RISK MANAGEMENT MAGAZINE

Vol. 18, Issue 3
September – December 2023

EXCERPT

https://www.aifirm.it/rivista/progetto-editoriale/

Data Analytics for Credit Risk Models in Retail Banking: a new era for the banking system

Adamaria Perrotta, Andrea Monaco, Georgios Bliatsios

Data Analytics for Credit Risk Models in Retail Banking: a new era for the banking system

Adamaria Perrotta (UCD University College Dublin), Andrea Monaco (UCD University College Dublin), Georgios Bliatsios (UCD University College Dublin)

Corresponding author: Adamaria Perrottta (adamaria.perrotta@ucd.ie)

Article submitted to double-blind peer review, received on 21st November 2023 and accepted on 13th December 2023

Abstract

Given the nature of the lending industry and its importance for global economic stability, financial institutions have always been keen on estimating the risk profile of their clients. For this reason, in the last few years several sophisticated techniques for modelling credit risk have been developed and implemented. After the financial crisis of 2007-2008, credit risk management has been further expanded and has acquired significant regulatory importance. Specifically, Basel II and III Accords have strengthened the conditions that banks must fulfil to develop their own internal models for estimating the regulatory capital and expected losses. After motivating the importance of credit risk modelling in the banking sector, in this contribution we perform a review of the traditional statistical methods used for credit risk management. Then we focus on more recent techniques based on Machine Learning techniques, and we critically compare tradition and innovation in credit risk modelling. Finally, we present a case study addressing the main steps to practically develop and validate a Probability of Default model for risk prediction via Machine Learning Techniques.

Keywords: Credit Risk Management, Risk Prediction, Machine Learning, Loan Defaults.

1. Introduction

Estimating the risk profile of clients is one of the most important activities in the banking sector. The number and variety of methodological approaches to model credit risk in the banking system has strongly grown in the last 20 years, due to the increasing relevance that risk has for firms' business models. Indeed, a strategic, central role has been given to credit risk for the banking system by the Regulator. Therefore, large internal teams of specialists have been tasked with the development and validation of sophisticated credit risk models for default prediction subject to well established regulatory frameworks. Such frameworks are being created by Central Banks and other banking supervisory entities and are usually the product of forums where these authorities meet and exchange ideas. An example of such a forum is the *Basel Committee on Banking Supervision (BCBS)* which was formed in 1974.

In 1988, the BCBS issued the first Basel Capital Accord (Basel I) which had the main objective of ensuring that all financial institutions operating in an international environment hold enough cash reserves. In 1999, the same entity published several proposals for changes to the current regulations to reinforce and level out the risk management process. Such proposals heavily contributed to the New Capital Agreements, known as Basel II and Basel III accords (see (Bank for International Settlements, 2005a) and (Bank for International Settlements, 2011)). In (Bank for International Settlements, 2005a), the Committee created a robust risk management framework by introducing the two main approaches currently in us to determine the minimum regulatory capital requirements for credit risk.

The first approach is known as *Standardized Approach,* and it allows banks to use prescribed estimates for the key risk factors involved in the calculation of the Risk Weighted Assets (RWA). The second and more common approach is known as *Internal Rating Based (IRB) Approach* and it permits banks to use internally developed models for estimating the RWA. Of course, the IRB is subject to regulatory approval to be put in use. The IRB approach can be further divided into two sub-approaches; one referred to as *Foundation – IRB Approach (FIRB)* and the other one as *Advanced - IRB Approach (AIRB)*. Within the AIRB approach, banks can use their own models to estimate the likelihood that a credit obligation will not be met while the remaining risk factors are prescribed by the Regulator. In view of the AIRB approach, all risk factors can be estimated using internal models. Credit institutions which are allowed to apply the FIRB or AIRB usually combine quantitative and qualitative techniques in their estimation method to get the best risk predictions. The key credit risk factors for which a significant number of models is developed on a regular basis are the Probability of Default (PD), Exposure at Default (EAD) and Loss Given Default (LGD); they contribute to the estimate of the RWA (Bank for International Settlements, 2005b) and the Economic Capital (Bank for International Settlements, 2009). Furthermore, under the *IFRS 9 Regulatory Framework* (Bank for International Settlements, 2017) the same risk factors are computed via different estimation methods and used in the calculation of *Expected Credit Loss* (ECL) (EBA/GL/2017/06, 2017) and provision management.

In addition to the new Regulation, another driver for the growth of credit risk modelling is the increased capability of exploiting, collecting, and processing of large dataset of portfolio debtors. Such information allows us to detect the underlying credit risk dynamics which have been completely ignored until now. Obviously, since credit risk models are predominantly data driven, the underlying training dataset should be of the best possible quality to ensure the robustness of results and reliability of the estimate (European Central Bank, 2019). Moreover, since the availability and the distribution of the data can heavily influence the design of a model, it is important to ensure that a set of rules and policies regarding the collection, structure, cleansing, transformation, storage, and assessment of the available dataset is in place prior to commencing model development. In particular, these rules can offer guidance with respect to how to handle missing values, when and if exclusions should be applied, how to treat outliers and several other controls depending on the nature of the data.

In literature, a big number of classification algorithms for borrowers' creditworthiness assessment have been proposed, jointly with some comparison studies (see as an example (Baesen, et al., 2003) and (Lessmann, Baesens, Seow, & Thomas, 2015)). Assuming that a sufficiently large data set is in place, based on the nature of the risk factor and the regulation governing the model's design, a wide range of statistical tools is available for developing credit risk models. Such tools include, but are not limited to, Logistic Regression, Generalized Linear Classifiers, Random Forests, Time Series Analysis, Bayesian Inference, Artificial Neural Networks, and hybrid methods which combine various continuous and/or discrete probability distributions.

Once the risk model has been developed, its predictive ability and discriminatory power must be assessed and monitored on a regular basis. This can be done through various validation techniques which, depending on the nature of the risk factors and the development approach, might include the use of confidence intervals, root mean square error, confusion matrices, accuracy ratio, receiver operating

characteristic curve (Irwin & Irwin, 2012), Kendall's Tau and other statistical indicators. Additionally, the model performance can also be assessed by benchmarking its outputs against the ones produced by alternative models and/or against historical observations.
The purpose of this contribution is to critically review the most used model development and validation procedures for credit risk in retail banking. In Section 2 we focus on the impact of Regulator on Credit Risk Methodologies and briefly motivate the introduction of ML approaches for regulatory capital estimation. In Section 3 we review some of the traditional statistical methods to PD and LGD estimates. Then in Section 4 we focus on more recent studies based on Artificial Intelligence and Machine Learning techniques. We conclude the paper with a case study, presented in Section 5, showing some of the basic and practical steps performed in the financial industry to develop and validate a PD model via Machine Learning Techniques. Conclusions will be discussed in Section 6.

2. The impact of Regulator on Credit Risk Methodologies

As pointed out in the Introduction, within Basel's Accords the Regulator defined a general framework of assumptions and models to estimate credit risk.
The first set of requirements to estimate credit risk is the one known as Basel I Accord (1998): in such agreement, the Regulator stated the definition of *minimum capital requirement* in terms of *Risk weighted assets*. In Basel I framework, the regulatory capital was computed estimating unexpected portfolio losses, while credit risk evaluation was modelled through the combined effect of a systematic and an idiosyncratic default rate risk factor. The Regulator also required financial institutions to customize the credit risk model with respect to the risk profile of the portfolio's constituents, defining five possible categories: corporate, sovereign, bank, retail, equity. Finally, the Basel Committee introduced in credit risk modelling the so-called through cycle PD: the average default rate performance for a particular customer over an economic cycle.
Basel I approach has been heavily criticized for being insufficiently granular, so the Basel II Accord (2004) provided a more refined methodology proposing two possible approaches:

- *the Standard Approach*: it is based on a simple categorization of debtors, without considering their actual credit risks. This approach relies on external credit ratings.
- *the Internal Ratings-Based (IRB) approach*: within this approach, financial institutions are allowed to use internal models to compute the regulatory capital requirement for credit risk.

Basel II accords set the reference parameters (i.e., the parameters of the IRB approach) that any credit risk model must estimate to compute the minimum capital requirement. Thus, each institution had to independently define the levels of the PD, LGD and EAD, starting from its own score models. Basel II also reinforced the validation process setting up new rules:

- Banks must have a robust system to validate the accuracy and consistency of rating systems, processes, and the estimation of all relevant risk components.
- A bank must demonstrate to the Regulator that the internal validation process enables it to assess the performance of internal rating and risk estimation systems consistently and meaningfully.

Basel I and II methodological framework constitute a constraint that has strongly influenced the development of credit risk modelling since any possible change/improvement must be in line with the Regulation to be effectively implemented. The Basel framework mainly relies on the estimate of three key risk factors: PD, LGD and EAD since they constitute the main ingredients to compute the minimum capital requirement.
The traditional methodological approach to credit risk is based on the explicit modelling of these quantities. However, with increased computing capabilities and access to larger amounts of data, the current literature describes many improvements upon simpler traditional risk models, which are limited in their ability to incorporate big data. The availability of large data sources and a higher level of info details allows both to increase the accuracy of traditional models to better fulfil regulators' requirements and to apply Machine Learning techniques to analyze and forecast credit risk. For this reason, there is increased interest in the development of machine learning models that can capture nonlinear relationships between variables of a dataset (Sadhwani, Giesecke, & Sirignano, 2021). However, ML models have their drawbacks. They can be computationally expensive and difficult to interpret; moreover, they can bring to ethical tradeoff (Lee & Floridi, 2021). In addition to that, ML approaches are distant from those covered by the current regulatory framework. This is the reason why the switch from traditional approaches to the full use of ML algorithms to risk decision-making is still far from the implementation.

3. Traditional Approaches to Credit Risk Analysis

3.1 Traditional Approaches to PD Models

The large variety of PD models used by banks to estimate credit risk can be classified into two main categories: *Structural Models* and *Reduced Form Models* (Duffie & Singleton, 2012). *Structural Models* estimate the probability of default starting from assumptions about the value of its assets and liabilities. The basic idea is that a company defaults if the value of its assets is less than the debt of the company. Conversely, the *Reduced Form Models* assume default as an external cause regardless of the value of the assets/liabilities of the company. In this section we will present an overview of the principal Structural and Reduced Form Models used in credit risk management.

3.2 Structural Models

The class of structural models identifies the family of models that describes the dynamics of credit event with the structure of a company's debt: the event of default occurs when the market value of a company's assets $A(t)$ falls below a threshold set by the nominal value of the debt $D(t)$. According to this model, the default event will depend on the dynamics of the assets $A(t)$ and the dynamics of the default barrier level represented by debt $D(t)$. The default event occurs at the time τ when $A(\tau) < D(\tau)$ and in this case

the equity value is equal to zero. Therefore, within the structural models setting the default event is translated into a barrier event estimation problem: we should estimate the probability of the default event in the time horizon [0, T], where T is the maturity of the debt, that is $P[A(\tau) < D(\tau) \mid \tau \in [0,T]]$.
The first example of structural model is the well-known Merton model (Merton, 1974). The basic assumptions of the Merton model are:

- *A(t)* follows a log-normal stochastic distribution.
- The default is evaluated on a fixed time horizon T.
- *A(t)* has both an equity component *E(t)* and a debt component *D(t)*: *A(t) = E(t) + D(t)*.
- The debt *D(t)* is represented by a zero-coupon bond with maturity T.
- the default can occur only at maturity T of the debt, and it happens if *A(T) <D(T).*

The Merton model can be easily applied if we agree on its strong hypotheses. This model provides in fact an explicit formula for risky bonds (Merton, 1974). Such formula can be used to estimate the PD of the firm as well the credit spread structure of the risky bond. However, the model's "oversimplified" hypotheses are responsible for some limitations in its use. Firstly, the Merton model simplifies the liabilities' structure assuming the debt as a single zero-coupon bond. This hypothesis allows the model to assess the occurrence of the default only at maturity T. Then, the "oversimplified" dynamics for the asset liability is responsible for a zero value of credit spread at very short maturities. Such a prediction is not confirmed by market observations. In conclusion, it is more than clear that some improvements to the Merton model are needed to get a more realistic estimate of PD.
Under this light, there are different models that extend the Merton model. Each of these models improves the original Merton framework by removing some assumptions; in the following we report the most representative ones.

- **The Black-Cox model** (Black & Cox, 1976). This model extends the Merton model by assuming that the default can occur before the maturity of the liability T. According to this model, the default barrier is a function of time *H (t)* of exponential type. This model belongs to the larger family of *first hitting time models*.
- **The KMV model** (Bharath & Shumway, 2006). The KMV model improves the Merton model exploiting historical data on defaults. This model is based on historical estimation of the probability of default, allowing to simulate the default over different time horizons. A big advantage of the KMV approach is the high accuracy of PD estimates in case of extreme events simulation. This great performance is due to the historical calibration of the model.
- **The Credit Grade Model** (Finger, et al., 2002). This model extends the structural models introducing a stochastic dynamic to the debt/barrier. The model in fact attributes a log-normal stochastic dynamics to both the asset *A(t)* and the recovery rate *RR(t)*[1], resulting in a stochastic dynamic also for the debt *D(t).* The model also allows us to estimate the survival probabilities in closed form therefore one can calibrate the model on the level of market spreads to market data. This model, as well as the KMV, has a level of accuracy in the simulation of extreme events much more accurate than the Merton model.

Despite such attempts to potentiate the Merton Model performances, the main drawbacks of this class of model are still present. Those models require to estimate the firm's asset value, which is non-observable. Moreover, structural-form models cannot incorporate credit-rating changes that occur quite frequently for default-risky corporate debt. These limitations can be partially removed by adopting a *Reduced Form* approach to model PD.

3.3 Reduced Form Models

The class of Reduced Form models, also known as *intensity models*, describes the process of default through a minimum set of hypotheses. The Reduced Form models are particularly suited to model credit spreads and their basic formulation makes them easy to be calibrated on corporate bond data or Credit Default Swaps (CDS)[2]. Given a time interval [t, t + Δt) and a default time τ, the probability that time τ falls within of the interval [t, t + Δt) is equal to λ(t)Δt, where λ(t) is generally named *hazard* or *intensity rate.* Since Reduced Form models assume that PD ~ λ(t)Δt, then the default event follows a Poisson process. Moreover, since the probability of default is completely specified by the function λ(t)Δt, the Reduced Form models differ between each other in relation to different hypotheses made on λ(t). We can consider three sub-classes of Reduced Form models:

- **The Time Homogeneous models**: the intensity of default λ(t) is deterministic and constant over time. Therefore, the dynamics of credit spreads results constant and deterministic. Within this sub-class it is immediate to infer the survival probability within a fixed horizon from the level of market spreads CDS. Its main limitation is that Time Homogeneous models are not able to describe the complex structure of credit spread, implicit in quoted instruments.
- **The Time Variant models**: this sub-class of Reduced Form models is based on the hypothesis of deterministic intensity but variable over time. Very often this family of models is specified in terms of cumulative intensity. This sub-class allows to reproduce, with more accuracy respect to time homogeneous models, credit spread dynamics implicit in quoted financial instruments.

[1] Recovery rate is the extent to which principal and accrued interest on defaulted debt can be recovered, expressed as a percentage of face value. The recovery rate enables an estimate to be made of LGD = 1 – RR.
[2] A credit default swap (CDS) is a financial derivative that allows an investor to "swap" or offset his or her credit risk with that of another investor.

- **The Stochastic Intensity models**: the intensity of default $\lambda(t)$ is stochastic (as an example, a CIR process). These models allow us to consider the market volatility of credit spreads. The default intensity dispersion can be obtained from credit spread options market prices or from CDS time series analysis.

The Reduced Form models family was originated by Jarrow and Turnbull (see (Jarrow & Turnbull, 1992) and (Jarrow & Turnbull, 1995)), and Duffie and Singleton (Duffie & Singleton, 1999). This family of models has been intensively studied in literature (see (Duffie & Singleton, 2012) and (Elliott, Jeanblanc, & Yor, 2000)). Some authors point out an intrinsic connection between Structural and Reduced Form models; Reduced Form models are intended as Structural Models in a different information filtrations: structural models are based on the firm's management information, while Reduced Form models are based on the information available on the market see (Duffie & Lando, 2001) and (Jarrow & Protter, 2004)).

3.4 Traditional Approaches to LGD models

Reduced Form models introduce separate explicit assumptions on the dynamic of the PD and the recovery rate RR. Despite these two variables - PD and RR - are mostly independently modelled, in literature they did not receive the same scientific attention. There are few studies on RR dynamics compared to the PD ones. The reason for this asymmetry relies on the basic assumption of credit risk models. Traditionally, credit risk models assume that RR depends on individual features like collateral or seniority and is not influenced by systematic factors; therefore, it has been considered as independent of PD.
Most recently, Basel Regulation has pointed out the relevance of this indicator, so RR modelling has attracted the interest of analysts and researchers. According to the Basel Accords, in fact, the recovery rate is used in the capital requirement formulas in a linear way. Its estimate is therefore crucial to model LGD with high accuracy. Moreover, the possibility provided by the IRB approach to set the LGD values tailored to bank's portfolios has increased research works on RR. To build LGD models some phenomenological peculiarities need to be considered. In particular, the observed historical distributions of RR are bimodal: the recovery rates are concentrated either in high values (around 70-80%) or low ones (around 20-30%). Moreover, RR values are strictly dependent on the industrial sector of obligors: tangible asset-intensive industries, especially utilities, have higher recovery rates than service sector firms, with some exceptions such as high tech and telecom. In the remaining part of this Section we will provide an overview of the different approaches to estimate LGD and of the prevalent literature available for modelling purposes.

3.4.1 LGD Estimate

There are three main different approaches to computing LGD: Market LGD, Workout LGD, Implied Market LGD.

- **The Market LGD**: This approach is based on market sentiment. Even once the default occurs, defaulted bonds and loans can still be traded on the market. Therefore, their prices include investor expectations on the entire recovery process: capital recovery of the restructuring costs and the related uncertainty.
- **The Workout LGD**: This approach is based on the historical analysis of defaulted loans to predict future values of LGD rates. All cash flows generated by the recovery process, all recoveries, as well as all costs are taken into account in the period ranging from the day of the credit event to the final recovery. These cash flows must be discounted; however, it is not obvious which discount rate needs to be applied, in case of debt restructuring actions via the issuance of risky assets such as equity or warrants.
- **The Implied Market LGD**: This approach is based on the analysis of market prices of bonds or loans before default. The spread between a loan-specific interest rate and the risk-free interest rate is equal to the expected loss thus the LGD value is deduced from the ratio between the spread and default probability.

In the case of bonds, the estimate of the LGD is less straightforward. The spread above risk-free rate is an indicator of the risk premium demanded by investors to price PD and LGD, as well as liquidity premiums.

3.4.2 LGD Modelling

Due to the complex phenomenology underlying the recovery process as well as the existence of three different approaches to compute LGD, there are several methods available in literature to model recovery dynamics. Here we briefly describe the most used ones.
In Frye's structural framework it is addressed that the value of LGD is affected by the PD level and therefore it cannot be modelled independently from it (see (Frye, 2000a) and (Frye, 2000b)). Such studies had a huge impact on regulatory practice since they suggested changes in guidelines of Basel accords. Indeed, in the model proposed by Frye in (Frye, 2000a) and (Frye, 2000b), defaults are driven by a single systematic factor: the correlation between the RR and PD, which derives from their mutual dependence on the systematic factor. However, the distribution of RR is shown to be different in high-default periods from low-default ones. The negative correlation between default rates and RR relies on the dependence of loans collateral from systematic factor. If the economy experiences a recession, then the default rate increases while the value of collateral decreases as well as the associated RR.
In a similar way to Frye's approach, the Jarrow model (Jarrow R. A., 2001) correlate both RR and PD to the state of the macroeconomy. The model introduces the liquidity premium and equity prices in a calibration procedure where RR and PD are explicitly separated.
Jokivuolle and Peura in (Jokivuolle & Peura, 2000) propose a model where the collateral drives the recovery dynamics. According to such a model, the collateral value is correlated with the PD while the credit event is triggered by the total asset value.
Tasche in (Tasche D. , 2004) introduces a single risk factor model, where LGD volatilities can be statistically estimated. Moreover, the model considers defaulted obligors. This model allows us to compute the capital charges according to Basel definition with an accurate numerical approximation, and it is one of the reasons why it is heavily implemented in banking practice.
It is important to underline that all these models consider the different characteristics of the obligor by segmenting the portfolio in terms of default period, loan-to-value ratio, customer type, credit score, etc.; these factors may be different according to the portfolio under analysis. Obviously, the setup of LGD models based on obligors' information benefits from the increases of available

information on obligors. This is the reason why in the last few years the use of ML techniques for LGD modelling, such as Regressions and Neural Networks approaches, has hugely increased. Consequently, several research studies have been conducted to address the validity of these approaches to PD and LGD modelling.

4. Machine Learning Methodologies to Credit Risk Analysis: an overview

4.1 Main Applications in Credit Risk: pros and cons

In recent years, the possibility of collecting and storing big datasets as well as the increase of computer performances provided the opportunity to increase the use of ML techniques in finance. Indeed, while conventional econometric methods fail to exploit nonlinear relations between features and hidden information deduced by unstructured data sources, ML models allow detecting peculiar patterns and dependencies from these new datasets. Thus, ML techniques promise to make data analysis for managing financial risk more efficiently. On the other side, this gain is not a "free lunch". Large financial datasets are in fact characterized by increased noise and nonlinear patterns, so they infer significant statistical challenges to modelling.

Financial institutions have developed new systems based on ML to drive expert decisions in the domain of credit risk modelling. The most common use of ML in credit risk is in credit decisions/pricing, followed by credit monitoring and collections. These techniques are mainly used by large firms due to their benefits of scale, access to data and large resources. The most active sector of ML applications is risk management, compliance, customer engagement and credit. In particular, the main applications ML techniques in credit risk management ca be synthesized as follow:

- Model validation: ML models are used as a benchmark to the standard model for capital requirements calculation.
- Data improvements: ML techniques can be used to improve data quality allowing us to clean, pre-process and analyze rich datasets.
- Variable selection: ML techniques allow to detect explanatory variables with useful predictive capacities within large datasets.
- Risk differentiation: compared to traditional PD model, ML models increase the differentiation of risks while computing the probability of default.

One of the main advantages of using ML for risk analysis is the improvement of risk differentiation. ML techniques may increase the discriminatory power of model allowing to identify risk drivers with better accuracy compared to traditional models. Therefore, ML models are used to optimize portfolio segmentation and take data-driven decisions. Moreover, ML can be used in a hybrid modality, confirming the selection of data features used in traditional models.

While ML techniques improve quantitative performances of credit models, the choice of appropriate economic theories and assumptions supporting the ML methodology could constitute a challenge. Indeed, applying IRB models requires understanding and interpreting underlying model dynamics, which is not always so straight: this could obviously prevent the use of ML techniques, despite their great performances.

Additionally, new ML models allow us to base credit score predictions on a broader range of variables than those traditionally included in the classic statistical models (Sadok, Sakka, & El Hadi El Maknouz, 2022). However, financial analysts must understand whether predictions based on big datasets could make credit available to individuals or companies that were previously considered ineligible using traditional methods. This is a very important item for driving right decisions, since a ML application could cause unintended negative consequences for the banking system, causing economic instability if not implemented with care (Eitel-Porter, 2021). Finally, it is important to investigate how the use of ML in credit risk models affects issues surrounding ethics choices, bias and discrimination in the lending market, because this could heavily contribute to housing inequality and racial disparities among minorities (Zou & Khern-am-nuai, 2022).

4.2 A literature review of ML in Credit Risk

For completeness and critical understanding of this research work, we have performed an extensive literature review to select the principal ML techniques used for credit risk assessment. In the following we present the outcomes of such selection. From now on, we will use the word "feature" to refer to the variables of a financial dataset. In literature, the word "feature" is equivalent to "factor" or "variable" or "characteristic". Features are the basic building blocks of datasets. The quality of the features in a dataset has a major impact on the quality of the insights one will gain, especially when ML algorithms are employed.

- **Linear Regression and Logistic Regression**: given a set of observations (i.e., the dependent variable or regressand) and some feature map of explanatory variables (i.e., regressor), linear regression set the best parameter configuration able to minimize residual errors of a linear objective function. Linear regression takes continuous inputs (e.g., profit margin, efficiency ratio, cash ratio, debt ratio, earnings per share, etc.) and outputs a continuous variable. Logistic regression uses a logit function to model extend linear regression to a binary dependent variable. These methodologies are used to forecast a company's financial distress (credit scoring) and then its PD level, taking continuous input related to the company profile (see (Altman, 1968), (Orgler, 1970), (West, 2000)).
- **Artificial Neural Network**: this technique is inspired by the functioning of the human brain. Individual neurons are modelled by logistic regression therefore the neural network is a multiple layer infrastructure that connects logistic regression classifiers. These algorithms are designed to recognize patterns and make predictions involving a large number of parameters. Main application of this methodology is the credit risk analysis and forecasting of borrower's credit profile dynamics (see (Hsieh, 2005), (Abdou, Pointon, & Elmasry, 2008), (Angelini, Tollo, & Roil, 2008), (Pang & Gong, 2009)).

- **Support Vector Machine**: this method allows to model nonlinear classification problems. The technique identifies hyper-plane multidimensional surfaces to separate classes of dataset. The boundary decision made by SVM is accurate compared to other techniques. However, due to the use of a kernel function it is not easy to attribute each prediction to an individual variable. SVM in credit applications is a supervised learning methodology that analyze data to perform credit risk scoring (see (Baesen, et al., 2003), (Huang, Chen, Hsu, Chen, & Wu, 2004), (Schebesch & Stecking, 2005), (Shin, Lee, & Kim, 2005)).
- **K-Dimensional Tree**: this technique allows solving classification problems using a binary tree structure (i.e., a sequence of nodes and branches), which sub-divides dataset by a hyperplane. Most of the effort to apply the method relies on the setting of the tree structure. It is quite flexible, allowing us to handle any probability distribution and non-linearity in the model. The usability of the method relies on a rich out-of-sample dataset, in fact one of the main limits of the technique is the risk of data over fitting. The method is often used to classify the credit profile of a company (Breiman L. , 2001).
- **Decision Tree**: the method allows to implement "if-and-else" questions to solve a specific classification problem. The Decision Tree has a binary tree structure like a K-D Tree, the model prediction is obtained through a sequence of nodes and branches that allow to easily interpret the decision-making logic. In a decision tree the size of the tree can grow to adapt to the complexity of the classification problem. The methodology is applied to forecast company profile conditioned to info and events using categorical and continuous inputs (Galindo & Tamayo, 2000).
- **Random Forest**: this technique can be classified as an ensemble learning method based on multiple decision trees. The decision trees are randomly generated in an iterative mode allowing to obtain a forest. The classification result is defined as the class selected by most of the trees. Random forests use the same training set to train multiple decision trees allowing to reduce the variance of the output. This method as the previous mentioned ones allows to classify credit profile of a company (Breiman L. , 2001) .
- **Boosting**: this technique can be classified as an ensemble method to reduce bias and variance in supervised learning. The method is based on the exploitation of ensemble characteristics of models to increase accuracy of forecast/description performances compared to the use of a single methodology: convert weak learners to strong ones. Boosting algorithms iteratively train single classifiers with respect to a distribution adding them to a final strong classifier. Once added the classifiers are re-weighted to increase accuracy. There are many boosting algorithms, these mainly differ by method of weighting training data (see (Amin, Islam, & Murase, 2009), (Nanni & Lumini, 2009), (Yu & Wang, 2009)).

5. Probability of Default: Elements of Model Development and Validation Within the Framework of Machine Learning

In this section we present a case study showing the development and validation process of a Probability of Default PD model via ML techniques. Before focusing on a specific ML method, we want to describe the main operative steps to perform in *any* ML method for risk management. This would facilitate the understanding of the case study and will shed light on how ML algorithms are changing the model development and validation in the banking sector.

Applying ML techniques in credit risk results in performing a multi-step process, where the different stages are:

1. Pre-Processing of the dataset: this step refers to any type of processing performed on raw data to prepare it for another data processing procedure.
2. Features Selection: in this step the analyst selects the most relevant variables to forecast the PD.
3. Data Modelling: in this step the analyst chooses the ML model to implement and the specific dataset to calibrate the model parameters.
4. Model Validation: once a model has been calibrated and implemented, the model performances must be tested in different market regimes.
5. Model Deployment: If the model has passed steps 1-4, it goes to production. At this stage potential operating risks must be managed.

As already pointed out in the previous sections, ML algorithms relies on large datasets; in particular, usually three different samples are introduced to set, validate, and finally test the model:

1. Training Sample: this dataset is used to calibrate the model parameters (step 3 above).
2. Validation Sample: this dataset is used to determine the values of the parameters of the model (step 4 above).
3. Test Sample: this dataset is used to measure the performance of the model (step 4 above).

In the rest of the paper, we present a case study showing how to implement steps 1-5 above to develop a PD model via ML techniques. To this aim, we built an artificial dataset of 1,000 home mortgages with 25 variables of account, demographic, and sociological nature (see Table 9 in the Appendix section for details).
The dataset was created in a way to mimic a real word database: we couldn't avail ourselves of a real one since banks don't provide actual obliger's information due to data protection regulations unless the research has been conducted with the financial institution.

To make the case study as authentic as possible, the variable names follow the format presented in the ECB's Loan-Level data templates[3].
Among the above-mentioned ML methods for credit risk assessment, we decided to compare a linear with a non-linear classifier: Logistic Regression (LR) vs k-nearest neighbor (k-NN). We employed Logistic Regression (LR) since on the one hand, the target variable **Default Indicator** (see Table 9) is of a binary nature and, on the other hand, LR is a well-established, simple to interpret and the most widely used approach in the banking sector. Among the non-linear classifiers, we have referred to k-NN since it is a relatively simple approach that is completely nonparametric. There are only two choices a user must make: the number of neighbors k and the distance metric to be used. We referred to the most common choice of distance metrics, i.e., the Euclidean distance. Validation techniques for measuring the predictive strength of the model are also being discussed.
We begin this section by introducing the raw data, the preprocessing of the dataset and the feature selection procedure which will lead us to the final dataset used for developing the model (Section 5.1). Then, we discuss the model fitting process and present the resulting model coefficients (Section 5.2). Although the calibration of the rating system process is out of scope from our case study given that we are developing the model on a simulated dataset, we will describe some of the basic methodological steps which are involved in this process. We conclude this section with the description of the validation procedure aiming to measure the performance of the model (Section 5.3).

5.1 Data Preprocessing and Features Selection

In this section we perform an initial data analysis and cleaning of the raw data as well as a features selection to get the final set of features for developing the PD model. In the Appendix – Table 9 - the final set features have been highlighted in bold. To address the borrower's default prediction problem, we define our target variable as Default Indicator, a numeric binary variable, describing whether the obligor is in default (1) or not (0). As already mentioned above, the dataset is the most important component of any ML model since the robustness of its output is strongly related to the quality of the inputs used for training. Therefore, several quantitative and qualitative checks are required to assess whether the dataset is fit for purpose. Thus, a preliminary step to derive useful insights regarding the distribution of the data and potential relationships among the variables is a *visual data exploration*. For example, in our dataset there is a downward trend regarding the number of defaulted borrowers with respect to the variable "Age" (see Figure 2). Therefore, one can infer that default rate and age appear to be negatively correlated and such relationship is expected to be reflected in the model, assuming that these variables are chosen for development purposes.
The next step is to control potential missing data and the presence of outliers among the variables. It is worth noting that there can be a few valid reasons for missing values for some variable and therefore when such a phenomenon is present, qualitative judgement should be employed. In case there is more than 5% of missing data for a given variable a decision whether to replace or keep the missing values needs to be made. For the features having less than 5% of missing rate, we decided to perform data imputation by substituting the nulls with the most frequent value of each feature (see (Siddiqi, 2017), (Joenssen & Bankhofer, 2012)). In case of replacement, a different choice is made in relation to the type of the variable: the mode, the mean or the median of the sample are potential candidates for missing numeric data while for nominal data types, the most frequently occurring values could be used (Siddiqi, 2017). When missing rates are higher than 30%, the variable has been dropped from the modelling dataset (Siddiqi, 2017).

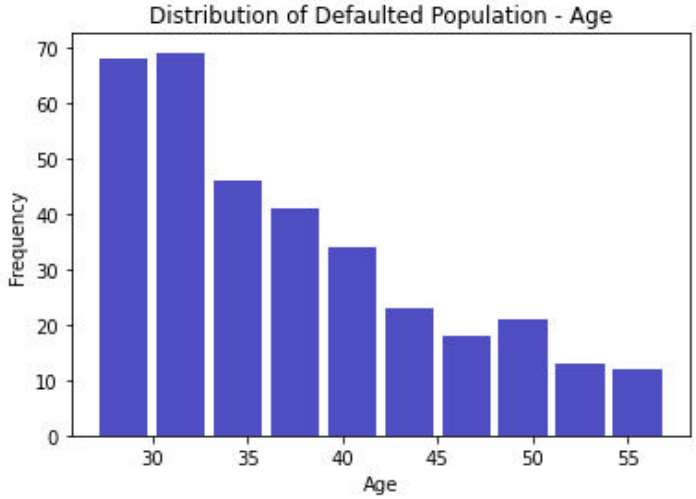

Figure 2: Distribution of Defaulted Population - Age

For what concerns the numerical features, the presence of outliers, i.e., observations that are significantly different from most of the data points in a set, can have negative impact on model training.
Therefore, their detection and treatment are important as well. There are several methods for dealing with outliers. In our case study, we now discuss the *Z-scores* and *Inter Quartile Range (IQR)* fences, also known as Tukey's fences.
Assuming a set of observations $X = \{x_1, \dots x_n\}$ has mean $\bar{x}$ and standard deviation σ, the Z-score of the observation $i \in \{1, \dots n\}$ is defined as:

$$Z_i = \frac{x_i - \bar{x}}{\sigma} \quad (1)$$

[3] https://www.ecb.europa.eu/paym/coll/loanlevel/transmission/html/index.en.html

By construction, the distribution of the Z-scores is standard normal *N(0,1)*. Within the Z-scores approach, the common practice is to classify as outliers those observations with absolute value greater than three.
The IQR approach measures the dispersion in the dataset by calculating the difference between the third (Q_3) and the first (Q_1) quartile of the data's distribution. The formula for calculating the Tukey's fences is given by:

$$[Q_1 - kIQR, Q_3 + kIQR] \quad (2)$$

where $IQR = Q_3 - Q_1$ and k is a positive constant. In our case study, we refer to the IQR method to detect outliers and we choose k=1.5, since this is the most common selection in banking practice. Once detected, there are several approaches for treating outliers including deleting, capping/flooring, replacing, or transforming (e.g., logarithmic transformation) the data. In our example, we chose to floor and cap the outliers in each case by using the lower and upper bounds of the Tukey's fences given in eq. (2). In Table 1 we present the number of available data, missing and outlier rate per variable in our dataset. As discussed at the beginning of this section, the expected total number of observations per variable is 1,000; however, as can be seen in the following table, this is not always the case.

Feature Name	Number of Observations	Missing Rate	Outliers Rate
Additional Loans	1,000	0%	N/A
Age	1,000	0%	0%
Application Date	1,000	0%	N/A
Application ID	1,000	0%	N/A
Burreau Score Value	1,000	0%	0%
Current Interest Rate Index	1,000	0%	N/A
Default Indicator	1,000	0%	0%
Employment Contract Type	997	0%	N/A
First Time Buyer	540	46%	N/A
Foreign National	688	31%	N/A
Interest Rate	1,000	0%	0.7%
Interest Rate Type	1,000	0%	N/A
Loan Term	955	5%	0%
Loan To Value	1,000	0%	0%
Marital Status	958	4%	N/A
Number of Debtors	1,000	0%	0%
Payment Schedule	987	1%	N/A
Post Code	668	33%	N/A
Primary Income	1,000	0%	0%
Principal	1,000	0%	1.5%
Property Rating	1,000	0%	N/A
Property Type	1,000	0%	N/A
Savings Size	961	4%	N/A
Secondary Income	236	76%	0%
Secondary Income Index	1,000	0%	N/A

Table 1: Data Quality

From Table 1 we observe that the variable "Secondary Income" has a missing rate of 76%. As discussed earlier, variables with such a high percentage of missing values should be excluded from the training sample. However, for this one this shouldn't be the case since from Table 9 we can infer that there is a direct one-to-one relationship between the features "Secondary Income", and "Secondary Income Index".
Additionally, in Figure 3 the variable "Secondary Income Index" has no missing values and out of the 1,000 total entries, 236 are equal to "Y". Thus, this implies that the true missing rate of "Secondary Income" is in fact zero.

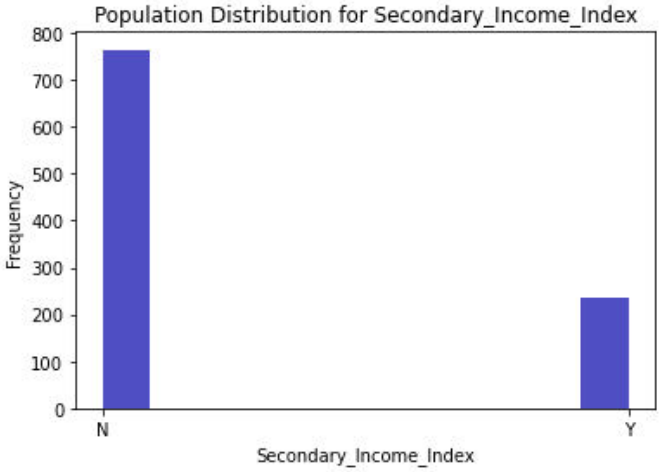

Figure 3: Population Distribution for Secondary Income Index

Table 9 (in bold) provides the final list of candidate variables. The variables "First Time Buyer", "Foreign National" and "Post Code" have been dropped from the list as they have a missing rate greater than 30% and any outliers for the remaining variables have been treated as discussed above. Moreover, the variable "Application ID" has also been removed as it bears no real information since it represents the unique identifier of the borrower in the dataset. Furthermore, our earlier qualitative analysis showed that the variables "Secondary Income" and "Secondary Income Index" are equivalent. Therefore, we do not include both in the analysis since in the case they are chosen for model development, it will have an unwanted impact on the regression coefficients and will generate misleading results. For this reason and without significant loss of generality, in our case study we chose the numeric variable "Secondary Income" for further analysis. Once the quality of the dataset has been verified, the next step is to identify which of the variables have sufficient predictive power to be considered for PD modelling purposes. In our approach, we initially determine the set of variables which we will use to fit the LR model as described in the following and then we further refine this list by considering: the level of statistical significance of each variable once LR has been fitted and the presence of multicollinearity among the variables. We underline that we chose Logistic Regression (LR) as our preferred modeling method because: it is well suited for such type of PD problems since it limits the output in the (0,1) space (the target is of binary type), it is straight forward to implement, it has low computational cost, and it is widely used in the banking sector for such type of problems. Let $A = \{feature_1, \dots feature_n\}$ be the set of all the predictive features, then the logistic regression equation, in terms of the logarithm of the odds, is given by:

$$log(odds) = log\left(\frac{PD}{1-PD}\right) = \alpha + \sum_{i=1}^{n} \beta_i \, feature_i, \qquad PD \in (0,1), \quad (3).$$

The regression coefficients α and β_i are estimated through the non-linear least square method.
LR can be applied using either the raw data of each feature or by grouping it into a number of bins/categories, therefore converting all variables to categorical ones. The latter approach is usually preferred due to the benefits it offers including but not limited to:

- Reducing the number of categories in the case of discrete variables and thus reducing computational time.
- Ensuring monotonicity in terms of predictor variables given their linear relationship with the dependent variable.
- Providing better understanding of the behavior of each predictor variable.

In our case study, we adopt a quantile-based grouping approach and then we implement the Weight of Evidence Encoding (WoE) (Baesens, Roesch, & Harald, 2016). WoE is a univariate measure describing the relationship between a predictor variable and the dependent one. Assuming that we have consolidated the values of a feature into m distinct bins (or equivalently "buckets" or "groups"), the WoE is given by:

$$WoE_i^{feature} = ln\left(\frac{Distribution\ of\ Performing\ cases}{Distribution\ of\ Defaulted\ cases}\right), i \in \{1,2, \dots m\} \quad (4)$$

Once the WoE for all predictor variables has been calculated, eq. (3) becomes:

$$log(odds) = log\left(\frac{PD}{1-PD}\right) = \alpha + \sum_{i=1}^{n} \beta_i \, WoE^{feature_i}, \qquad PD \in (0,1). \quad (5)$$

In our dataset we decided to group the continuous variables into 10 bins while we use the set of distinct values of each of the categorical ones as per our initial segmentation. Then, we calculate the Weight of Evidence per bin per feature as defined by eq. (4). Next, we further adjust the bins until strict monotonicity with respect to WoE is achieved. It should be noted however that fine tuning of the bins should always be performed in a careful manner and in line with Subject Matter Expert input since crude enforcement of such an approach could sometimes lead to a biased model behavior. In addition to improving model performance, this encoding process allows for smooth resolution of the missing values problem since, when applicable, a separate bin can be created to group them together. Following the fine tuning of the bins, for each feature we computed the Information Value (IV); it is measure of the predictive strength of a feature, given by the following formula:

$$IV_{feature} = \sum_{i=1}^{m} (Distribution\ of\ Performing\ Cases_i - Distribution\ of\ Defaulted\ Cases_i) WoE_i^{feature} \quad (6).$$

Referring to the above formula, a variable is classified as weak, medium or strong predictor based on the following widely used bandwidths:

- IV < 0.1: Weak Predictor.
- 0.1 ≤ IV < 0.3: Medium Predictor.
- IV ≥ 0.3: Strong Predictor[4]

In our case study, a variable is included into the list of candidates for modelling if and only if its IV is greater than or equal to 0.1.
We conclude this paragraph providing an example of how to calculate the WoE and IV for the variable "Burreau Score Value" described in Table 9.
As mentioned earlier, we start our approach by creating ten bins of equal size each containing 100 observations each and then we calculate the number of performing and defaulted obligors per bin (Default Indicator = 0). Then, we compute the distribution of both performing and defaulted population segments.
This is done by dividing the number of obligors per bin per segment (performing/defaulted) by the total number of obligors belonging to the same segment. Finally, the WoE per bin is calculated using eq. (4).

Bin	Observation	Lower Bound	Upper Bound	Performing Cases	Defaulted Cases	Dist. Perf. Cases	Dist. Def. Cases	WoE
1	100	500	539	59	41	7.52%	19.07%	-0.9311
2	100	539	585	61	39	7.77%	18.14%	-0.8477
3	100	586	629	70	30	8.92%	13.95%	-0.4477
4	100	631	670	79	21	10.06%	9.77%	0.0299
5	100	670	695	80	20	10.19%	9.30%	0.0912
6	100	695	719	89	11	11.34%	5.12%	0.7957
7	100	719	743	82	18	10.45%	8.37%	0.2213
8	100	744	769	85	15	10.83%	6.98%	0.4396
9	100	769	804	91	9	11.59%	4.19%	1.0186
10	100	804	900	89	11	11.34%	5.12%	0.7957

Table 2: WoE for the variable "Burreau Score Value", Bins = 10

As can be seen from Table 2, a reversal trend is observed between bins 6 and 7 and bins 9 and 10. Furthermore, the WoE for bins 1 to 3 has the same sign while bins 4 and 5 have similar value. Thus, it would make sense to group together bins 1 - 3, 4 - 5, 6 - 7 and 8 - 10. In our case, this can be done by generating the respective quartiles for this variable.

Bin	Observation	Lower Bound	Upper Bound	Performing Cases	Defaulted Cases	Dist. Perf. Cases	Dist. Def. Cases	WoE
1	250	500	607	157	93	20.0%	43.26%	-0.771
2	250	607	695	192	58	24.5%	26.98%	-0.098
3	250	695	753	214	36	27.3%	16.74%	0.487
4	250	754	900	222	28	28.3%	13.02%	0.775

Table 3: WoE for the variable "Burreau Score Value", Bins = 4

[4] When a variable has IV greater than 0.5 usually is considered as a suspiciously strong predictor. To verify the validity of such an outcome, it is suggested that further data investigation is conducted coupled with qualitative analysis regarding its impact and importance in the modelling process.

From Table 3 we can see that when 4 bins are used, the WoE for this variable is strictly monotonous as opposed to the case of 10 bins (see Table 2) implying better predictive performance. Finally, if we use the values Table 3 in equation (6) we have that $IV_{\text{Burreau Score Value}} = 0.35 > 0.1$, therefore "Burreau Score Value" is a candidate variable for modelling. In addition, since we implemented WoE encoding, each unique value of this variable is mapped to one and only one bucket as defined in Table 3.
Table 4 below provides the final number of bins along with the calculated IV for each of the final variables in our dataset. Remember that the IV cut-off rate for including a variable in the analysis has been set to 0.1 meaning that from now on we focus only *on the first eight variables* from the list below[5]. Finally, when possible, we have created a separate bin for the missing values allowing for smooth encoding process.

Feature Name	IV	# Bins
Loan To Value	*0.716827*	*4*
Additional Loans	*0.626395*	*2*
Secondary Income	*0.490542*	*3*
Age	*0.394578*	*5*
Burreau Score Value	*0.351431*	*4*
Number of Debtors	*0.259039*	*3*
Savings Size	*0.157953*	*3*
Property Rating	*0.104202*	*4*
Application Date	0.097471	4
Principal	0.057218	5
Primary Income	0.043727	5
Property Type	0.028451	4
Interest Rate	0.026268	5
Interest Rate Type	0.026173	2
Marital Status	0.020673	4
Loan Term	0.013827	4
Payment Schedule	0.007772	3
Current Interest Rate Index	0.003292	3
Employment Contract Type	0.000378	4

Table 4: Information Value

5.2 Model Development

The aim of this case study is to address borrowers' default prediction problem via estimating the Probability of Default (PD) of each individual loan. The approach we propose in this article consists of a *regression problem,* i.e., we aim to predict the value of a depended variable (the PD) via modeling its relationship with one or more independent variables (the features). In this section we show how to fit a *Logistic Regression model* on the final list of candidate variables subject to the two selection criteria mentioned in paragraph 5.1 (the level of statistical significance of each variable once LR has been fitted and the presence of multicollinearity among the variables). Specifically, a variable will remain in our model if it is statistically significant, i.e. p-value ≤ 0.05 and its Variance Inflation Factor (VIF) is less than 5 as this is a commonly used cut-off threshold in banking practice (see (Ron Johnston, Jones, & Manley, 2018)). The VIF measures the degree of multicollinearity in our multiple regression model. In general, multicollinearity means that two or more variables of a set can linearly approximate some other variables from the same set. If a high degree of multicollinearity is present,

[5] We note that for the two variables having IV greater than 0.5 no data quality issues have been raised. Furthermore, in view of Table 9, it is evident that they contain significant qualitative information for our modelling purposes.

there could be various negative effects such as making the estimates of the regression coefficients unstable or inflating their standard errors. This would imply a bias in the interpretation of the relationship between a predictor variable and a dependent variable. The formula for calculating the VIF is given by:

$$VIF_{variable_i} = \frac{1}{1 - R^2_{variable_i}} \quad (7)$$

where $R^2_{variable_i}$ is the R - squared value resulting from regressing the i[th] variable against all other available variables. We underline that in our case we have implemented the WoE encoding. This approach shifts the focus from inferring relationships between the independent variables and the target one towards enhancing the predictive power of the model. Therefore, the interpretation of the regression coefficients is not as straightforward as it would be had we used the actual variable values.
Table 5 below provides the estimated coefficients via LR and the respective p-values for each of our eight candidate variables.

#	Feature Name	Coefficient	Standard Error	p-Value
0	Constant	-1.4571	0.1682	<0.001
1	Loan To Value (WoE)	-1.5686	0.1847	< 0.001
2	Additional Loans (WoE)	-1.1834	0.1512	< 0.001
3	Secondary Income (WoE)	-1.8103	0.4623	0.0001
4	Age (WoE)	-0.4895	0.2466	0.0472
5	Burreau Score Value (WoE)	-0.9739	0.2878	0.0007
6	Number of Debtors (WoE)	**0.8648**	**0.4907**	**0.078**
7	Savings Size (WoE)	2.3549	0.8501	0.0056
8	Property Rating (WoE)	-1.0258	0.3531	0.0037

Table 5: LR Model Coefficients (Initial fit)

The variable "Number of Debtors" is not statistically significant given that its p-value is outside the tolerance (p-value ≤ 0.05). Therefore, it will be dropped from the list and the model will be refit. The outcomes of the updated model after removing the variable "Number of Debtors" are reported in Table 6, where all the variables are statistically significant and have VIF < 5.

#	Feature Name	Coefficient	Standard Error	p-Value	VIF
0	Constant	-1.6259	0.1421	<0.001	-
1	Loan To Value (WoE)	-1.5382	0.1827	< 0.001	1.104557
2	Additional Loans (WoE)	-1.1786	0.1506	< 0.001	1.051574
3	Secondary Income (WoE)	-1.8208	0.4684	< 0.001	2.921492
4	Age (WoE)	-0.4856	0.2464	0.0487	1.923134
5	Burreau Score Value (WoE)	-0.9582	0.2864	< 0.001	1.945571
6	Savings Size (WoE)	2.3085	0.8596	0.0072	2.8061
7	Property Rating (WoE)	-1.0139	0.3585	0.0041	1.206845

Table 6: LR Model Coefficients (Final)

At this stage, we compute the PD for each individual borrower using equation (5). Once we have identified the features that have a significant predictive power for the LR, we decided to further investigate the default prediction problem using also a non-linear classifier. Classifiers using non-linear algorithm are as an example the support vector machine, artificial neural network, k-nearest neighbour, naïve Bayes, random forest, and Bayesian network. As mentioned above, in this paper we decided to compare LR with k-NN since this approach is computationally affordable and completely nonparametric. The basic principle behind this method is that a given instance within a data set will generally exist in proximity with other instances sharing similar properties. Hence, additional information about an instance can be obtained by observing other instances that are close to it, that is, the Nearest Neighbours (NNs).

If the instances within a data set are tagged with a classification label, then the class of a new instance can be determined by observing the classes of its NNs. The advantage of nearest-neighbour classification is its simplicity. There are only two choices the modeler must make: the number of neighbours k and the distance metric to be used. We decided to use the Euclidean distance as measure. we have implemented the k-NN algorithm for k =3, 5, 7 using the Euclidean distance (Sun & Huang, 2010).
As mentioned in the introduction, the calibration of the model is out of the scope of this article; despite that, we want to conclude this paragraph describing the main steps of the calibration procedure, since it would be the final task to perform in building a PD model before validating it.
In real world banking practice, retail portfolios of medium size usually contain tens of thousands of borrowers. Therefore, the risk profile at portfolio level is easier to understand if the obligors are classified into grades and a PD - representative of the population – is assigned to each grade. The calibration sample usually contains hundreds of thousands of data while the observation period extends to more than 10 years.
This process is commonly referred to as calibration of a rating system and it is usually performed in a hybrid manner where statistical tools and expert judgement are combined in such way that:

1. All grades are homogeneous, meaning that obligors of similar risk profile are grouped together.
2. All grades have distinct risk characteristics and differ from one another.
3. All grades have a significant number of borrowers.
4. The default rate across the grades displays a monotonous behaviour.
5. The PD assigned to each grade is representative of the long run average of the portfolio's default rate.

5.3 Model Validation

In this section we introduce the concept of Model Validation and its relevance for every financial institution. We will briefly introduce the most used statistical tools used in this process and we apply them to our case study. Model validation is a well-regulated process with the main goal to verify whether a model developed for assessing any financial risk (in our case credit risk) addresses in a satisfactory and appropriate manner the business needs and its design objectives. Therefore, all financial institutions have in place large and independent teams tasked with ensuring that every model used for risk management is quantitatively and qualitatively sound and fully compliant with the latest regulatory requirements. For this purpose, the data used for model development, the model design, the model documentation, and the policy framework on which the model was built are subject to investigation and validation. In other words, no model will go live in a production system without having successfully passed the validation process. The process of validating a newly developed model is referred to as "Initial Validation". Once the model has been deployed to production, it is then subject to periodic reviews to ensure that its performance remains fit for purpose and is still aligned with the current regulatory framework, which is frequently updated. The process of periodically assessing the quality of a live model is referred to as "Periodic Validation". If a model fails some periodic validation review, for example due to changes in the macro-economic environment, then depending on the severity of the identified issues either it will have to be re-trained by incorporating the latest available data or completely re-built. Given that most models are mainly statistically developed, the validation process has a strong quantitative component. This is especially true for PD models for credit risk assessment. In the quantitative analysis, the most prominent areas of interest are to verify whether the population used during model development is representative of the current population and whether the model has good predictive ability and discriminatory power.
In literature, there are two types of samples used for validation purposes: the *in-time* sample and the *out-of-time* sample. The *in-time* sample randomly reserves a portion, usually from the 20% to the 30%, of the dataset used for the model development to the validation. This means that, the model is trained on the remaining percentage (from 70% to 80%) and validated on the reserved sample. The *out-of-time* sample contains information that are outside the time frame of the data used for model development. Indeed, the validation is performed on a later dataset than that on which the model has been fitted. To check the representativeness of the training dataset upon the *out-of-time* validation sample, the Population Stability Index PSI is used. The PSI is an index that measures how much a variable has shifted over time and is used to monitor applicability of a statistical model to the current population. However, PSI statistic intends to capture any significant shifts in the distribution of the development sample over time; since our model is not subject to calibration and it is based on a simulated dataset, those two tests are not applicable to our case study. For this reason, we have applied an *in-time* validation with 70 / 30 split, meaning that 70% of the dataset was used for model development and 30% for validation purposes.
The discriminatory power of a PD model measures its ability to differentiate well between defaulted and non-defaulted borrowers. The most popular tools available for this purpose are the Receiver Operating Characteristic (ROC) curve, the Area Under the ROC Curve (AUC), the Cumulative Accuracy Profile (CAP) and the Accuracy Ratio (AR). We decided to use the Area Under the Curve (AUC) and the Accuracy Ratio (AR), since they are robust performance measures for a large number of classifiers (including LR) and they are extensively used in the literature related to ML methods applied to credit risk (see for example (Tasche D. , 2008), (Tang & Chi, 2005), (Fantazzini & Figini, 2009), (Kruppa, Schwarz, Arminger, & Ziegler, 2013), (Addo, Guegan, & Hassani, 2018)). The AUC is a performance measurement statistic describing the strength of the classifier in terms of assigning a lower PD to a true random performing observation than a true random defaulted observation. In general, the performance of a classifier like the AUC can be described through a confusion matrix of the following form:

		Observed	
		Default	**Performing**
Predicted	**Default**	True Positive	False Positive
	Performing	False Negative	True Negative

Table 7: Confusion Matrix

Here, True Positive (TP) represents the number of obligors that the model classified as in default and were actual defaults, False Positive (FP) represents the number of borrowers that the model classified as in default but were in performing status, False Negative (FN) represents the number of obligors that the model classified as performing but were in default status and finally, True Negative (TN) represents the number of obligors that the model classified as performing and were in performing status. Starting from the values contained in the confusion matrix, one can calculate the following two rates:

$$True\ Positive\ Rate\ (TPR) = \frac{TP}{TP + FN} \quad (8)$$

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{FP + TN} \quad (9)$$

Based on this set-up, the Receiver Operating Characteristic (ROC) curve is then defined as the set of all points (FPR, TPR) across all possible cut-off probability thresholds, i.e., the level representing the probability of a true prediction. Notice that all points along the diagonal line represent a model for which the TPR is equal to the FPR for each cut-off threshold, thus, a random model with no substantial discriminatory power. Furthermore, points (0,0) and (1,1) imply cut-off thresholds of 1 and 0 respectively. Simply put, in the first case all points have been classified as in performing status meaning that TP= FP =0 therefore, (FPR, $TPR) = (0,0). In the second case, all points have been classified as in default status meaning that FN=TN=0 therefore (FPR, TPR) = (1,1). Once the ROC curve has been generated, the AUC statistic is then simply defined as the area between the FPR axes and the ROC curve. The discriminatory power of a model is defined as follows:

- AUC = 0.5: No Substantial Discrimination.
- 0.5 < AUC < 0.7: Weak Discrimination.
- 0.7 ≤ AUC < 0.8: Moderate Discrimination.
- AUC ≥ 0.8 High Discrimination.

In our case study, we consider the results obtained in Table 6 and we compute the ROC curve in case of LR and the k-NN algorithm for k =3, 5, 7 (see Figure 4 and Figure 5 below).

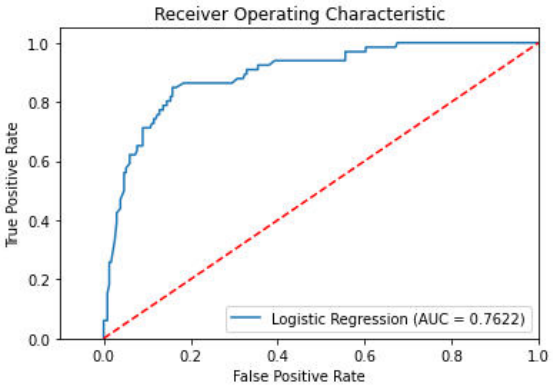

Figure 4: Logistic Regression, 7 Features, AUC

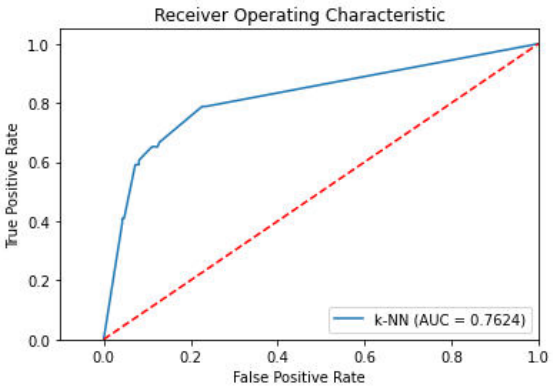

Figure 5: k-NN, k = 3, 7 Features, AUC

Table 8 shows the AUC values in case of LR and the k-NN algorithm for k =3, 5, 7. Both models have a moderate discrimination power, the AUC difference between LR and k-NN is negligible for k = 3, while the LR algorithm performs better than the k-NN when k becomes larger (k = 5, 7).

Number of Features	LR	3NN	5NN	7NN
7	0.7622	0.7624	0.7213	0.7494

Table 8: Model Comparison – AUC – 7 features

Another popular tool used for measuring the discriminatory power of a model is the Cumulative Accuracy Profile (CAP) (see (Tasche D. , 2008)). This curve differs from the ROC one in the sense that instead of plotting the TPR versus the FPR across all possible regression thresholds, it represents the cumulative percentage of default rate against the corresponding cumulative population percentage. When evaluating a PD model via CAP, it is assumed that the population has been sorted in descending order PD wise. Furthermore, similarly to the ROC curve, the diagonal line described by the equation y=x represents a random model with no substantial discriminatory power, while the perfect model is described by a CAP curve in which the maximum response rate is achieved at the lowest possible end of the population's distribution. As an example, Figure 6 displays three different models of poor, medium and high discriminatory power. Note that all three of them are bounded by the random and perfect model CAP curves.

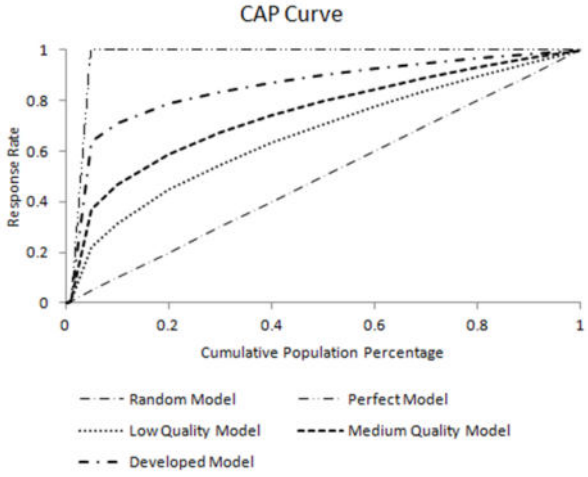

Figure 6: Sample CAP Curve

Starting from the CAP curve, the discriminatory power of a model can be estimated using the Accuracy Ratio (AR) which is defined as the quotient of the area between the CAP curves of the developed and the random model over the area defined by the CAP curves of the perfect and the random model. Specifically, if A denotes the area between the developed and the random model and B the area between the perfect and the developed model:

$$AR = \frac{A}{A+B} \quad (10)$$

In a successful model, the AR ranges as $0 \leq AR \leq 1$, and the higher the value is, the stronger the discriminatory power of the model is. Observe that equation 10 is equivalent to AR = 2AUC – 1. In our case study, since AUC ranges from 0. 7213 to 0.7624, then AR ranges from 0.4426 to 0.5248, confirming the moderate predictive power of the PD model.

Typically, once the model validation is completed, in the event of a positive outcome the model can be either enabled to the production system or can continue to be used as is until the next periodic review. However, if this is not true then depending on the severity of the recommendations that have been called out by the validation team, the model will be subject to several corrective actions spanning from minor adjustments to some of its components up to full re-built. To this point, it should be clear that model development and validation are two distinct and independent functions both serving the same objective, i.e., to ensure that all models used by a financial institution are of the best possible quality.

6. Conclusions

In this paper we have performed a critical review of the most used model development and validation procedures for credit risk in retail banking. After having described the impact of Regulator on Credit Risk Methodologies and briefly motivated the introduction of ML approaches for regulatory capital estimation, we have reviewed some of the traditional statistical methods to PD and LGD estimates. Then we have focused on more recent studies based on Machine Learning techniques. As next step, we have constructed a case study showing how to develop and validate a PD model via a Machine Learning Techniques. We have used a simulated dataset to show the data pre-processing, cleansing and the application of the Weight of Evidence Encoding, a powerful technique well suited for Logistic Regression problems since it introduces a monotonic relationship between the target variable and the predictors. Finally, we have compared the LR model against non-linear classifiers, the k-NN algorithm (for k= 3, 5 & 7) and we measured the predictive power of both models using the AUC and AR. The results indicate that on the one hand, the LR classifier coupled with WoE performed in a similar way to the k-NN for k =3, while it outperformed LR for k=5 and k = 7. Given the interpretability and simplicity that the LR method offers coupled with less computational effort and complexity as opposed to non-parametric machine learning models, we conclude that the choice of an LR model leveraging the WoE technique is very well suited and produced good predictive power in comparison to k-NN.

Machine learning methods are of major importance in the retail banking sector since most, if not all, of the models used for risk management make us of these techniques. As Science and Technology advance year over year and at a rapid pace, as well as the availability of large dataset is increasing, financial institutions are constantly strengthening their decision-making processes by developing sophisticated algorithms which can consume large pools of data in a very short time and are in line with the current regulation as well. Since banks play a very important role in the global economy, better decision-making means better risk management which in turn means greater financial stability. It is always worth remembering that despite the increase of ML techniques and availability of large dataset, risk management remains human activity and effective risk management would not be possible without strong critical reflection and experts' judgement at every step of the process.

Appendix

The following table provides the feature name, description and information regarding whether it is included or not in the model development phase. The variable names follow the format presented in the ECB's Loan-Level data templates. In bold you can find the features that have been included in the development process.

#	Feature Name	Description	Type
1	**Additional Loans**	**Nominal variable, describing whether the obligor has additional loans. Values: Y/N**	**Account Type**
2	**Application Date**	**Ordinal variable, describing the date of the application. Values between January 2010 to December 2019.**	**Account Type**
3	Application ID	Nominal variable, describing the obligor's unique identifier.	Account Type
4	**Burreau Score Value**	**Numeric variable (Integer) describing the obligor's credit score. Values between 500 to 900.**	**Account Type**
5	**Current Interest Rate Index**	**Nominal variable, describing the index on which the mortgage was written on. Values: Euribor_3M, Euribor_6M, No_Index.**	**Account Type**
6	***Default Indicator (Target Variable)***	***Numeric variable, describing whether the obligor is in default or not. Values: 1 (Default), 0 (Performing).***	***Account Type***
7	First Time Buyer	Nominal variable, describing whether the obligor is a first-time buyer. Values: Y /N.	Account Type
8	**Interest Rate**	**Numeric variable (float), describing the interest rate applied to the mortgage. All interest rates have been rounded to the first decimal place. Values from 4.3% to 7.5%.**	**Account Type**
9	**Interest Rate Type**	**Nominal variable, describing whether the interest rate is fixed or floating. Values: Fixed / Floating.**	**Account Type**
10	**Loan Term**	**Numeric variable (Integer), describing the maturity of the mortgage. Values: 20, 25 & 30.**	**Account Type**
11	**Loan To Value**	**Numeric variable (float), describing the percentage of the value of the mortgage with respect to the value of the property. Values (rounded to the second decimal place) range from 77% to 85%.**	**Account Type**
12	**Number of Debtors**	**Numeric variable (Integer), describing the number of obligors on which the mortgage is written on. Values 1, 2 or 3.**	**Account Type**
13	**Property Type**	**Nominal variable, describing the purpose for which the mortgage was given. Values (1) Holiday/second home (2) non-owner-occupied/buy-to-let (3) Other (4) Owner-occupied.**	**Account Type**
14	**Payment Schedule**	**Nominal variable, describing the interest payment schedule. Values: 3M / 6M.**	**Account Type**
15	**Principal**	**Numeric variable (Integer), describing the face value of the mortgage. Values (rounded to thousands) from 144,000 to 648,000.**	**Account Type**
16	**Property Rating**	**Ordinal variable, describing the quality of the property. Values (from best to worst): (1) CAT_1 (2) CAT_2 (3) CAT_3 (4) CAT_4**	**Account Type**
17	Secondary Income Index	Nominal variable, describing whether a secondary income exists. Values: Y / N	Account Type
18	Foreign National	Nominal variable, describing whether the obligor is of foreign nationality. Values: Y / N	Demographic Type
19	Post Code	Nominal variable, describing the zip code of the obligor's address.	Demographic Type
20	**Age**	**Numeric variable (Integer), describing the obligor's age at application date. Values from 27 to 57.**	**Sociological Type**
21	**Marital Status**	**Nominal variable, describing the marital status of the obligor. Values: M/C (Married or Cohabitating), D (Divorced), S (Single).**	**Sociological Type**
22	**Primary Income**	**Numeric variable (Integer), describing the obligor's annual salary. Values (rounded to thousand): 36,000 to 95,000**	**Sociological Type**
23	**Saving Size**	**Ordinal variable, describing the obligor's saving status. Values: Above 50k / Below 50k.**	**Sociological Type**
24	**Secondary Income**	**Numeric variable (Integer), describing the size of the obligor's annual secondary income. Values (rounded to thousand) from 4,000 to 15,000.**	**Sociological Type**
25	**Employment Contract Type**	**Nominal variable, describing the obligor's occupation type. Values: (1) Fixed Term Contract (2) Permanent Contract (3) Self-Employed**	**Sociological Type**

Table 9: Variable Description

Bibliography

EBA/GL/2017/06. (2017). *Guidelines on credit institutions' credit risk management practices and accounting for expected credit losses.*

Abdou, H., Pointon, J., & Elmasry, A. (2008). Neural Nets Versus Conventional Techniques in Credit Scoring in Egyptian Banking. *J. Expert Systems with Applications, 35(3)*, 1275-1292.

Addo, P., Guegan, D., & Hassani, B. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. *University Ca' Foscari of Venice, Dept. of Economics Research Paper Series No. 08/WP/2018*.

Altman, I. E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*, 589-611.

Amin, M. F., Islam, M. M., & Murase, K. (2009). Ensemble of single-layered complex-valued neural networks for classification tasks. *Neurocomputing, 72*, 2227-2234.

Angelini, E., Tollo, G. D., & Roil, A. (2008). A Neural Network Approach for Credit Risk Evaluation. *The Quarterly Review of Economics and Finance, 48(4)*, 733-755.

Baesen, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Van-Thienen, J. (2003). Benchmarking state-of-art classification algorithm for credit scoring. *Journal of the Operational Research Society, 54*, 627-635.

Baesens, B., Roesch, D., & Harald, S. (2016). *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS.* Wiley.

Bank for International Settlements. (2005a). *Basel Committee on Banking Supervision, International convergence of capital measurement and capital standards: a revised framework.* Basel, Switzerland.

Bank for International Settlements. (2005b). *An Explanatory note on Basel II IRB Risk Weight Functions.* Basel, Switzerland.

Bank for International Settlements. (2009). *Range of practices and issues in economic capital frameworks.* Basel, Switzerland.

Bank for International Settlements. (2011). *Basel III: A global regulatory framework for more resilient banks and banking systems.* Basel, Switzerland.

Bank for International Settlements. (2017). *IFRS 9 and expected loss provisioning - Executive Summary.* Basel, Switzerland.

Bharath, S. T., & Shumway, T. (2006). Forecasting default with the KMV-Merton model. *AFA 2006 Boston Meetings Paper*.

Black, F., & Cox, J. C. (1976). Valuing corporate securities: Some effects of bond indenture provisions. *The Journal of Finance 31(2)*, 351-367.

Breiman, L. (2001). Random Forest. *Machine Learning, 45(1)*, 5-32.

Duffie, D., & Lando, D. (2001). Term structures of credit spreads with incomplete accounting information. *Econometrica, 69(3)*, 633-664.

Duffie, D., & Singleton, K. (1999). Modelling term structure of defaultable bonds. *Rev. Financial Stud., 12*, 687-720.

Duffie, D., & Singleton, K. (2012). *Credit risk: pricing, measurement, and management.* Princeton University Press.

Eitel-Porter, R. (2021). Beyond the promise: implementing ethical AI. *AI Ethics 1*, 73-80.

Elliott, R. J., Jeanblanc, M., & Yor, M. (2000). On models of default risk. *Mathematical Finance, 10(2)*, 179-195.

European Central Bank. (2019). *ECB Guide to Internal Models.*

Fantazzini, D., & Figini, S. (2009). Random Survival Forests Models for SME Credit Risk Measurement. *Methodology and Computing in Applied Probability volume, 11*, 29-45.

Finger, C., Finkelstein, V., Pan, G., Lardy, J., Ta, T., & Tierney, J. (2002). *Credit Grades. Technical Document.* . New York: Riskmetrics Group.

Frye, J. (2000a). Collateral Damage. *Risk, April, 13(4)*, 91-94.

Frye, J. (2000b). Collateral damage detected. *Federal Reserve Bank of Chicago Working Paper, Emerging Issues Series, October*, 1-14.

Galindo, J., & Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. *Computational Economics, 15(1)*, 107-143.

Hsieh, N. (2005). Hybrid mining approach in design of credit scoring model. *Expert Systems with Applications, 28*, 655-665.

Huang, Z., Chen, H. C., Hsu, J., Chen, W. H., & Wu, S. (2004). Credit Rating Analysis with Support Vector Machines and Neural Networks: A Market Comparative Study. *Decision Support System, 37(4)*, 543-558.

Irwin, R. J., & Irwin, T. C. (2012). *Appraising Credit Ratings: Does the CAP Fit Better than the ROC?* IMF Working Paper, 12/122.

Jarrow, R. A. (2001). Default parameter estimation using market prices. *Financial Analysts Journal, 57(5)*, 75-92.

Jarrow, R., & Protter, P. (2004). Structural versus reduced-form models: A new information based perspective. *J. Investment Management, 2(2)*, 34-43.

Jarrow, R., & Turnbull, S. (1992). Credit risk: Drawing the analogy. *Risk Magazine, 5(9)*, 51-56.

Jarrow, R., & Turnbull, S. (1995). Pricing Derivatives on financial securities subject to credit risk. *The Journal of Finance, 50(1)*, 53-85.

Joenssen, D., & Bankhofer, U. (2012). Hot Deck Methods for Imputing Missing Data. *Machine Learning and Data Mining in Pattern Recognition. MLDM 2012. Lecture Notes in Computer Science. 7376*, pp. 63-75. Berlin: Springer.

Jokivuolle, E., & Peura, S. (2000). A model for estimating recovery rates and collateral haircuts for bank loans. *Bank of Finland Research Discussion Paper, 2*.

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications, 40(13)*, 5125-5131.

Lee, M., & Floridi, L. (2021). Algorithmic Fairness in Mortgage Lending: from Absolute Conditions to Relational Trade-offs. *Minds & Machines*, 165-191.

Merton, R. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance 29(2)*, 449-470.

Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 36*, 3028-3033.

Orgler, Y. (1970). A credit scoring model for commercial loans. *J. Money Credit Bank., 2*, 435-445.

Pang, S., & Gong, J. (2009). C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks. *Systems Engineering - Theory and Practice, 29(12)*, 94-104.

Ron Johnston, R., Jones, K., & Manley, D. (2018). Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. *Quality and Quantity, 52*, 1957–1976.

Sadhwani, A., Giesecke, K., & Sirignano, J. (2021). Deep Learning for Mortgage Risk. *Journal of Financial Econometrics*, 313-368.

Sadok, H., Sakka, F., & El Hadi El Maknouz, M. (2022). Artificial intelligence and bank credit analysis: A review. *Cogent Economics & Finance, 10(1)*.

Schebesch, K. B., & Stecking, R. (2005). Support vector machine for classifying and describing credit applicants: Detecting typical and critical regions. *Journal of the Operational Research Society, 56*, 1082-1088.

Shin, K. S., Lee, T. S., & Kim, H. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications, 28(1)*, 127-135.

Siddiqi, N. (2017). *Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards, 2nd Edition.* Wiley.

Sun, S., & Huang, R. (2010). An adaptive k-nearest neighbor algorithm. *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, 91-94.

Sun, S., & Huang, R. (n.d.). Sun, An adaptive k-nearest neighbor algorithm. *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, 91-94.

Tang, T., & Chi, L. (2005). Predicting multilateral trade credit risks: comparisons of Logit and FuzzyLogic models using ROC curve analysis. *Expert Systems with Applications, 28(3)*, 547-556.

Tasche, D. (2004). The single risk factor approach to capital charges in case of correlated loss given default rates. *arXiv preprint cond-mat/0402390*.

Tasche, D. (2008). Validation of internal rating systems and PD estimates. *The Analytics of Risk Model Validation*, 169-196.

West, D. (2000). Neural network credit scoring models. *Computers and Operations Research, 27(11-12)*, 1131-1152.

Yu, L., & Wang, S. L. (2009). An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: the case of credit scoring. *European Journal of Operational Research, 195*, 942-959.

Zou, L., & Khern-am-nuai, W. (2022). AI and housing discrimination: the case of mortgage applications. *AI Ethics*.