Probability of Default Modeling with Machine Learning: A Comparison of Techniques Using Real Banking Data

Stefano Bonini[1] *– Giuliana Caivano*[2]

[1]University of Bologna - Lumsa University
[2]University of Milan – University of Bologna

ABSTRACT

In recent years, Machine Learning (ML) and Artificial Intelligence (AI) have experienced significant growth, driven by the increasing volume and variety of data, the availability of tools and software with greater computational power, and the reduced cost of data storage (e.g., cloud storage). In Credit Risk Management, the estimation of Probability of Default (PD) has attracted considerable research interest, with recent studies showing that advanced AI methods achieve better performance than traditional statistical methods based on simpler ML techniques. This study empirically investigates the outcomes of applying different advanced ML techniques to the estimation and calibration of Probability of Default. The research was conducted on a large dataset comprising over 800,000 retail customers from a panel of European banks under ECB supervision, covering 10 years of historical data (2012–2022) and analyzing 300 variables for each customer. We used the same data for both traditional and advanced ML techniques to compare their performances using the historical data currently available in banks' databases. The study shows that neural networks produce higher accuracy in ranking population risk, with an Accuracy Ratio of 71%, although this is only 5 percentage points higher than that of traditional logistic regression, which has an AR of 66%.

Key Words:
Risk Management, Credit Risk, Machine Learning, Big Data, Data Analysis, Advanced Predictive Analytics, Risk Modeling, Probability of Default, Credit Scoring

1. Introduction

Nowadays, Artificial Intelligence (AI) has become a buzzword to watch. The banking sector seems to have adopted this field of study only recently, although its foundations were laid when researchers began exploring the concept of computers learning from data. This is the fundamental premise of Machine Learning (ML, or automatic learning)—that computers can learn to perform tasks simply by identifying patterns and relationships within data, becoming more effective as the volume and availability of information increase.

The last few years have been marked by a technological and digital revolution that has created new opportunities to enhance efficiency within existing operational practices. This evolution has led to the adoption of more advanced methodological approaches across various research fields. AI and ML are gaining renewed momentum in an increasingly competitive environment with shrinking profit margins. This shift is driven by the growth in the volume and variety of data, the availability of processing tools with high computational power, and the decreasing costs of data storage (e.g., cloud storage). ML plays a crucial role in both technology and business, enabling financial institutions to manage large datasets more effectively and to facilitate model adaptation and recalibration.

In recent years, ML techniques have been developed to estimate binary variables across various scientific fields. In the context of Credit Risk Management, and particularly in modeling the Probability of Default (PD), the application of AI techniques has enhanced traditional statistical methods and improved performance, both in solving credit scoring problems ([18], [31], [39]) and in estimating the Probability of Default (see [3], [4], [8], [9], [23], [27]).

Despite the high predictive performance of ML methods, understanding and explaining the outputs and the relationship among variables is not straightforward [38]. For this reason, authorities and regulators have started to monitor the risks associated with adopting AI methods. From a European perspective, the European Commission proposed the AI Act, a set of regulatory principles for AI focusing on accuracy, explainability, fairness, and robustness [19]. Additionally, the Italian regulator published a discussion paper [6] on the adoption of Machine Learning in credit scoring.

The purpose of this article is to highlight the importance of choosing the right algorithms and parameters, selecting relevant variables (features), and understanding the role of evaluation criteria and expert contributions in defining the Probability of Default (PD) for a loan portfolio.

The literature reviewed in this paper primarily focuses on the application of ML techniques to datasets with many observations from a single financial institution. It also explores the creation of a cardinal measure for default events [29] using generalized classification algorithms and classification tree techniques. As discussed later in this paper, the research emphasizes: a) sorting and ranking capabilities, specifically how different ML and deep learning algorithms predict default events; b) the selection of variables by each algorithm as risk drivers; and c) the calibration power of estimates obtained through unsupervised techniques for defining rating classes.

In terms of methodology, this research builds on algorithms applied in existing literature. For example, in [12], the authors use decision trees, logistic regression, and random forests to analyze the level of consumer delinquency using data from six different banks. In another example, [26], the authors apply decision trees to a mortgage portfolio to predict default events and compare the results with k-nearest neighbors (KNN), artificial neural networks (ANN), and probit models.

Among these, classical regression models (both logistic and linear, as in example [14]) are the most used analytical tools in the banking industry. Logistic regression is used here as a benchmark to compare its fitting capacity (in terms of Accuracy Ratio and Correct Classification Rate) with that of other non-parametric

models—specifically, ML and deep learning techniques highlighted in recent literature. This comparison was conducted using three different approaches: classification trees (or decision trees), random forests, and deep learning (neural networks). These methods were applied to a sample of over 800,000 retail customers from a panel of European banks under the supervision of the ECB, with 10 years of historical data (2012–2022). The evaluation focused not only on the fitting capacity of each model compared to logistic regression but also on the set of variables selected by each model.

The paper is structured as follows: Section 2 reviews the literature on the main methodological approaches underlying the use of algorithms and explains how these methods operate in relation to specific objective or target variables to be modeled. Section 3 outlines the criteria used for classifying and comparing results. Section 4 presents the dataset used in this empirical study and discusses the main results obtained. Finally, Section 5 provides the study's conclusions and summarizes the potential future directions of the research.

2. Theoretical Aspects of the Most Widely Used Methodologies in the Literature

This section outlines two applications of machine learning algorithms:

a) Building a ranking/scoring model for the population (supervised learning);

b) Calibrating scores and defining the probability of default for each rating class (unsupervised learning).

2.1. Supervised learning for scoring definition

Our first research objective was to build a population ranking model using supervised learning.

Supervised ML algorithms involve a learner that operates with input/output datasets derived from historical data. Input data consists of attributes critical to defining the outputs, while the outputs are derived from these inputs. In many supervised ML algorithms, input data is associated with support vectors, which can contain either continuous or discrete features.
Supervised learning can be based on either classification or regression algorithms. Classification algorithms are used when the outputs are categorical, whereas regression algorithms are applied to predict continuous variables.
Once the input/output datasets are established, the learner assesses these vectors by processing historical data. While this approach is effective for mapping input/output data within a training subset, it may not perform well for predicting output values for new, unseen input data, or in cases where the training data contains noise. In such situations, supervised ML models aim to construct functions that generalize well to out-of-sample data.

In our study, we developed and compared various population ranking/scoring models using four supervised learning approaches:

a. Logistic regression;
b. Decision tree (CART);
c. Random Forest;
d. Neural network (deep learning).

Logistic regression is considered an extension of linear regression models built upon linear relations turned into exponential functions - so-called logistic transformations. In particular, logistic regressions analyze relations between multiple independent variables and single dependent dichotomic variables (this study case focuses on the variable “good/bad”) by means of estimates of probability scores, in order to discriminate at most the two groups identified by the dichotomous variable.

$$y_i = f(w_i) = \frac{1}{1+e^{-w_i}} \quad (1)$$

In Equation (2), the independent variable is defined as a linear function of the selected indicators.

$$w_i = \alpha + \sum_{j=1}^{m} \beta_j x_{i,j}$$

By combining the defined equations and including the error term, the following logit model is obtained:

$$y_i = \frac{1}{1+e^{-\alpha-\sum_j \beta_j x_{i,j}}} + \varepsilon_i \quad (3)$$

In the above equation, the values produced by the logistic function (the codomain) fall within the range (0,1). Consequently, the dependent variable y_i ranges between 0 and 100% and represents the probability of default.

On one hand, linear models such as logistic regression offer the advantages of producing accurate results and providing clear interpretations of linear relationships. On the other hand, these models struggle with categorical variables and high correlations between variables, which can cause issues. Additionally, they may perform poorly with non-linear relationships and are often prone to underfitting.

Classification and Regression Trees (CART) are widely used techniques in machine learning because they can address both regression and classification problems (see [34]). In these models, a dependent variable (either discrete or continuous) is related to a set of independent variables through recursive sequences based on binary splits (trees). These recursive sequences divide the multidimensional space of independent variables into distinct "regions," each defined by specific parameters. As illustrated in the regression tree below, each region is characterized by unique parameters.

In Fig. 1, the CART model includes two non-negative independent variables (x1, x2) referred to as the "feature vector," and a categorical dependent variable with values "good" and "bad." The sequence of binary splits in the tree, as shown in Fig. 1, partitions the domain of (x1, x2) into five distinct regions, determined by the parameters L1 through L4.
For the initial condition x1<L1 and x2<L2, the values of (x1, x2) are classified as a "bad" outcome. Conversely, for the condition x1<L1 and x2>=L2, the values of (x1, x2) are classified as a "good" outcome.

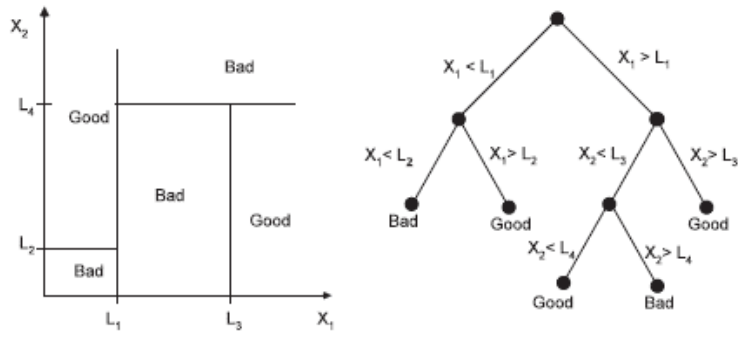


Figure 1 – Example of classification tree with a binary target variable

The choice of parameters such as (Lj) is aimed at minimizing the gap within each class or subset of the dependent variable while maximizing the separation between different categories. Both objectives are incorporated into the objective function.

$$D = \sum_{i=1}^{K} \left\{ \sum_{j \in S_i} \left(y_i - \widehat{\beta}_i \right)^2 \right\} = \sum_i D_i, \tag{4}$$

In algorithm (4) y_i represents the values of the target variable (0 and 1 in this study), and $\widehat{\beta}_i$ denotes the associated parameter within one of the KKK subspaces of the dataset S_i. This process is applied to all subsets until an optimal threshold value for the variable is identified.
The process is iterated until specific conditions are met, such as creating a subspace or cluster of data based on predefined criteria. This involves grouping data to ensure that subsets share the same category or target variable value, as seen in Classification and Regression Trees.

An algorithmic tree composed of multiple nodes and leaves often represents complex results. Due to the numerous splits and values, the outcomes can be difficult to interpret and sometimes inaccurate.
In such cases, a crucial step is to apply **tree pruning**: a process that sequentially removes nodes and leaves that are either unnecessary for the overall evaluation or carry irrelevant or inaccurate information.

This is achieved by defining a loss function computed as follows:

$$C_\alpha(K) = \sum_{i=1}^{K} D_i + \alpha K \tag{5}$$

In the loss function, K defines the dimension of the tree throughout the process, while α\alphaα represents the model's computational cost. The loss function is designed to facilitate the **pruning** of leaves and nodes whose removal does not significantly affect the overall function $\sum_{i=1}^{K} D_i$.

Decision trees are among the most widely used algorithms and are considered particularly effective machine learning tools. Their logical simplicity and clear structural rules enable them to reveal the major driving factors behind evaluation procedures.

In this context, decision trees are regarded as an optimal methodology for reducing high-dimensional data to its essential variables.

Another advantage of this model is its reduced computational complexity when handling large volumes of observations and variables. For this reason, CARTs are frequently used as primary tools for combining different estimation techniques.
While decision trees are generally easy to interpret and build, they have notable drawbacks, including a tendency to overfit and a high dependency on the characteristics of the training dataset.

An evolution of the CART models is the Random Forest technique. A Random Forest is an ensemble classifier composed of multiple decision trees, each functioning as an independent, identically distributed random vector. These trees are built to represent the most prevalent input classes in the datasets analyzed [37].

Each tree within a Random Forest is built and trained on a random subset of the data, drawn from the overall training set. In this approach, individual trees do not analyze the entire dataset; instead, they focus on the optimal attributes within randomly selected subsets.

Randomization is a learning method for classification designed to diversify datasets and minimize the number of underlying relationships between them. The result of Random Forest (RF) modeling can be presented either as an average of predictions from all the individual trees or as a grouping based on the majority vote among the trees (clustering).
Literature reviews on RF techniques indicate that these methods produce highly consistent probabilistic estimates (see [11], [31], [32], [36]). They have been tested across various types of data and frequently compared to classical parametric methods [24]. However, compared to single decision trees, RF models are less intuitive and harder to explain, and their output parameters can be challenging to calibrate over time.

Finally, many machine learning techniques are based on neural networks. The term "neural network" originates from the mathematical modeling of what was once thought to be the primary functioning mechanism of the animal brain [25].

A neural network is a multistage ([1], [41]), non-linear regression model [29], consisting of layers of neurons connected by synaptic links. These connections enable the network to relate input variables to output variables. A neuron in this context can be viewed as a mathematical function (or primitive function) of the explanatory variables ([12], [14], [21]). Neural networks determine the coefficients of the network functions - such as the sigmoid function ([25], [20], [16]) - that connect neurons and express the relationships between input and output variables. This is achieved by minimizing an objective function ([15], [13]), which is typically the average squared deviation between the actual output values and the predicted values ([37], [7]).

While neural networks can capture non-linear and non-monotone relationships between the Probability of Default (PD) and the explanatory variables, they have significant drawbacks. These include the arbitrariness in choosing numerous parameters and, most notably, the difficulty in interpreting the results, often referred to as "black boxes."

2.2. Unsupervised learning

Unlike supervised learning, unsupervised learning involves working with unlabeled data. These techniques are particularly useful for exploring the structure of the data and extracting meaningful information. Notably, unsupervised learning methods do not rely on pre-established objective functions, which distinguishes them from supervised learning approaches.

This study implements k-means algorithms to cluster unlabeled datasets [33]. K-means is a widely used clustering technique that partitions a set of observations (N) into K clusters. Observations are grouped based on their proximity to the nearest cluster mean, with the goal of minimizing the total intra-cluster variation.

Specifically, given a set of observations $(x_1, x_2, \dots, x_N)$ representing real vectors with DDD dimensions, k-means algorithms aim to partition these observations into K subsets $S = \{S_1, S_2, \dots, S_K\}$. The objective is to minimize the Within-Cluster Sum of Squares (WCSS), a measure of the total variation within each cluster.

This function is represented as follows

$$\min_S \sum_{i=1}^{K} \sum_{x \in S_i} \|x - \mu_i\|^2 = \min_S \sum_{i=1}^{K} |S_i| \operatorname{Var} S_i \tag{6}$$

μ_i represents the mean of the values in cluster S_i.The algorithm uses iterative error functions to initialize K cluster means $m_1^{(1)}, \ldots, m_k^{(1)}$. The process of assigning data values to clusters involves the following stages:

1. **Assigning data values (Assignment step):** Each observation (x) is assigned to the cluster whose mean (m) is closest, as determined by the Euclidean distance. Mathematically, this involves partitioning the observations using a Voronoi diagram generated by the cluster means.

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \right\} \forall j, i \tag{7}$$

For each observation x_p assign it to exactly one cluster $S^{(t)}$

2. **Updating Cluster Centroids (Update Step):** Recalculate the means of the clusters by finding the new centroids, which are the averages of the sample observations assigned to each cluster.

$$m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \in S_i^{(t)}} x_j \tag{8}$$

The algorithm converges when the configuration of cluster centroids no longer changes. Several variations of the original k-means algorithm have been developed to improve performance, including K-medians clustering, K-means++, and soft k-means (Fuzzy C-means). These approaches are among the most popular enhancements.

3. Application of ML to Estimation of the Probability of Default

The objective of our study is to apply the multivariate techniques of supervised machine learning (ML) discussed earlier to quantify the creditworthiness (Probability of Default - PD) of customers when granting new credit. To achieve this, we utilize cluster analysis to create a discrete representation (rating scale) that reflects the creditworthiness of each customer.

3.1. Data Sample description

This research was conducted on a dataset comprising over 800,000 retail customers from a panel of European banks under ECB supervision, utilizing 10 years of historical data collected from January 2012 to December 2022. Each customer's data was analyzed based on its informative and predictive attributes to create the following input datasets:

- *Credit Bureau:* Information related to the following product categories was analyzed:
  - *Cards:* Card usage, residual amount, number of active contracts, number of cards in possession;
  - *Non-installment Products:* Number of active contracts, agreed amount, amount used, overrun amount;
  - *Installment products:* Monthly installment amount, residual installment amount, overdue and unpaid installment amount, total number of active contracts.

- *Overall data (bank and system):* Number of creditors in the system, number of active contracts at the institution, number of active contracts in the system, total overdue unpaid amounts, Credit Bureau score, presence of non-performing loans in the system, presence of protests in the system, presence of prejudicial records in the system, presence of negative notes;
- *Product information:* Used data related to monthly installment amounts, installment-to-income ratio, amount requested, value of the property, degree of mortgage, type of property (in the case of a mortgage), loan duration, and purpose of the loan;
- *Socio-demographic information:* Nationality, geographical area of residence, years of residence at the current address, banking seniority, workplace seniority, age, type of employment agreement held by the applicant (e.g., fixed-term, permanent), counterparty type (individual, joint account holder, guarantor), profession, sector of economic activity, housing status (e.g., owner, renter), marital status, credit card holder status (additional), net annual salary, net annual income (including other earnings), and possession of real estate.

3.2. Input vector construction

The target variable in this study reflects the pattern of default conversion for loans disbursed on a monthly basis, capturing defaults at 90 days past due, unlikely to pay statuses, and non-performing loans, over the period from January 2012 to December 2022.

Given the main objective of this study — to identify ML methodologies that best fit our case study — the aforementioned input datasets have been used to estimate the probability of default. The following representations highlight the relationships between individual variables and the default rate across the entire sample analyzed. Please note that the graphs below are based on the variables most relevant to the classes/subsets examined.

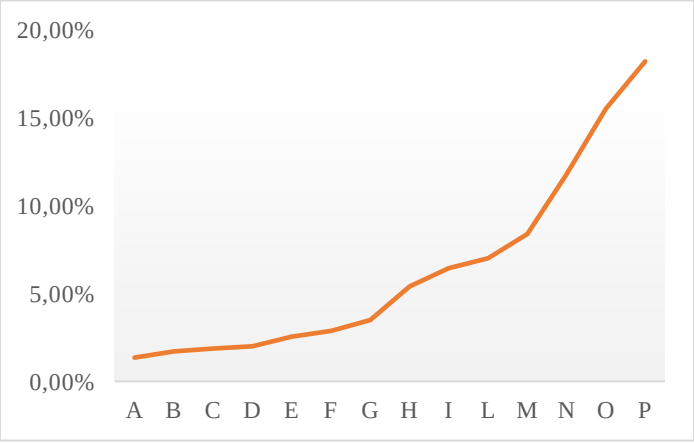


Figure 2 – CBScore and Default rate distribution

In Fig.2., the Credit Bureau score demonstrates a positive trend in relation to the default rate: as the score or rating class increases, the observed riskiness also increases.

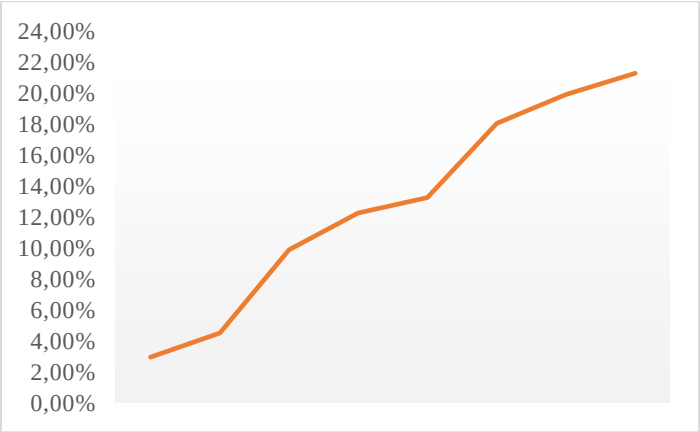


Figure 3 – Overdue accounts and Default rate trends

Overdue accounts also appear to follow a similar trend.

Figure 4 – Installment amount on salary rate and Default rate trend

To select the variables for the multivariate analysis using the aforementioned supervised ML techniques, input variables were first subjected to standard normalization procedures to identify missing and anomalous values. Variables with more than 15% missing values were excluded from the process.

Individual variables were selected by combining the graphical analysis (see Figs. 2, 3, and 4) with logistic regression techniques and predefined constraints set for each variable analyzed:

- Consistency of the sign of the coefficient with the expected economic meaning for both the variable and the default rate;
- Statistical significance of the coefficient (p-value less than 5%);
- Predictive power of each variable measured through an Accuracy Ratio greater than 10%.

The identified variables were then subjected to correlation analysis to exclude those with a correlation coefficient greater than |0.5|. The table below shows the variables used for constructing the multivariate model.

Table 1 – Intermediate short list of variables

Model inputs	
Sociodemographical	***Credit Bureau***
Nationality	**Cards**

Geographic area of residence	Residual credit card plafond amount
Years of residence at the current address	Number of active contracts
Bank seniority	Number of owned credit cards
Work seniority	**Non - installment products**
Age	Number of active contracts
Numbers of guarantors for each contract	Granted loan amount
Total number of contract members (joint holders and guarantors)	Loan overdue amount
Type of work (time, permanent, etc.)	Loan used amount
Type of NDG (physical person, joint holders, guarantor, etc.)	**Installment products**
Profession	Monthly installment amount
Sector of economic activity	Amount of residual loan
Housing situation (property, rent etc.)	Monthly installment overdue unpaid
Marital status	Number of active contracts
Number of credit cards owned	**Credit system data**
Net annual income from work	Number of custodian banks
Net annual overall income	Number of active contracts with the current bank
Real estate ownership	Number of active contracts with system banks
Product information	Total amount of overdue loans
Monthly installment amount	Credit Bureau score
Installment amount on income	Non performing status within system
Loan amount required	Presence of system protests
Real Estate market value	Presence prejudicial to the system
Loan duration	Presence of system negative notes
Mortgage degree	
Loan purpose	
Type of Real estate	

3.3. Multivariate ML models: main results

This section describes the results obtained from applying ML algorithms to build default prediction models, using an input sample of assessments conducted between January 2012 and December 2022. The sample includes retail customers from a panel of European banks supervised by the ECB.

As discussed in Section 2.1, building models to predict default probabilities is a common problem in supervised learning, one of the most widely used ML techniques. In supervised learning frameworks, a "learner" is represented by pairs of input and output values derived from historical data, where input data is used to determine the corresponding output values.

Input data are typically represented as vectors and may include continuous and/or discrete values, with or without missing data, depending on the learning algorithm used. Supervised learning involves a "regression" problem when the output is continuous, and a "classification" problem when the output is discrete. The goal

of the "learner" is to identify a function that maps input vectors to output values. One approach to this mapping is to build a record of all previous input/output pairs.
While this approach is effective for mapping input/output pairs within a training dataset, it may not reliably predict output values for input data that differ from those in the training dataset or if the training dataset contains noise.

Supervised learning aims to construct a function that can accurately map input/output pairs for out-of-sample datasets, even those not present in the original input sample.

In our model, the output is a continuous variable ranging between 0 and 1. This value can be interpreted, under certain conditions, as an estimate of the probability of default occurring within 12 months from the date the agreement is signed, based on specific input variables.

Definition of the multivariate model

For constructing the forecast model, we built and compared three ML algorithms:

- **A three-layer neural network**: This model uses a backpropagation algorithm and is fully connected and feed-forward;
- **A CART model**: This model employs the Gini index as the objective function for tree pruning;
- **A Random Forest model**: This ensemble method combines multiple decision trees to improve prediction accuracy.

These methodologies were selected due to their prevalence and effectiveness in the literature. The results obtained were evaluated based on criteria such as overall performance and predictive capabilities of the models. The output values from these models were then compared with those obtained from traditional supervised ML techniques, including logistic regression.

The table below summarizes the performance metrics obtained from the different models:

Table 2 – Performances of different ML models

ML Technique	Accuracy Ratio	CCR
Neural Network	71%	86%
Random Forest	68%	81%
Classification tree	66%	79%
Logistic Regression	66%	77%

By comparing the performance of the models using Accuracy Ratio (AR) and Correct Classification Rate (CCR), the neural network model stands out with the highest performance. It achieved an AR of 71% and a CCR of 86%. In comparison, the Random Forests model showed slightly lower performance, with AR values of 68% and 66% and CCR values of 81% and 79%, respectively. Classification trees, while performing similarly to Random Forests, had AR values of 66% and CCR values of 79%.

However, this study indicates that neural networks yield more accurate results when the detection is based on standard indicators. The neural network algorithm shows its strength when additional information or attributes are incorporated, especially in cases involving less structured data.

In terms of methodology, classification trees have proven effective for mapping input and output variables, as well as identifying their ranking values, which are essential for calibrating Probability of Default (PD) estimates.

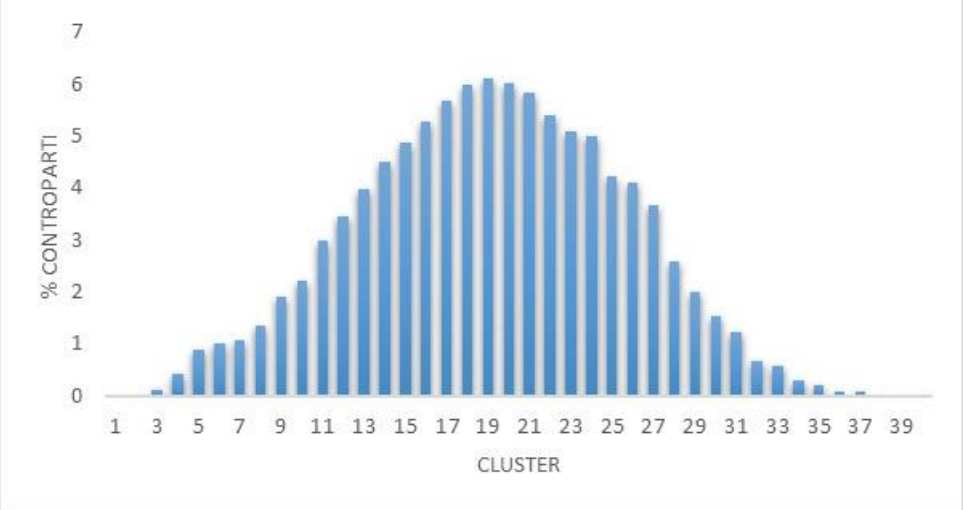
3.4. Model calibration

The final step in the Probability of Default (PD) estimation model is the calibration of all scores to convert them into PD values, achieved through the creation of rating classes. This process involved calibrating the scores obtained from the multivariate model—identified using classification tree techniques—through a Bayesian-type method. The output scores were then anchored to a long-term central tendency. Ultimately, we established the optimal rating scale in compliance with regulatory requirements.

The unsupervised ML approach, specifically k-means clustering techniques, was employed to create rating scales with the following parameters:

- **Initial Set of Parameters:** An initial set of 40 clusters was identified, with the final scale constrained to a maximum of 11 rating classes;
- **Cluster Splitting Criteria:** Clusters were split based on a concentration threshold greater than 15%.

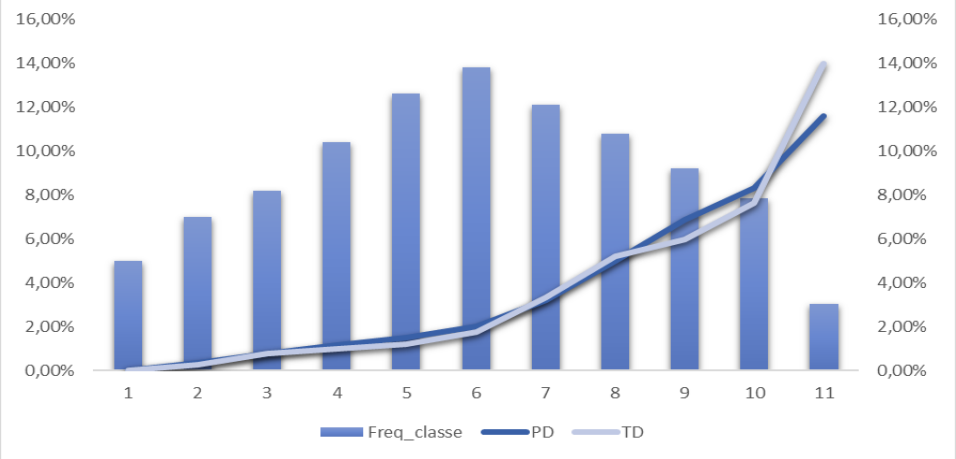
Figure 5 – Cluster population distribution



The clusters identified through k-means clustering were analyzed using an iterative combinatorial algorithm to evaluate:

- **Rating Scale Shape:** Assessing the symmetry and shape of the bell curve to ensure a well-structured rating scale;
- **Class Concentration:** Verifying the presence of appropriate concentration within each class, avoiding excessively high or low concentrations;
- **Calibration Tests:** Performing binomial and chi-square tests to validate the calibration of the rating scale;
- **Monotonicity of PD Trend:** Ensuring the default probability (PD) trend and default rate exhibit a consistent and logical monotonic relationship.

Figure 6 – Final Rating classes after machine learning application



4. Conclusions

This paper has demonstrated the application of various machine learning (ML) methodologies for estimating the Probability of Default (PD). We compared several commonly used techniques with advanced methods, including Neural Networks, Classification Trees, and Random Forests.

Our analysis was based on a substantial dataset, encompassing over 800,000 retail customers from European banks supervised by the ECB, with data spanning from January 2012 to December 2022.

Key findings include:

- **Classification Trees**: These techniques, while showing slightly lower predictive performance compared to Random Forests and Neural Networks, offer interpretability advantages. They were utilized to create rating classes via k-means clustering in the final stage of our analysis.
- **Neural Networks**: These models showed limited effectiveness when applied to standard indicators. However, they hold potential for significant improvements when integrating new or less structured data.
- **Random Forests**: This technique demonstrated robust performance but was less interpretable compared to Classification Trees.

Looking ahead, further research will focus on applying additional ML techniques to datasets from corporate portfolios, aiming to enhance model accuracy and applicability in diverse contexts.

5. Bibliography

[1] ALTMAN, E., BARBOZA, F., KIMURA, H. (2017). *ML models and bankruptcy prediction*. Expert Systems with Applications 83: 405-417

[2] ADDO P., GUEGAN D., HASSANI B. (2018). *Credit Risk Analysis using Machine and Deep Learning*. Documents de travail du Centre d' Economie de la Sorbonne

[3] ANTUNES F., RIBEIROA B., PEREIRA F. (2017). *Probabilistic modeling and visualization for bankruptcy prediction.* Applied Soft Computing 60: 831–843.

[4] ARMINGER G., KRUPPA J., SCHWARZ A., ZIEGLER, A. (2013). *Consumer credit risk: Individual probability estimates using ML.* Expert Systems with Applications.

[5] AUGUSTIN L., BOULESTEIX A., T STROBL C. (2007). *Unbiased split selection for classification trees based on the Gini index.* Computational Statistics & Data Analysis.
[6] Banca d'Italia (2022). *Artificial intelligence in credit scoring: an analysis of some experiences in the Italian financial system.* Occasional Paper – Questioni di Economia e Finanza n. 721
[7] BHAVSAR H., PANCHAL M. H. (2012*). A review on support vector machine for data classification.* International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 1(10), pp-185.
[8] BONINI S., CAIVANO G. (2013). *Survival analysis approach in Basel2 Credit Risk Management: modelling Danger Rates in Loss Given Default parameter.* Journal of Credit Risk, 9 (1).
[9] BONINI S., CAIVANO G. (2016). *Estimating loss-given default through advanced credibility theory.* The European Journal of Finance 22 (13).
[10] BONINI S., CAIVANO G. (2021). *Artificial Intelligence: the Application of Machine Learning and Predictive Analytics in Credit Risk*. Risk Management Magazine 16 (1).
[11] BREIMAN L. (2001). *Random forests.* ML 45 (1): 5-32.
[12] BUTARU F., QUINGQUING C., BRIAN C., SANMAY D., ANDREW W. L., and AKTARE S. (2016). *Risk and risk management in the credit card industry.* Journal of Banking and Finance 72: 218–39.
[13] CHAUDHURI A., DE K. (2011). *Fuzzy support vector machine for bankruptcy prediction.* Applied Soft Computing 11: 2472–2486.
[14] CHEN C., SCHWENDER H., KEITH J., NUNKESSER R., MENGERSEN K., MACROSSAN P. (2011). *Methods for identifying SNP interactions: a review on variations of Logic Regression, Random Forest and Bayesian logistic regression.* IEEE/ACM Trans Computer Biol Bioinform 8: 1580–1591.
[15] CHEN S., HARDLE W., MORO R.A. (2006). *Estimation of Default Probabilities with Support Vector Machines.* SFB 649, Economic Risk, Berlin, SFB 649 discussion paper, No. 2006-077.
[16] CHEN M.L., TSAI C.F. (2010). *Credit rating by hybrid ML techniques.* Applied Soft Computing
[17] CHEN Y., CALABRESE, R., BARRAGAN, B.M. (2024). *Interpretable machine learning for imbalanced credit scoring datasets.* European Journal of Operational Research, 312 (1)
[18] DWYER D.W., STEIN R.M. (2006). *Inferring the default rate in a population by comparing two incomplete default databases.* Journal of Banking & Finance 30: 797–810.
[19] European Commission (2020). On artificial intelligence - A European approach to excellence and trust. White Paper
[20] FONSECA P., LOPES H. (2017). *Calibration of ML Classifiers for Probability of Default Modelling.* James Finance, Crowd Process Inc.
[21] GHODSELAHI A. (2011). *A hybrid support vector machine ensemble model for credit scoring.* International Journal of Computer Applications, 17(5), 1-5.
[22] HARDLE W., MORO R.A., SCHAFER D. (2005). *Predicting Bankruptcy with Support Vector Machines.* SFB 649, Economic Risk, Berlin, SFB 649 discussion paper, No. 2005-009.
[23] HARDLE W., MORO R.A., HOFFMANN L. (2010). *Learning Machines Supporting Bankruptcy Prediction*. SFB 649, Economic Risk, Berlin, SFB 649 discussion paper, No. 2010-032
[24] HERNÁNDEZ-LOBATO D., SHARMANSKA V., KERSTING K., LAMPERT C.H., QUADRIANTO N. (2014). *Mind the nuisance: Gaussian process classification using privileged noise.* Advances in Neural Information Processing Systems 27: 837–845.
[25] HOSMER D. W., LEMESHOW S. (2000). *Applied logistic regression.* 2nd ed. New York: Wiley.
[26] HUANG Z., CHEN H., HSU C.J., CHEN W.B., WU S. (2004). *Credit rating analysis with support vector machines and neural networks: a market comparative study.* Decision Support Systems 37: 543-558.
[27] Hurlin, C., Perignon, C., & Saurin, S. (2022). The fairness of credit rating models. arXiv Preprint.
[28] KABIR M.J., KANG B.H., LIU Y., WASINGER R., ZHAO Z., XU S. (2015). *Investigation and improvement of multi-layer perception neural networks for credit scoring.* Expert Systems with Applications.
[29] KHANDANI A.E., KIM J., LO A.W. (2010). *Consumer credit-risk models via machine-learning algorithms.* Journal of Banking & Finance.
[30] KHASHMAN A. (2010). *Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes.* Expert Systems with Applications.

[31] 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.
[32] LIAW A., WIENER M. (2002). *Classification and regression by random forest.* R News 2 (3): 18-22.
[33] LOH W.Y. (2011). *Classification and regression trees.* WIREs Data Mining Knowledge Discovering 1: 14–23.
[34] MIN J. H., LEE Y. (2005). *Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters.* Expert Systems with Applications, Vol. 28, Issue 4, pp.603-614.
[35] STEINBACH M., TAN P.N. (2009). kNN: *k-nearest neighbors.* In: Wu, X., Kumar, V. (eds.). The top ten algorithms in data mining. Chapman & Hall/CRC: 151–162.
[36] STEINBERG D. (2009). *CART: classification and regression trees*. In: Wu, X., Kumar, V. (eds.). The top ten algorithms in data mining. Chapman & Hall/CRC: 180–201.
[37] STROBL C., BOULESTEIX A., AUGUSTIN L. T. (2007). *Unbiased split selection for classification trees based on the Gini index.* Computational Statistics & Data Analysis 52 (1): 483-501.
[38] Teodorescu, M., & Yao, X. (2021). *Machine learning fairness is computationally difficult and algorithmically unsatisfactorily solved.* IEEE Publication 2021
[39] Tripathi, D., Shukla, A. K., Reddy, R., Bopche, G. S., & Chandramohan, D. (2022). *Credit scoring models using ensemble learning and classification approaches: A comprehensive survey.* Wireless Personal Communications, 123 (1), 785–812
[40] TSAI C. F., CHEN M. L. (2010). *Credit rating by hybrid ML techniques.* Applied soft computing, 10(2), 374-380
[41] Wang, J., Xu, J., Cheng, Q., Kumar, R. (2024). *Research on finance Credit Risk Quantification Model Based on Machine Learning Algorithm*. Academic Journal of Science and Technology, 10 (1)
[42] Wang, J., Gambacorta, L., Huang, Y., Qui, H. (2024). *How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm.* Journal of Financial Stability, 73
[43] WRIGHT M. N., ZIEGLER A. (2017). *Ranger: A fast implementation of random forests for high dimensional data in C++ and R.* Journal of Statistical Software 77:1-17