Journal of
Risk and Financial Management

MDPI

Article

Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach

Nicolas Suhadolnik [1,2], Jo Ueyama [1] and Sergio Da Silva [3,*]

[1] Institute of Mathematics and Computer Science, University of Sao Paulo, Sao Carlos 13566-590, Brazil; npsuhadolnik@gmail.com (N.S.); joueyama@icmc.usp.br (J.U.)
[2] Regional Bank for Development of the South Region, Curitiba 80030-900, Brazil
[3] Graduate Program in Economics, Federal University of Santa Catarina, Florianopolis 88049-970, Brazil
* Correspondence: professorsergiodasilva@gmail.com

Abstract: Financial institutions and regulators increasingly rely on large-scale data analysis, particularly machine learning, for credit decisions. This paper assesses ten machine learning algorithms using a dataset of over 2.5 million observations from a financial institution. We also summarize key statistical and machine learning models in credit scoring and review current research findings. Our results indicate that ensemble models, particularly XGBoost, outperform traditional algorithms such as logistic regression in credit classification. Researchers and experts in the subject of credit risk can use this work as a practical reference as it covers crucial phases of data processing, exploratory data analysis, modeling, and evaluation metrics.

Keywords: credit risk; computer methods; machine learning

Citation: Suhadolnik, Nicolas, Jo Ueyama, and Sergio Da Silva. 2023. Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach. *Journal of Risk and Financial Management* 16: 496. https://doi.org/10.3390/jrfm16120496

Academic Editors: Daniel Oliveira Cajueiro and Regis A. Ely

Received: 23 October 2023
Revised: 21 November 2023
Accepted: 25 November 2023
Published: 27 November 2023

1. Introduction

Recent technological advancements and regulatory changes have led to new financial intermediation models, increasing competition, and reducing loan costs. Central banks are promoting programs for financial system modernization and efficiency, including the deployment of central bank digital currencies and open banking for data and service exchange (Araujo 2022). This environment fosters the use of credit financial technologies, balancing transformation and value creation with stability and transparency.

The surge in computer methods and data volumes has transformed credit judgments, placing machine learning techniques at the forefront of credit risk assessments (Louzada et al. 2016). The application of appropriate credit risk analysis tools is essential for the functioning of financial institutions. However, the opacity and "black box" nature of machine learning algorithms have raised concerns regarding their implications for financial stability (Chakraborty and Joseph 2017).

This study delves into the significant role of machine learning in credit risk assessment. It provides a comprehensive overview of current methodologies and introduces a novel application of these methods using extensive real data from a financial institution. Our analysis, covering a longer timeframe and a larger dataset than recent credit scoring research, reveals that ensemble techniques, particularly XGBoost, consistently outperform other methods in both imbalanced and balanced datasets. We emphasize the importance of exploratory data analysis in understanding complex, imbalanced data.

Our contributions to the field are twofold. First, we demonstrate the effectiveness of ensemble methods in credit risk classification, supported by comprehensive data analysis. Second, we contextualize our findings within the broader credit scoring literature, comparing our results with significant recent studies (Teply and Polena 2020; Xia et al. 2020; Malekipirbazari and Aksakalli 2015). This paper follows a structured approach: Section 2 discusses the key aspects of credit risk analysis and differentiates between traditional and

machine learning methodologies. Section 3 outlines the methods used, while Section 4 details data collection and pre-processing. Section 5 presents the experiments, major findings, and a literature comparison. Finally, Section 6 concludes with our final thoughts.

2. Literature Review

This review has three sections. First, it covers basic ideas and aspects of credit risk analysis. Then, it contrasts machine learning with conventional statistical methods. Lastly, it provides a summary of current machine learning uses in evaluating credit risk.

2.1. Credit Risk Analysis Dimensions

Credit decisions are typically made using judgments based on the prior knowledge of human analysts. This method is very susceptible to subjectivity, consistency issues, and the influence of certain analyst preferences (Abdou and Pointon 2011). Complex statistical and computational approaches started to take up more and more space as the credit market expanded and financial institutions' competitiveness increased due to new technologies being applied in the financial services industry. As a result, these techniques started to supplement or, in some situations, replace human judgment.

Financial institutions making credit decisions usually assess applicant risk using the "5 Cs of credit" method. This involves five key areas: "Character" examines past defaults, legal issues, and trustworthiness indicators; "Capacity" evaluates financial health, focusing on debt-to-income ratios and management skills; "Conditions" assesses external macroeconomic factors outside the borrower's control; "Capital" looks at equity to gauge commitment and reduce lender risk; and "Collateral" considers guarantees, requiring detailed valuations due to diverse asset properties.

Credit risk assessment involves creating a synthetic indicator from information available at the time of the credit request. This process evaluates key factors affecting borrower behavior to decide on loan approval (Hand and Henley 1997). The data for credit risk comes from various sources, and there is no fixed number of attributes for scoring models; this varies based on data type and specific economic and social contexts (Abdou and Pointon 2011). Despite abundant data, limited access to financial information has led to initiatives like open banking, which seeks to standardize data sharing and give clients control over their data, thereby reducing information asymmetry (Vicente 2020).

2.2. Traditional Algorithms vs. Machine Learning

Traditional models like linear and logistic regression work well for economic issues, but they may fall short with large datasets where relationships are more complex than linear ones. Overreliance on these old methods can lead to irrelevant hypotheses, questionable results, and poor handling of current problems. Therefore, using a broader range of tools is recommended for data-driven problem solving (Breiman 2001). In this context, machine learning algorithms could be more effective for modeling complex interactions (Varian 2014).

Econometric and statistical methods differ from machine learning in their primary goals. Econometrics, assuming data come from a known stochastic model, focuses on the significance of estimated parameters, confidence intervals, and causal inference (Athey and Imbens 2019). In contrast, machine learning prioritizes developing algorithms for making predictions or identifying key units with minimal information. Consequently, in machine learning, the key measure of success is often out-of-sample predictive performance (Bazarbash 2019).

Many machine learning models produce parameters that are hard to interpret, making it challenging to understand their results. This lack of clarity affects financial institutions' strategies, often resulting in fully automated credit assessment processes (Bazarbash 2019). Supervised learning models are methodologically effective in classifying individuals as reliable or unreliable payers. Recent studies show a broad range of machine learning

applications in credit risk assessment, and this paper presents an overview of the main methodologies and their key findings.

2.3. Credit Scoring: An Overview

Machine learning has become a key tool in credit decision making in recent years. Numerous studies evaluate its predictive power or develop new classification and regression methods for specific scenarios. Its influence extends to various finance areas, including cross-sectional return prediction for stocks and other assets, as shown by Gu et al. (2020) and Bali et al. (2023); market timing (Cakici et al. 2023; Zhou et al. 2023); and risk prediction (Drobetz et al. 2021). These applications highlight machine learning's transformative role in finance, offering insights beyond traditional models.

Louzada et al. (2016) conducted a thorough evaluation of the literature on the theory and use of credit risk rating models. They observe that the most prevalent purpose among authors was to propose a new credit rating algorithm, based on a thorough review of 187 articles published between 1992 and 2015. Another goal was to compare various credit risk assessment methodologies, which has become less important in recent publications. The most common credit risk classification methods discovered in these studies are neural networks, support vector machines, fuzzy logic, linear regression, decision trees, logistic regression, and ensemble methods. Louzada et al. (2016) also compare the prediction performance of the various approaches on three different sets of data, with a focus on support vector machines and fuzzy logic.

In their study, Dastile et al. (2020) conducted a comprehensive literature review on statistical and machine learning models used in credit scoring. They also introduce a guiding machine learning framework for credit scoring. The review covers 74 primary studies published between 2010 and 2018, revealing that ensemble classifiers generally outperform individual ones. Notably, while deep learning models are not widely adopted in the credit scoring literature, they show promise in their results.

In a systematic review, Markov et al. (2022) analyzed 150 articles from 2016 to 2021 to discern trends in credit scoring methodologies. The article contributes to the understanding of how various statistical and machine learning techniques are applied in credit scoring at different stages. Their study reveals a growing preference for advanced methods like ensembles and neural networks over traditional techniques like decision trees and logistic regression, often leading to better predictive results.

Zhang and Yu (2024) offer an in-depth review of consumer credit risk assessment. They pinpoint a notable gap in research about data traits and stress the importance of multi-scenario modeling in machine learning. The study underscores the significance of grasping data traits' impacts on a model's predictive capacity. It also emphasizes the necessity of developing adept data processing techniques to decide the best learning method. The paper concludes by presenting a structured framework for consumer credit scoring, noting the consistent development of hybrid and ensemble classifiers as primary algorithms.

The selection of a data set is an important part of empirical investigation. Given the difficulty in obtaining information on organizations and people's credit histories, many researchers rely on data that are readily available in public archives (Louzada et al. 2016). We might specifically mention the usage of the Australian credit (AC) and German credit (GC) datasets, both of which are available in the UCI Machine Learning Repository (Dua and Graff 2017). The AC dataset has 690 observations with 14 features, whereas the GC dataset contains 1000 observations with 20 features. Despite the fact that these datasets are frequently used in credit risk assessment applications, the small number of observations might be a significant drawback in research comparing the predictive power of different classifiers.

With the rise of FinTechs, such as peer-to-peer (P2P) lending platforms, other data sources, such as "digital footprints," have been incorporated into credit risk assessment (Berg et al. 2020). Due to their impact on the inclusion and stability of the financial system, these institutions have garnered the interest of researchers and regulators (Bazarbash 2019;

Chakraborty and Joseph 2017). Some companies, such as Lending Club, one of the pioneering platforms in P2P lending, make transaction data available to the public; however, the models used are largely undisclosed.

3. Materials and Methods

Machine learning algorithms are increasingly being used in credit decision making. There is ongoing interest in the development of new tools for classifying credit risk. As a result, current research contains a wide range of algorithms and applications (Louzada et al. 2016). With this in mind, our methodology is based on a review of studies with similar goals and data sets to ours. Following that, we seek to identify the algorithms that perform the best in each of the many methods. We gathered traditionally used predictive performance indicators in order to compare our findings to those of similar applications. Table 1 summarizes the main credit scoring methodologies using data from Lending Club.

Table 1. Compilation of related works.

Paper	Sample	n	Features	Algorithm	Performance Metric
Serrano-Cinca et al. 2015	2008–2011	3788	5	Logistic regression	Accuracy Hosmer–Lemeshow test Nagelkerke's R^2
Malekipirbazari and Aksakalli 2015	2012–2014	68,000	15	Logistic regression *K*-nearest neighbors Random forests Support vector machines	Accuracy Area under the curve Root mean square error
Xia et al. 2020	2011–2013	64,139	15	Logistic regression Decision tree Random forests Artificial neural networks Gradient boosting decision tree XGBoost CatBoost	Accuracy False positive rate False negative rate Area under the curve Hirsch index
This work	2007–2018	1,305,402	18	Logistic regression Decision tree *K*-nearest neighbors Support vector machines Artificial neural networks Random forests Extra trees AdaBoost Gradient boosting decision tree XGBoost	Accuracy Precision Recall F1-score Area under the curve

3.1. Credit Risk Rating: Selected Algorithms

The major aspects of the supervised learning algorithms we chose for credit risk classification are summarized here. Although the focus of this study does not provide a comprehensive description of the methodology used, several of the listed sources do.

Logistic regression is a technique that financial institutions still employ to make credit judgments. Table 1 shows that logistic regression appeared in all of the studies examined. Despite being a relatively simple technique, it is adequate for dealing with binary classification problems, providing greater prediction accuracy in some circumstances than more complex techniques (Teply and Polena 2020). The logistic regression model considers a set of independent variables $X = \{X_1, \ldots, X_n\}$ and a categorical dependent variable $Y \in \{y_1, y_2\}$. It is specifically designed for binary classification tasks, where the dependent variable represents two possible outcomes or classes. The model applies the logistic function to transform the linear combination of independent variables into probabilities. Thus, if we consider y_1 as the category of interest (for instance, charged-off

loans), the model can be expressed as $\log \frac{\pi}{1-\pi} = X\beta$, where $\pi = P(Y = y_1)$ and β is the vector containing the model's coefficients. Therefore, the model can be represented by

$$\pi_i = \frac{e^{X\beta}}{1 + e^{X\beta}},$$

where π_i is the probability of the ith individual to belong to category y_1. The logistic regression coefficients can be estimated using the maximum likelihood method.

Decision trees are generally simple to implement and produce findings with a higher degree of interpretability, which is an important quality for credit risk classification models. They are made up of decision rules that are structured in the form of a tree architecture. In general, the goal is to create a series of if-then-else conditions that cover all possible combinations in a hierarchical structure similar to a flowchart, where subsequent decisions are dependent on previous ones and the final result is obtained from the sequence of all decisions from the root node to the terminal or leaf node. At each node, a partitioning decision is taken in order to optimize the purity measure. In general, this metric is computed using the Gini index or entropy. As a result, partitions are created in each node to ensure that a group of individuals or businesses with comparable charged-off loans remain in the same region. To avoid overfitting, a size constraint for the tree must be set. When compared to other algorithms, one advantage of decision trees is the higher interpretability of the results obtained (Bazarbash 2019).

Random forests can be thought of as a strategy that combines several decisions trees that differ in two ways. To begin, each tree is built from a subsample (called bagging) of the main sample. Second, the partitions at each node are optimized over a random subset of the characteristics rather than all available features. These two changes provide enough variance in the generated trees and smooth the outcomes, which are derived by averaging the results of each tree. A majority rule can be used to determine the ultimate result in classification problems. As a result, random forest algorithms outperform isolated tree algorithms in terms of predictive capacity (Athey and Imbens 2019).

K-nearest neighbors (KNN), regarded as one of the simplest and most popular classification algorithms, can be defined as follows:

$$h(x) = \text{majority}_{i \in \aleph_x} y_i$$

where $\aleph_x$ is the set of the k observations closest to x. Thus, it seeks the most frequent class, y_i, observed among the k-nearest neighbors of x, considering a specific distance measure, such as the Euclidean distance, applied to the set of relevant attributes.

The purpose of a support vector machine (SVM) algorithm in a binary classification problem is to identify the best classification function to separate the members of each class. Geometrically, the measure for determining the best classification function can be obtained. A linear classification function corresponds to a separating hyperplane for a linearly separable dataset. Given the large number of alternative linear hyperplanes, the SVM finds the best separation function by maximizing the margin between the two classes. The margin is intuitively described as the space or separation between the two classes as specified by the hyperplane. As shown in Figure 1, the margin corresponds to the shortest distance between the hyperplane and the nearest point of each class, also known as support vectors (Wu et al. 2008). When the functional form that determines the line of separation is linear, the SVM is said to be based on a linear kernel. Through the use of nonlinear kernels, notably the radial and polynomial basis functions, several SVM extensions enable us to cope with datasets that are not linearly separable. Among the key benefits of SVM is its capacity to handle high-dimensional datasets, as well as being less prone to overfitting (Bazarbash 2019).

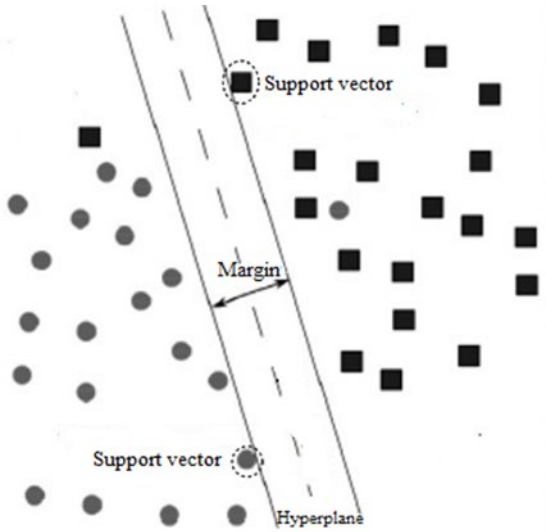

Figure 1. Hypothetical example of a separation hyperplane defined by a SVM algorithm.

An artificial neural network is made up of numerous processing nodes or neurons and is inspired by the brain structures of sentient species that learn through experience. As shown in Figure 2, the architecture of neural networks is arranged into layers that are linked by a hierarchy of relative weights. The characteristics are utilized in the first layer, or input layer, to calculate the values of the nodes, which are subsequently used as inputs in the calculation of the nodes in the second layer. Each node processes information by applying a linear or non-linear activation function. In addition to an output layer where the final results are obtained, a neural network might comprise one or more intermediate layers (hidden layers). Artificial neural networks differ in general based on the number of intermediate layers and the activation function utilized (Louzada et al. 2016). In practice, models of artificial neural networks with dozens of layers might be utilized, meaning thousands or millions of parameters that can necessitate significant computer resources. The fundamental advantage of neural networks is their ability to deal with complex interactions in vast datasets. However, because deciphering how the results are acquired is challenging, artificial neural networks are regarded as "black boxes." Deep learning, an extension of neural network algorithms, is one of the subfields of machine learning that has garnered considerable interest in recent years due to the impressive results gained in a wide range of applications (Bazarbash 2019).

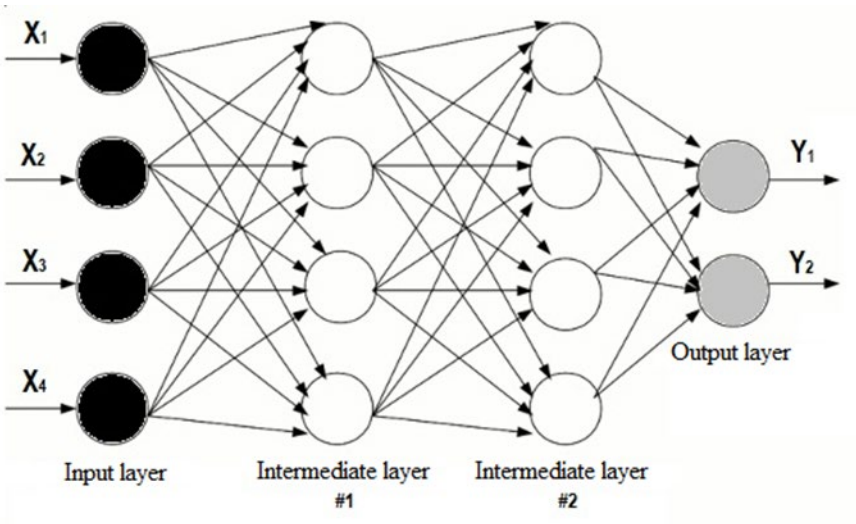

Figure 2. Artificial neural network with four attributes, two intermediate layers, and binary output.

There are numerous ways to combine algorithms and generate ensemble models. Boosting methods can be used to sequentially combine a group of lesser performing classifiers in order to generate a classifier with higher predictive power. In general, the boosting approach involves training numerous models sequentially, and the error function used to train a specific algorithm is determined by the performance of prior algorithms. AdaBoost, Gradient Boosting, Extra Trees, and XGBoost are some of the most used approaches (Bishop 2006; Xia et al. 2020). Another method for combining models that is simpler is to average the predictions of a set of independent models, like in the case of bagging random forest algorithms.

3.2. Performance Indicators

Considering the performance measures commonly employed in credit risk classification studies (Table 1), measurements generated from the confusion matrix and area under the curve are used in this work. The confusion matrix compares the outcome of the algorithm's classification with the actual classes observed in the data set. A misclassification happens in the context of binary credit risk scoring when the algorithm assigns a class (charged-off or fully paid loan) to individual or company i that differs from the true class observed in the data set. Table 2 depicts the confusion matrix, where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives, and $TP + TN + FP + FN = N$, where N is the total number of observations in the data set.

Table 2. Confusion matrix.

	Predicted	
Observed	Charged-off	Fully paid
Charged-off	*TP*	*FN*
Fully paid	*FP*	*TN*

We can get the following measurements from the confusion matrix: accuracy, recall, specificity, precision, and F1-score. Accuracy is a global predictive performance metric that correlates to the algorithm's share of correct classifications, denoted as

$$\text{Accuracy} = \frac{TP + TN}{N}$$

Despite its widespread use, accuracy may be insufficient to deal with uneven data sets that benefit a majority class. The true positive rate, also known as recall, is the proportion of positive cases (charged-off loans) properly classified by the algorithm in relation to the total number of true positives:

$$\text{Recall} = \frac{TP}{TP + FN}$$

The proportion of negative cases (fully paid loans) correctly classified by the algorithm in respect to the total number of genuine negatives is referred to as specificity, also known as the true negative rate:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Precision is the fraction of positive cases (charged-off loans) accurately identified by the algorithm in respect to the total number of positives:

$$\text{Precision} = \frac{TP}{TP + FP}$$

The F1-score is

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Through its harmonic mean, the F1-score provides a synthetic measure of precision and sensitivity. This measure outperforms accuracy in the situation of unbalanced datasets.

The area under the curve (AUC) is utilized in addition to the performance measures mentioned above. The curve in question is the receiver operating characteristic (ROC) curve, which is obtained by plotting the true positive rate (sensitivity) on the y axis and the false positive rate (1—specificity) on the x axis, and taking this relationship into account for different cutoff points on the probability estimated by the classification algorithm. In Figure 3, the wine dataset (toy dataset from scikit-learn) was utilized to train a random forest classifier and generate a visual plot for comparison with the support vector machines for classification (SVC), using the ROC curve and AUC.

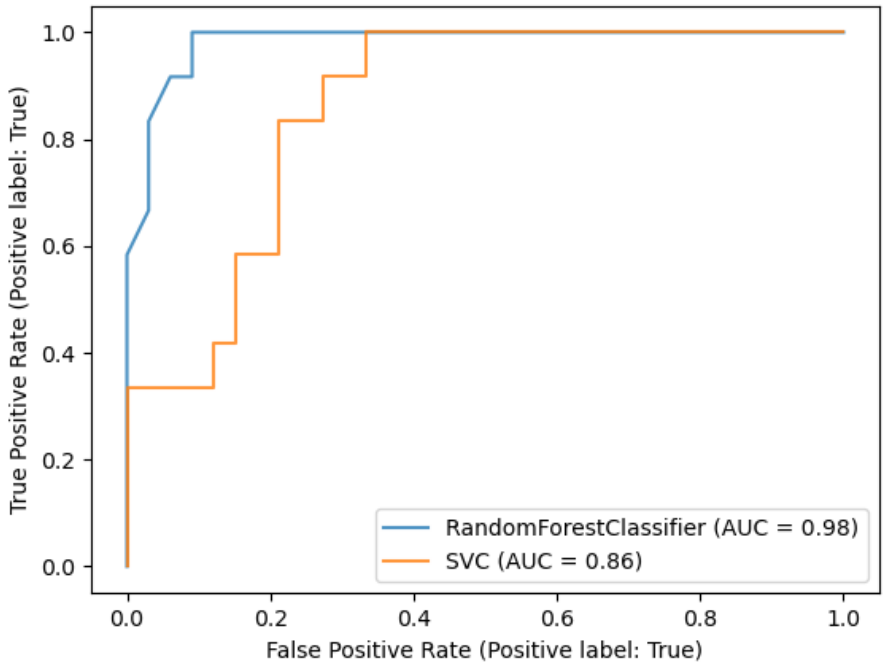

Figure 3. The higher the AUC value, or the closer the ROC curve is to the upper left corner, the better the classifier's performance. Source: https://scikit-learn.org/stable/auto_examples/miscellaneous/plot_roc_curve_visualization_api.html#sphx-glr-auto-examples-miscellaneous-plot-roc-curve-visualization-api-py (accessed on 15 November 2023).

4. Data Analysis

The selection of features is a vital step in applying machine learning algorithms to credit risk classification. In this section, we first provide an overview of the original dataset. Following that, we detail the pre-processing treatment and transformations, as well as the reasons for picking the variables that comprise the data set used in the simulations. Finally, we provide a description of each of the selected features.

4.1. Overview

When selecting the dataset, we first attempted to broaden the time of analysis and the number of observations in comparison to credit scoring research reported in the literature (Teply and Polena 2020; Malekipirbazari and Aksakalli 2015). We used the Lending Club loan dataset from the Kaggle repository (George 2018) for this, which originally gathered loans made by the Lending Club from 2007 to 2018, with 2,260,701 observations (loans) and 151 variables totaling 1.55 GB.

The original data set contains loan requests from individuals that were accepted by the Lending Club platform in the P2P lending model. Requests that were rejected, that is, those that did not satisfy the parameters of the platform's credit policy, were not taken into account in this study. Although this posed a risk of introducing survivorship bias,

this procedure was necessary due to the nature of the classification problem and the data available. Collective loans, or loans with more than one borrower, were also eliminated.

The target variable loan status was modified to indicate a binary result in order to deal with the proposed supervised classification challenge. Only loans that were fully paid or charged-off were considered in this change. As a result, the variable assumed the value 0 for fully paid and 1 for charged-off. Other scenarios, such as loans with status current, grace period, and in arrears of up to 120 days, were eliminated. Only loans that had already reached their full term and had been resolved, or those that had been charged-off with a delay of more than 120 days, were considered.

This initial treatment yielded a data set with 1,076,751 fully paid loans and 268,559 charged-off loans (Figure 4). Figure 5 also depicts the number of loans in each class over the investigated period. The decrease in the total number of loans observed in 2016 is due to the removal of operations in progress, as indicated above for the target variable loan status. The disparity between the fully paid and charged-off classes will be addressed later.

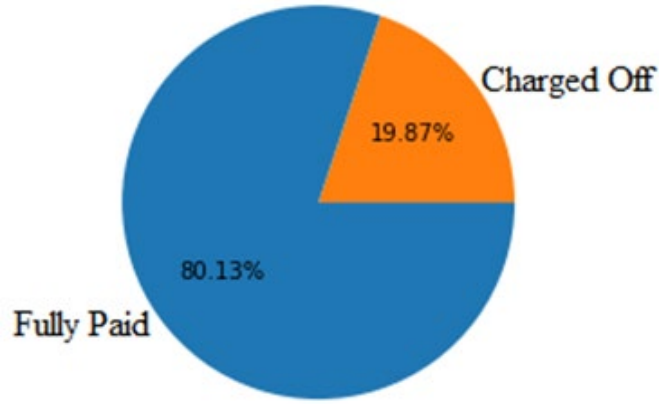

Figure 4. Loan status.

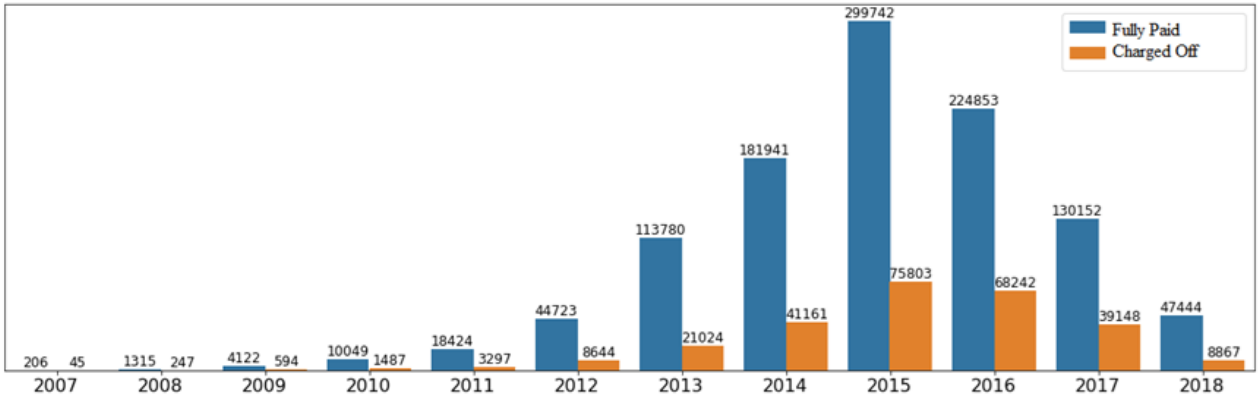

Figure 5. Loans made each year.

4.2. Data Pre-Processing

In general, a significant number of features had to be disregarded due to the amount of missing data that made any attempt to use them unfeasible. Figure 6 presents the missing values matrix for the original data set, with 44 features having 50% or more missing data. After a detailed analysis of each variable, it was found that some features had redundancy or a lack of informational value for the credit risk classification. This occurs, for example, with variables that represent transaction or borrower identification numbers. In addition, considering that the objective of this work is to evaluate the predictive capacity of credit risk classification algorithms, only variables available at the time borrowers requested the

loans were included. Thus, variables such as those identifying the borrower's payment history and schedule or their current credit score were not considered.

In addition to the target variable loan status, seven quantitative and eleven qualitative features were selected from a total of 151 accessible variables in the original data set. In the selection of features, we attempted to capture aspects related to the credit risk analysis dimensions outlined in Section 2.1. Thus, certain features are contingent upon the capacity, character, and circumstances of the borrower at the time of loan application. In addition, several features capture the loan's general features. Each of the 18 selected features is described in Table 3.

Table 3. Selected features.

Feature	Description	Type
Annual income	Self-reported by the loan applicant in US dollars at the time of registration	Numeric
Debt-to-income ratio	Monthly debt payments (excluding mortgages and the sought loan) as a percentage of monthly income	Numeric
Limit surpassed	Ratio between the amount of credit the borrower is using and all available revolving credit (e.g., credit cards)	Numeric
Credit availability	Total number of open credit lines reflected in the borrower's credit file	Numeric
Banking partnership	Total number of credit lines currently in the borrower's credit file	Numeric
Financial past	Time, in years, after the borrower opened his or her first credit line until the time of the request	Categorical (possible values: up to 5 years, 6–10 years, 11–15 years, 16–20 years, and over 20 years)
Credit score	The value of the lower limit of the borrower's score range (FICO® Score) at the time of the request	Numeric
Delayed payments	Indicator of the existence of payment commitments that are more than 30 days past due in the recent two years	Categorical (possible values: yes or no)
Credit applications	Number of credit inquiries in the last six months, excluding autos and mortgages	Categorical (possible values: 0, 1, 2, 3 or 3+)
Pending registration	Indicator of the presence of derogatory public records	Categorical (possible values: yes or no)
Tax liens	Indicator of the existence of outstanding tax issues in the borrower's history	Categorical (possible values: yes or no)
Employment length	Employment length in years	Categorical (possible values: up to 1 year, 2–3 years, 4–5 years, 6–10 years, 10+ years)
Housing type	Housing situation of the borrower at the time of application	Categorical (possible values: own, mortgaged, rented, other)
Income verification	Validation indicator of the borrower's informed worth or source of income at the time of the request	Categorical (possible values: verified, not verified, verified source)
Loan amount	The loan application in US dollars	Numeric
Loan interest rate	The loan's annual interest rate	Numeric
Loan term	Total loan term expressed in months	Categorical (possible values: 36 or 60 months)
Loan purpose	The borrower's chosen purpose or objective for the loan at the time of application	Categorical (possible values: debt restructuring, credit card, remodeling, purchases, health, small business, vehicle, moving, vacation, property, marriage, renewable energy, education, other)

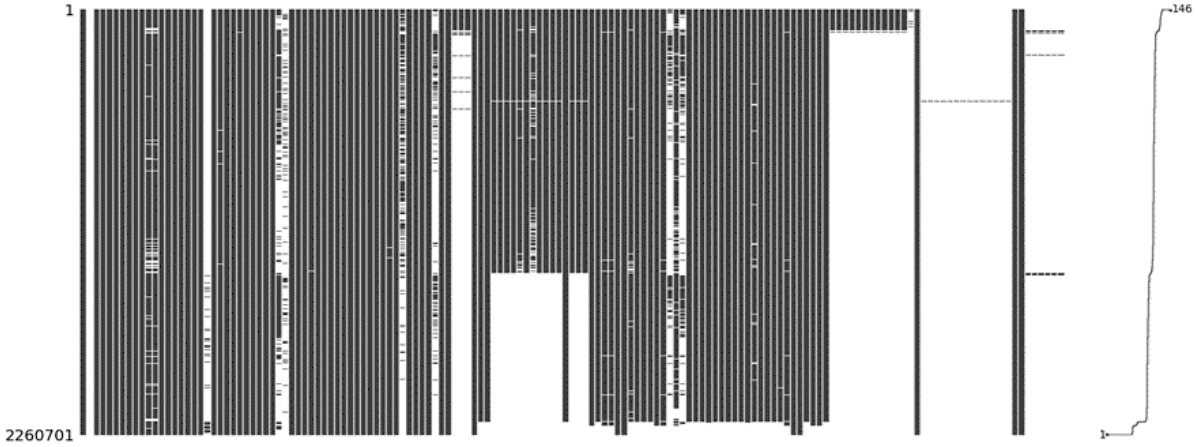

Figure 6. Missing matrix. Each column in the matrix represents a variable found in the original dataset, and any empty spaces denote missing values.

Transformations were performed in some of the selected features based on the identification of erroneous, missing, or conflicting values. Outliers higher than 2.5 standard deviations from the mean were detected and deleted for the annual income categorical variable (13,288 (1.01%)). A negative value for the debt-to-income ratio was found and eliminated, due to the fact that only positive or zero values are expected for this attribute. Two outlier values, both more than 200, were detected and deleted from the limit surpassed variable, which had a mean value of 52 and a third quartile value of 71 as a reference. The maximum value for the credit availability variable was set at 40 in order to decrease the number of possible values without sacrificing information. As a result, 1199 (0.09%) values greater than the maximum set for this characteristic were found and deleted. Similarly, the maximum value for the banking partnership variable was set at 80, hence 1281 (0.09%) values greater than the predefined maximum were found and deleted. The financial past variable was created by calculating the number of years elapsed between the borrower's initial line of credit and the time of the credit request. The attributes delayed payments, pending registration, and tax liens were converted into binary variables with possible values yes or no. Because of the huge number of categories in the original data set, the variables credit applications and employment length were discretized to reduce the number of possible values while keeping them informative.

4.3. Exploratory Data Analysis

The data set resulting from the previous step's pre-processing contains 1,305,402 observations (loans), of which 1,045,072 (80.06%) are fully paid and 260,330 (19.94%) are charged-off. Seven of the eighteen selected criteria are quantitative, five are continuous (annual income, debt-to-income ratio, limit surpassed, loan amount, loan interest rate), and two are discrete (credit availability, banking partnership). Table 4 provides a statistical summary of the most important summary metrics for each of the chosen quantitative attributes.

Figure 7 depicts a preliminary examination of the existing correlations between the quantitative features and the dependent variable loan status. In general, the exploratory analysis employs boxplots to discover changes in the summary measures of a quantitative feature within each category of the loan status target attribute as an indicative of the degree of correlation between the variables. Figure 7 shows that only the top two quantitative features appear to be related to loan status. The preliminary visual examination reveals a positive relationship between the debt-to-income ratio and charged-off loans; that is, a higher debt-to-income ratio appears to enhance the likelihood of charged-off loans. Similarly, visual inspection of the loan interest rate variable reveals that higher interest rates appear to be related to an increased likelihood of charged-off loans. However, a visual examination of the quantitative variables annual income, limit surpassed, credit availability, banking partnership, and loan amount does not reveal a significant relationship with the

loan status variable. This is because it is not possible to show a significant difference between the position and dispersion measures within the fully paid and charged-off loan classes by visually evaluating each attribute.

Table 4. Statistical summary of quantitative attributes.

	Annual Income	Debt-to-Income Ratio, %	Limit Surpassed, %	Credit Availability	Banking Partnership	Loan Amount	Loan Interest Rate, %
Mean	73.13	18.1	51.85	11.57	24.93	14.21	13.23
S.D.	38.55	8.35	24.44	5.42	11.91	8.57	4.75
Minimum	2	0	0	0	2	0.5	5.31
25%	46	11.85	33.5	8	16	7.75	9.75
50%	65	17.61	52.2	11	23	12	12.74
75%	90	23.97	70.7	14	32	20	15.99
Maximum	252.4	49.96	193	40	80	40	30.99

Note: Annual income and loan amount are in thousands.

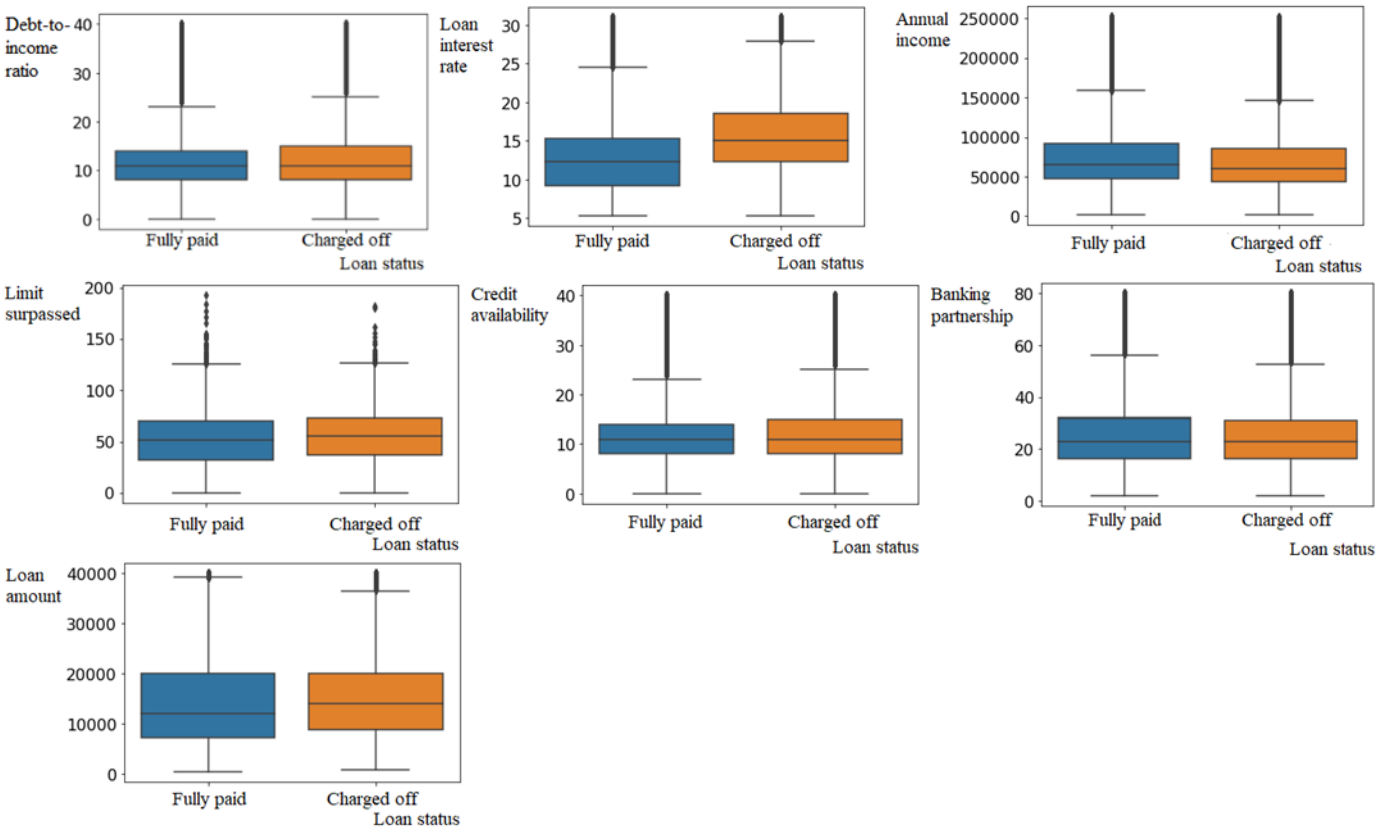

Figure 7. Only the top two quantitative features appear to be related to loan status.

Some indicators of linkage can be found by visually inspecting the joint distribution of the variables, taking into account each class of the loan status target characteristic (Figure 8). As a result, because the variables are independent, the same proportions should be expected for the fully paid and charged-off loan groups. First, comparing the proportions of fully paid and charged-off loans according to loan term reveals a considerable disparity within each class. As a result, a longer loan term (60 months) appears to be connected with a higher risk of charged-off loans. Borrowers with delayed payment, pending registration, or tax liens are also at a higher risk of having their loans charged-off. Second, there are a lot of categories in the credit score and loan purpose attributes, which makes visual examination in Figure 8 tough. However, based on the proportion of fully paid and charged-off loans found in each category, it is possible to see that a lower credit score appears to be related to a

higher risk of charged-off loans. In terms of loan purpose, loans aimed at small enterprises appear to have a higher likelihood of being charged-off. Third, it can be shown that the income verification provided by the borrower at the time of application does not appear to contribute to lowering the likelihood of charged-off loans. Borrowers with unconfirmed income, in contrast, appear to be less likely to have their loans charged-off.

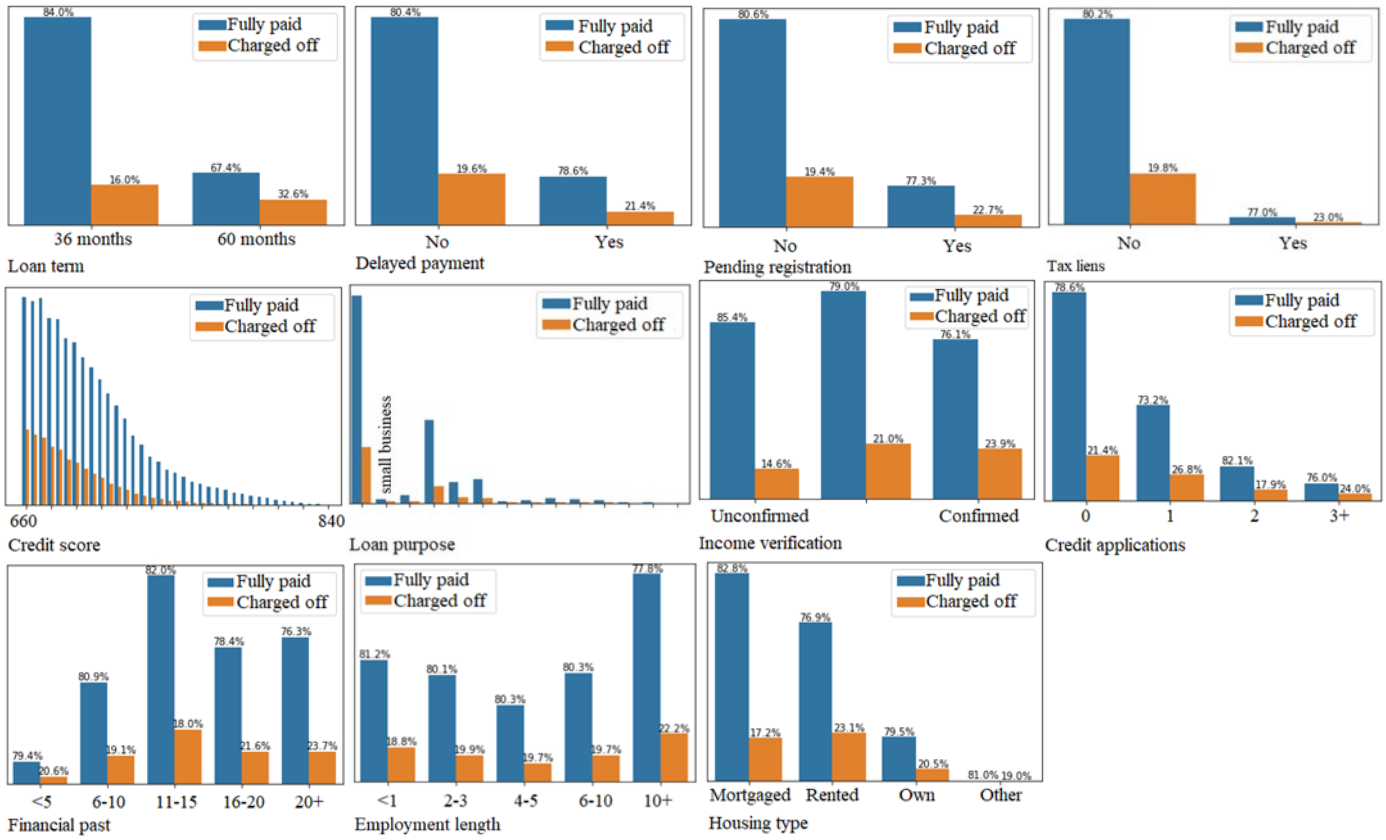

Figure 8. Selected features and loan status.

The study of the joint distributions, taking into account each class of the target attribute, does not enable establishing a link between the variables for the features credit applications, financial past, employment length, and housing type. Finally, the correlation matrix produced for the selected quantitative variables reveals only one value greater than 0.5 in absolute terms. The greatest association, with a value of 0.7, is discovered between the banking partnership and credit availability variables. To avoid the loss of potentially significant information for credit risk classification, none of the 18 previously specified features were eliminated from the data set.

5. Experimental Results

This section begins with the experiment design, with the goal of evaluating the prediction performance of the algorithms chosen for credit risk classification, as explained in Section 3.1. The findings are then compared to other studies that employ a similar data set. In terms of computing resources, the simulations were carried out on a computer with an Intel® CoreTM i5-1035G1 CPU running at 1.19 GHz and 8 GB of RAM memory, using the JupyterLab® software, version 3.2.1, and the Python programming language, version 3.9.13.

5.1. Experiment Design

As described in Section 3.1, the initial selection of algorithms was based on the key findings in the literature concerning the classification of credit risk. Thus, the ten algo-

rithms in Table 5 were chosen, which also displays the starting configuration of simulation hyperparameters.

Table 5. Selected algorithms and hyperparameters.

Algorithm	Hyperparameter
Logistic regression (LR)	class_weight="balanced"
Decision tree (DT)	algorithm=CART, class_weight="balanced", max_depth=7, min_samples_leaf=0.01
K-nearest neighbors (KNN)	n_neighbors=11
Support vector machines (SVM)	kernel="rbf", class_weight="balanced", max_iter=100,000
Artificial neural networks (ANN)	hidden_layer_sizes=(8,4), activation="relu", solver="adam", learning_rate= 0.001("constant")
Random forests (RF)	class_weight="balanced", max_depth=7, min_samples_leaf=0.01
Extra trees (ET)	class_weight="balanced", max_depth=7, min_samples_leaf=0.01
AdaBoost (ADA)	algorithm= SAMME.R
Gradient boosting (GB)	loss="log_loss", learning_rate=0.1, n_estimators=100, max_depth=3, min_samples_leaf=1
XGBoost (XGB)	booster=gbtree, learning_rate=0.3, gamma=0, alpha=0, min_child_weight=1, max_depth=6, sampling_method=uniform

We used the 5-fold cross-validation method in the simulations (Malekipirbazari and Aksakalli 2015). To do this, the pre-processed data set was initially separated into two subsets, the first containing 80% of the total observations and the second containing the remaining 20%. The training subset was then partitioned at random into five further subsets (folds) of the same size. In the experiments, each model was trained using four subsets resulting from the prior partitioning and evaluated using the subset that was not utilized for training. This was repeated five times, with each run selecting a different subset reserved for validation. The testing subset from the pre-processed data set was saved for later use in the fine-tuning stage and final algorithm evaluation, as shown in Figure 9.

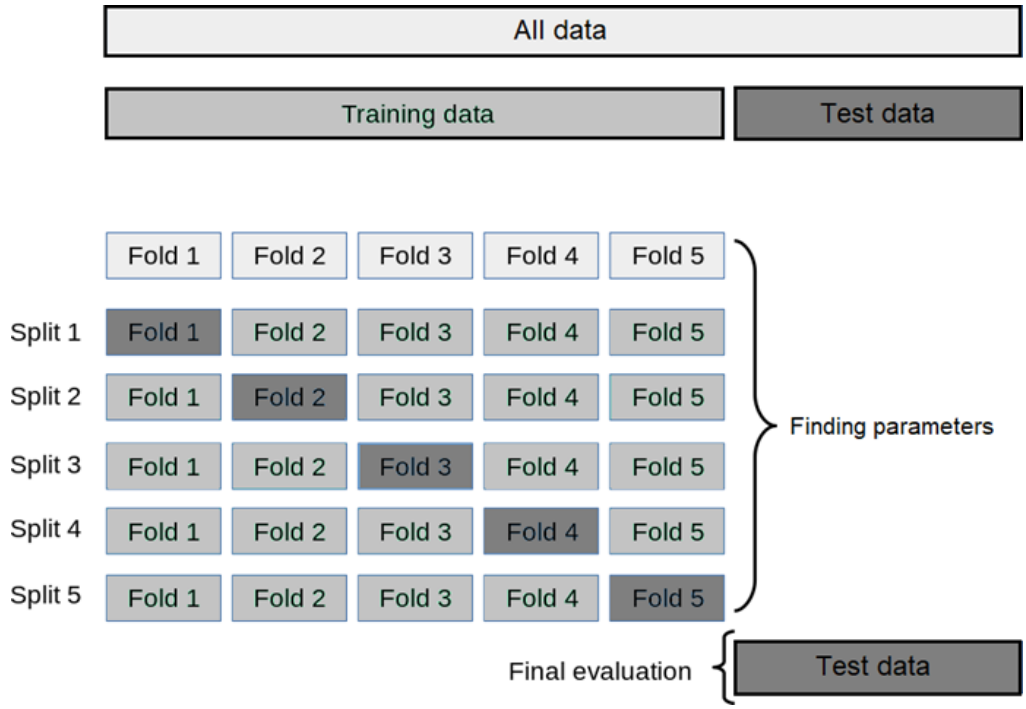

Figure 9. Selected features and loan status. Source: Adapted from https://scikit-learn.org/stable/modules/cross_validation.html. Access on 15 November 2023.

In addition to the data treatment outlined in Section 4.2, and taking into account the input formatting requirements of the chosen algorithms, all qualitative features were converted to numerical ones. Thus, integer values matching the ordering of the categories

of each variable were employed in the case of the financial past and employment length. Dummy variables were used to modify the nominal features housing type, income verification, and loan purpose. Finally, the data were normalized so that each feature was in the interval [0, 1] using the MinMaxScaler method from the scikit-learn v1.2.0 library.

5.2. General Results

After conducting the simulations using 5-fold cross-validation, the prediction performance of the models is compared in this section. To evaluate the chosen classifiers and facilitate comparisons with existing literature, we used metrics such as accuracy, precision, sensitivity, and F1, derived from the confusion matrix. Additionally, we utilized the AUC (area under the curve) metric, as given by the receiver operating characteristic (ROC) curve, as presented in Section 3.2.

The algorithms in this initial approach were generated without hyperparameter fine-tuning, that is, from the default configuration commonly found in the literature, with minor tweaks to minimize the possibility of overfitting and restrict the maximum run duration (Table 5). Table 6 summarizes the outcomes achieved by each model in the simulations performed. We give the mean and standard deviation achieved in the five iterations with cross validation for each of the performance measures employed.

Table 6. Model performance on the entire dataset.

Model	AUC	Accuracy	Precision	Recall	F1	Time * (s)
Logistic regression (LR)	0.7087 [0.0019]	0.6568 [0.0013]	0.3205 [0.0026]	0.6403 [0.0037]	**0.4272 [0.0030]**	14.0
Decision tree (DT)	0.6998 [0.0018]	0.6167 [0.0076]	0.3003 [0.0039]	0.6896 [0.0104]	0.4183 [0.0027]	8.0
K-nearest neighbors (KNN)	0.6438 [0.0012]	0.7921 [0.0007]	0.4176 [0.0048]	0.1028 [0.0016]	0.1650 [0.0023]	7806.0
Support vector machines (SVM)	0.5488 [0.0078]	0.3759 [0.0546]	0.2012 [0.0035]	**0.7127 [0.0762]**	0.3130 [0.0061]	18,453.0
Artificial neural networks (ANN)	0.7141 [0.0018]	0.8031 [0.0010]	**0.5640 [0.0115]**	0.0640 [0.0033]	0.1149 [0.0052]	43.0
Random forests (RF)	0.7032 [0.0012]	0.6283 [0.0022]	0.3058 [0.0025]	0.6773 [0.0030]	0.4214 [0.0026]	91.0
Extra trees (ET)	0.6919 [0.0011]	0.6496 [0.0005]	0.3098 [0.0019]	0.6138 [0.0020]	0.4118 [0.0020]	69.0
AdaBoost (AB)	0.7087 [0.0013]	0.8023 [0.0010]	0.5439 [0.0063]	0.0657 [0.0052]	0.1172 [0.0082]	83.0
Gradient boosting (GB)	0.7128 [0.0017]	0.8029 [0.0009]	0.5637 [0.0080]	0.0610 [0.0006]	0.1101 [0.0011]	351.0
XGBoost (XGB)	**0.7185 [0.0015]**	**0.8036 [0.0010]**	0.5549 [0.0081]	0.0857 [0.0019]	0.1484 [0.0031]	52.0

Notes: For each measure, the model with the best performance is bolded, and the standard error is displayed in brackets. * The mean time, expressed in seconds, that each cross-validation split takes to fit the estimator on the train set and assess it on the test set.

As previously stated, the data set used in this study has a disparity between the fully paid (80%) and charged-off (20%) loan classes. In this situation, there is a roughly 1:4 imbalance in favor of the majority class. As a result, the imbalance in some classification algorithms can benefit the majority class while presenting a weaker predictive performance for the minority class. As noted by Dastile et al. (2020), despite advancements in the applications of machine learning models in credit scoring, issues related to imbalanced datasets are still frequently overlooked. To address this issue, a subsampling technique was used, with the assumption that the amount of data is large enough to run simulations on the resulting data set. For this, observations from the fully paid majority class are randomly removed using the imblearn library's RandomUnderSampler method.

The resampled data set contains 416,528 observations evenly divided between fully paid (50%) and charged-off (50%) loan classes. Table 7 summarizes the results achieved by each model in the simulations run with the balanced data set using the same algorithms and hyperparameters used in the simulations with the whole data set. As expected, the resampling approach considerably improves the predictive performance of the minority charged-off loans class, as shown by the rise in the precision, recall, and F1 measures.

Table 7. Performance of models on the balanced dataset.

Model	AUC	Accuracy	Precision	Recall	F1	Time * (s)
Logistic regression (LR)	0.7076 [0.0015]	0.6499 [0.0017]	**0.6531 [0.0020]**	0.6393 [0.0031]	0.6461 [0.0024]	5.6
Decision tree (DT)	0.6988 [0.0013]	0.6431 [0.0016]	0.6366 [0.0031]	0.6670 [0.0139]	0.6513 [0.0053]	2.8
K-nearest neighbors (KNN)	0.6572 [0.0013]	0.6150 [0.0012]	0.6157 [0.0012]	0.6119 [0.0019]	0.6138 [0.0013]	1327.4
Support vector machines (SVM)	0.6111 [0.0117]	0.5335 [0.0362]	0.5309 [0.0378]	**0.8060 [0.1709]**	0.6279 [0.0353]	6951.2
Artificial neural networks (ANN)	0.7116 [0.0018]	0.6527 [0.0017]	0.6440 [0.0048]	0.6837 [0.0181]	**0.6631 [0.0063]**	33.7
Random forests (RF)	0.7026 [0.0008]	0.6459 [0.0015]	0.6390 [0.0025]	0.6709 [0.0039]	0.6545 [0.0019]	32.2
Extra trees (ET)	0.6925 [0.0008]	0.6369 [0.0009]	0.6468 [0.0011]	0.6030 [0.0035]	0.6241 [0.0022]	26.3
AdaBoost (ADA)	0.7078 [0.0012]	0.6500 [0.0011]	0.6441 [0.0014]	0.6704 [0.0030]	0.6570 [0.0019]	31.0
Gradient boosting (GB)	0.7118 [0.0010]	0.6531 [0.0016]	0.6476 [0.0022]	0.6717 [0.0015]	0.6594 [0.0018]	130.0
XGBoost (XGB)	**0.7153 [0.0013]**	**0.6563 [0.0018]**	0.6507 [0.0022]	0.6749 [0.0019]	0.6626 [0.0020]	16.5

Notes: For each measure, the model with the best performance is bolded, and the standard error is displayed in brackets. * The mean time, expressed in seconds, that each cross-validation split takes to fit the estimator on the train set and assess it on the test set.

Boosting-based ensemble methods like XGBoost and Gradient boosting generally outperform others on both total and balanced datasets. XGBoost particularly excels in AUC and accuracy on the entire dataset. Table 8 reveals that, except for recall, XGBoost ranks as the top or near-top performer on the balanced dataset metrics. Figure 10 further displays each model's performance on the balanced dataset.

Table 8. Ranking of models on the balanced dataset.

Model	AUC	Accuracy	Precision	Recall	F1
XGBoost (XGB)	1	1	2	3	2
Gradient boosting (GB)	2	2	3	4	3
Artificial neural networks (ANN)	3	3	6	2	1
AdaBoost (ADA)	4	4	5	6	4
Logistic regression (LR)	5	5	1	8	7
Random forests (RF)	6	6	7	5	5
Decision tree (DT)	7	7	8	7	6
Extra trees (ET)	8	8	4	10	9
K-nearest neighbors (KNN)	9	9	9	9	10
Support vector machines (SVM)	10	10	10	1	8

Note: The number indicates the model's position in each performance measure among the ten evaluated algorithms.

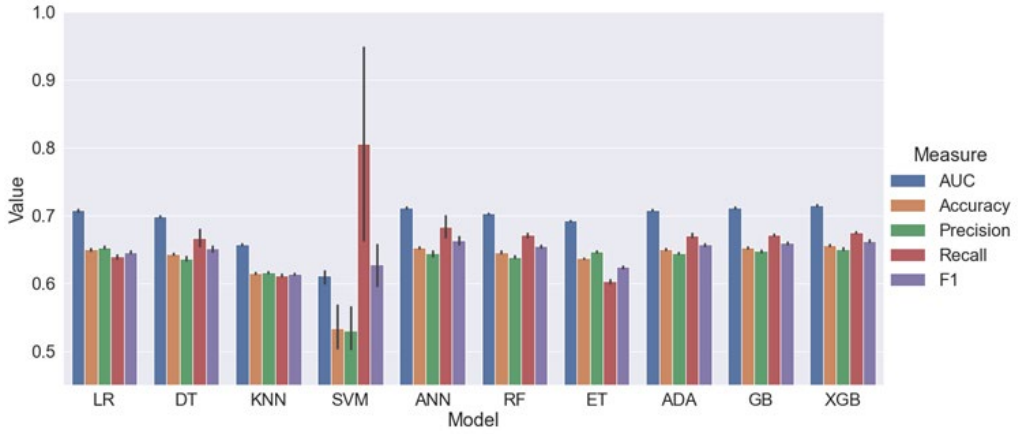

Figure 10. Model performance when resampled. The error bars represent the standard deviation.

The varying results among machine learning models in this study likely arise from different factors. Ensemble tree models like XGBoost, along with Artificial Neural Networks, outperform KNN, SVM, and logistic regression. This superiority is due to their better handling of complex datasets, capturing intricate data patterns. In contrast, KNN,

SVM, and logistic regression may not adapt as well to complex data, leading to lower performance. Recognizing these differences is key in practical applications, emphasizing the importance of selecting the most suitable model for credit scoring scenarios and dataset specifics to ensure optimal predictive performance.

5.3. Hyperparameter Optimization and Final Evaluation

In addition to the final evaluation, the following stage is devoted to hyperparameter optimization, or fine-tuning, based on the Section 5.2 results. For this purpose, the XGBoost algorithm was chosen since, among the previously selected models, it has the best predictive performance on the key measures evaluated. The optimization procedure entails exploring the hyperparameter space of the model in order to enhance the predictive performance based on the selected evaluation metric, which in this case is the AUC. Then, the final evaluation of the model is performed using the testing set coming from the initial division of the pre-processed data set, which has not been utilized for training or validating the models in any prior stage.

In general, grid search or random search methods are utilized to carry out the hyperparameter optimization procedure. The grid search is an exhaustive search for the optimal hyperparameter combination from a collection of previously stated values. In the case of random search, random hyperparameter combinations are conducted from a list of previously established values, with the number of iterations expressly limited. According to Bergstra and Bengio (2012), random search procedures are more efficient for hyperparameter optimization than grid search, finding models as good as or better with substantially shorter execution time.

In order to optimize the hyperparameters, we used the RandomizedSearchCV technique from the scikit-learn library. Table 9 shows the collection of hyperparameters and the value grid for the random search. The ideal value of each hyperparameter was determined using 5-fold cross-validation on a balanced data set, with a limit of 200 iterations for all conceivable configurations. The performance of the XGBoost model in the testing set was evaluated using the hyperparameters discovered during the optimization phase.

Table 9. Hyperparameters and values.

Hyperparameter	Value Grid	Optimal Value
max_depth	[3, 6, 8, 10, 12, 15, 20]	6
min_child_weight	[1, 3, 5, 7, 10]	1
gamma	[0, 0.0001, 0.001, 0.01, 0.1]	0.001
learning_rate	[0.01, 0.05, 0.1, 0.2, 0.3, 0.5]	0.3
alpha	[0, 0.0001, 0.001, 0.01, 0.1]	0.01
subsample	[0.1, 0.25, 0.5, 0.75, 1]	1
colsample_bytree	[0.1, 0.25, 0.5, 0.75, 1]	0.75
colsample_bylevel	[0.1, 0.25, 0.5, 0.75, 1]	0.75
colsample_bynode	[0.1, 0.25, 0.5, 0.75, 1]	1

Table 10 summarizes the major predictive performance measures of the XGBoost model following the hyperparameter optimization process with the testing data set. In general, the accuracy metric indicates that the model's global prediction performance is 0.65, implying that 65% of total classifications are right. When performance is broken down by class, the precision metric for fully paid loans, which corresponds to the proportion of negative cases properly classified by the model in relation to the total number of negative cases, reveals an 89% success rate. In terms of the charged-off loan class, the model has a worse performance, with only 32% of the total classifications made by the algorithm being right.

Table 10. XGBoost model performance with hyperparameter optimization.

Class	Precision	Recall	F1
Fully paid loan (0)	0.8880	0.6393	0.7434
Charged-off loan (1)	0.3151	0.6729	0.4292
Macro	0.6015	0.6561	0.5863
Balanced	0.7746	0.6459	0.6812
Accuracy			0.6459

According to the recall metric, the algorithm has a true positive rate of 67% for the charged-off loan class, which relates to the proportion of correctly classified positive cases in relation to the total number of true positives. Similarly, when it comes to the fully paid loan class, the algorithm has a genuine negative rate of 64%. Figure 11 also shows the AUC measure, which is represented by the area under the ROC curve and is obtained by plotting the false positive rate against the true positive rate and taking this relationship into account for different cutoff points in the probability determined by the algorithm.

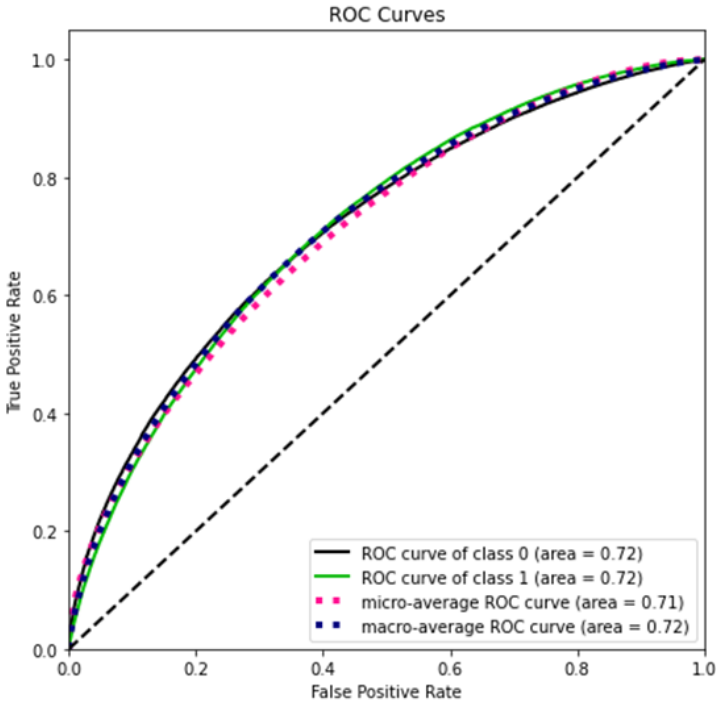

Figure 11. ROC curve and AUC. The AUC is the area under the ROC curve, calculated by plotting the false positive rate against the true positive rate.

One advantage of employing decision tree-based algorithms, such as XGBoost, is the ability to calculate the contribution of each feature to the model's success. The contribution of each feature in the creation of decision trees is measured by importance; that is, the bigger the average information gain in all divisions in which a feature is employed, the greater its relative importance. Because the importance of each existing feature in the data set is explicitly estimated based on mean decrease in impurity (MDI), it is easy to do a ranking by degree of importance for the five most significant features (Figure 12). The loan term is the most essential feature, accounting for 31% of the total, followed by the loan interest rate (15%).

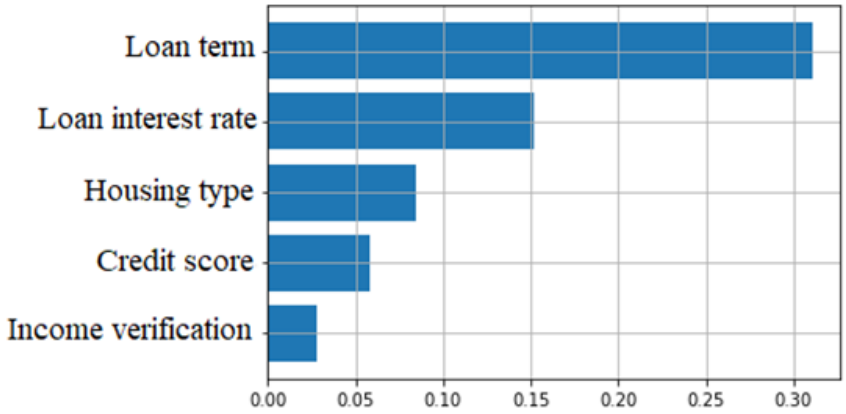

Figure 12. Relative importance of features.

5.4. Benchmarking Results and Practical Implications

Numerous studies have examined the performance of various statistical and classical machine learning models in credit scoring, but a consensus on the best model remains elusive (Dastile et al. 2020). A meta-analysis by Abdou and Pointon (2011) involving 214 studies found no optimal technique for variable selection, sample size, validation criteria, or model performance in credit scoring. They emphasize that no method currently guarantees optimal outcomes, regardless of the dataset used. Furthermore, comparisons reveal that advanced techniques like neural networks generally outperform traditional methods like logistic regression in terms of predictive capacity, but the differences are often minimal.

Lessmann et al. (2015) conducted comparison research on 41 classification approaches, taking into account eight data sets and six predictive performance measures, to provide an overview of the achievements made in the design of credit scoring models. The authors' findings do not support the claim that complicated procedures offer little advantage over simpler methods. In this regard, heterogeneous ensembles outperform simple classifiers in terms of predictive performance, particularly in respect to logistic regression models. The authors' findings provide some support for the search for new models, as well as research that take into account other aspects involved in credit risk assessments, such as variable selection and data quality.

Our findings are consistent with those of Xia et al. (2020), Finlay (2011), and Malekipirbazari and Aksakalli (2015), who find that boosting algorithms outperform classical algorithms in empirical applications of credit risk rating models. It is worth mentioning that the computational cost of the XGBoost model in terms of execution time is rather cheap, compared to single classifiers such as logistic regression. In addition to selecting the algorithm with the best performance, aspects related to data collection and processing, exploratory data analysis, performance metrics, and result interpretability should not be overlooked in the formulation of a guiding framework for machine learning in credit risk assessment.

The traditional human-based 5C's approach has limitations when dealing with a high volume of daily loan applications. Machine learning techniques have revolutionized credit decision making by automating tasks and enabling more efficient processes. This automation allows financial institutions to provide loans faster and at lower costs, promoting financial inclusion, especially among smaller borrowers like small- and medium-sized enterprises (Bazarbash 2019). Additionally, machine learning-based classifications serve as valuable complementary tools for handling complex credit decisions.

6. Conclusions

This paper examines the prediction performance of ten credit risk classification algorithms using a massive amount of real data from a financial institution. The original data set includes loans made by the Lending Club from 2007 to 2018, with 2260,701 observations (loans) and 151 variables totaling 1.55 GB. In comparison to other credit scoring studies

(Serrano-Cinca et al. 2015; Malekipirbazari and Aksakalli 2015; Xia et al. 2020; Teply and Polena 2020), this one greatly increases the period of examination and the volume of data.

After an in-depth data analysis, we used a pre-processed dataset with 1,305,402 loans and 18 features. We tested using 5-fold cross-validation and assessed classifiers using accuracy, precision, recall, F1, and AUC metrics. Due to data imbalance, we applied a subsampling technique and reassessed the model performance. The results suggest that ensemble models utilizing boosting strategies outperform traditional benchmark algorithms like logistic regression (LR) and decision trees (DT) in terms of prediction. Specifically, the XGBoost model gets the best results when all measurements are considered, as well as benefits associated with lower processing costs in terms of execution time compared to the analyzed individual algorithms. For interpretability, XGBoost calculates feature importance using mean decrease in impurity (MDI). The loan term is the most critical feature, followed by the interest rate.

It is crucial for future research to confirm if classifiers' high performance on specific datasets translates to real-world use. Incorporating macroeconomic factors like growth rate and unemployment can help create dynamic models that capture more than just borrower details. Even with tech advancements and affordable solutions, there are still regulatory and organizational hurdles to the widespread use of machine learning for better financial inclusion and credit decisions.

Author Contributions: Conceptualization, N.S., J.U. and S.D.S.; methodology, N.S.; software, N.S. and J.U.; validation, J.U. and S.D.S.; formal analysis, N.S.; investigation, N.S.; resources, N.S., J.U. and S.D.S., data curation, J.U.; writing—original draft preparation, N.S.; writing—review and editing, S.D.S.; visualization, N.S. and S.D.S.; supervision, J.U.; project administration, J.U.; funding acquisition, N.S., J.U. and S.D.S. All authors have read and agreed to the published version of the manuscript.

Funding: FAPESP [Grant number: 2022/09644-7], BRDE [Grant number: 2021/366], CNPq [Grant number: PQ 2 301523/2019 3], and Capes [Grant number: PPG 001] supported this work.

Data Availability Statement: Data are from Kaggle and are available at https://www.kaggle.com/datasets/wordsforthewise/lending-club. Accessed on 15 November 2023.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

Abdou, Hussein A., and John Pointon. 2011. Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management* 18: 59–88. [CrossRef]

Araujo, Fabio. 2022. Initial Steps towards a Central Bank Digital Currency by the Central Bank of Brazil. BIS Papers No. 123. pp. 31–37. Available online: https://www.bis.org/publ/bppdf/bispap123.pdf (accessed on 15 November 2023).

Athey, Susan, and Guido W. Imbens. 2019. Machine learning methods that economists should know about. *Annual Review of Economics* 11: 685–725. [CrossRef]

Bali, Turan G., Heiner Beckmeyer, Mathis Moerke, and Florian Weigert. 2023. Predicting option returns with machine learning and big data. *Review of Financial Studies* 36: 3548–602. [CrossRef]

Bazarbash, Majid. 2019. FinTech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk. IMF Working Paper No. 2019/109. Available online: https://www.imf.org/-/media/Files/Publications/WP/2019/WPIEA2019109.ashx (accessed on 15 November 2023).

Berg, Tobias, Valentin Burg, Ana Gombović, and Manju Puri. 2020. On the rise of FinTechs: Credit scoring using digital footprints. *Review of Financial Studies* 33: 2845–97. [CrossRef]

Bergstra, James, and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13: 281–305. Available online: https://dl.acm.org/doi/pdf/10.5555/2188385.2188395 (accessed on 15 November 2023).

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Berlin: Springer Nature B.V.

Breiman, Leo. 2001. Statistical modeling: The two cultures. *Statistical Science* 16: 199–231. [CrossRef]

Cakici, Nusret, Christian Fieberg, Daniel Metko, and Adam Zaremba. 2023. Do anomalies really predict market returns? New data and new evidence. *Review of Finance*. Forthcoming. Available online: https://ssrn.com/abstract=4557747 (accessed on 15 November 2023).

Chakraborty, Chiranjit, and Andreas Joseph. 2017. Machine Learning at Central Banks. Bank of England Working Paper No. 674. Available online: https://www.bankofengland.co.uk/working-paper/2017/machine-learning-at-central-banks (accessed on 15 November 2023).

Dastile, Xolani, Turgay Celik, and Moshe Potsane. 2020. Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing* 91: 106263. [CrossRef]

Drobetz, Wolfgang, Fabian Hollstein, Tizian Otto, and Marcel Prokopczuk. 2021. Estimating stock market betas via machine learning. *SSRN*. [CrossRef]

Dua, Dheeru, and Casey Graff. 2017. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences. Available online: https://archive.ics.uci.edu/ml/index.php (accessed on 15 November 2023).

Finlay, Steven. 2011. Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research* 210: 368–78. [CrossRef]

George, Nathan. 2018. All Lending Club Loan Data. Available online: https://www.kaggle.com/wordsforthewise/lending-club (accessed on 15 November 2023).

Gu, Shihao, Bryan Kelly, and Dacheng Xiu. 2020. Empirical asset pricing via machine learning. *Review of Financial Studies* 33: 2223–73. [CrossRef]

Hand, David J., and William E. Henley. 1997. Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 160: 523–41. [CrossRef]

Lessmann, Stefan, Bart Baesens, Hsin-Vonn Seow, and Lyn C. Thomas. 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research* 247: 124–36. [CrossRef]

Louzada, Francisco, Anderson Ara, and Guilherme B. Fernandes. 2016. Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science* 21: 117–34. [CrossRef]

Malekipirbazari, Milad, and Vural Aksakalli. 2015. Risk assessment in social lending via random forests. *Expert Systems with Applications* 42: 4621–31. [CrossRef]

Markov, Anton, Zinaida Seleznyova, and Victor Lapshin. 2022. Credit scoring methods: Latest trends and points to consider. *The Journal of Finance and Data Science* 8: 180–201. [CrossRef]

Serrano-Cinca, Carlos, Begoña Gutiérrez-Nieto, and Luz López-Palacios. 2015. Determinants of default in P2P lending. *PLoS ONE* 10: e0139427. [CrossRef] [PubMed]

Teply, Petr, and Michal Polena. 2020. Best classification algorithms in peer-to-peer lending. *North American Journal of Economics and Finance* 51: 100904. [CrossRef]

Varian, Hal R. 2014. Big data: New tricks for econometrics. *Journal of Economic Perspectives* 28: 3–28. [CrossRef]

Vicente, Julia. 2020. Fintech disruption in Brazil: A study on the impact of open banking and instant payments in the Brazilian financial landscape. *Social Impact Research Experience* 86. Available online: https://repository.upenn.edu/sire/86 (accessed on 15 November 2023).

Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, and et al. 2008. Top 10 algorithms in data mining. *Knowledge and Information Systems* 14: 1–37. [CrossRef]

Xia, Yufei, Lingyun He, Yinguo Li, Nana Liu, and Yanlin Ding. 2020. Predicting loan default in peer-to-peer lending using narrative data. *Journal of Forecasting* 39: 260–80. [CrossRef]

Zhang, Xiaoming, and Lean Yu. 2024. Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods. *Expert Systems with Applications* 237: 121484. [CrossRef]

Zhou, Xianzheng, Hui Zhou, and Huaigang Long. 2023. Forecasting the equity premium: Do deep neural network models work? *Modern Finance* 1: 1–11. [CrossRef]