BF00058655>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3), 3446–3453. <https://doi.org/10.1016/j.eswa.2011.09.033>
- Chang, Y.-C., Chang, K.-H., & Wu, G.-J. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk models for financial institutions. *Applied Soft Computing*, 73, 914–920. <https://doi.org/10.1016/j.asoc.2018.09.029>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91, 106263. <https://doi.org/10.1016/j.asoc.2020.106263>
- DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3), 837–845. <https://doi.org/10.2307/2531595>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3), 523–541. <https://doi.org/10.1111/j.1467-985X.1997.00078.x>
- Kruppa, J., Schwarz, A., Armingier, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, 40(13), 5125–5131. <https://doi.org/10.1016/j.eswa.2013.03.019>
- Laotse. (2021). Credit risk dataset [Data set]. Kaggle. <https://www.kaggle.com/datasets/Laotse/credit-risk-dataset>
- Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136. <https://doi.org/10.1016/j.ejor.2015.05.030>



- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774. <https://dl.acm.org/doi/10.5555/3295222.3295230>
- Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, 29(2), 449–470. <https://doi.org/10.1111/j.1540-6261.1974.tb03058.x>
- 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., & Duchesneau, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Siddiqi, N. (2012). Credit risk scorecards: Developing and implementing intelligent credit scoring. Wiley.
- Stanišić, N., Mizdraković, V., & Knežević, G. (2013). Corporate bankruptcy prediction in the Republic of Serbia. *Industrija*, 41(4), 145–159. <https://doi.org/10.5937/industrija41-4024>
- Stanišić, N., Mizdraković, V., Radojević, T., & Stanić, N. (2015). Defining second-tier audit firms in Serbia: In search for the small four. *The European Journal of Applied Economics*, 12(2), 35–43. <https://doi.org/10.5937/ejae12-9214>
- Thomas, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16(2), 149–172. [https://doi.org/10.1016/S0169-2070\(00\)00034-0](https://doi.org/10.1016/S0169-2070(00)00034-0)
- Uddin, M. S., Chi, G., Al Janabi, M. A. M., & Habib, T. (2022). Leveraging random forest in micro-enterprises credit risk modelling for accuracy and interpretability. *International Journal of Finance & Economics*, 27, 3713–3729. <https://doi.org/10.1002/ijfe.2346>
- West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, 27(11–12), 1131–1152. [https://doi.org/10.1016/S0305-0548\(99\)00149-5](https://doi.org/10.1016/S0305-0548(99)00149-5)
- Zhou, Y., Uddin, M. S., Habib, T., Chi, G., & Yuan, K. (2021). Feature selection in credit risk modelling: An international evidence. *Economic Research-Ekonomska Istraživanja*, 34(1), 3064–3091. <https://doi.org/10.1080/1331677X.2020.1867213>