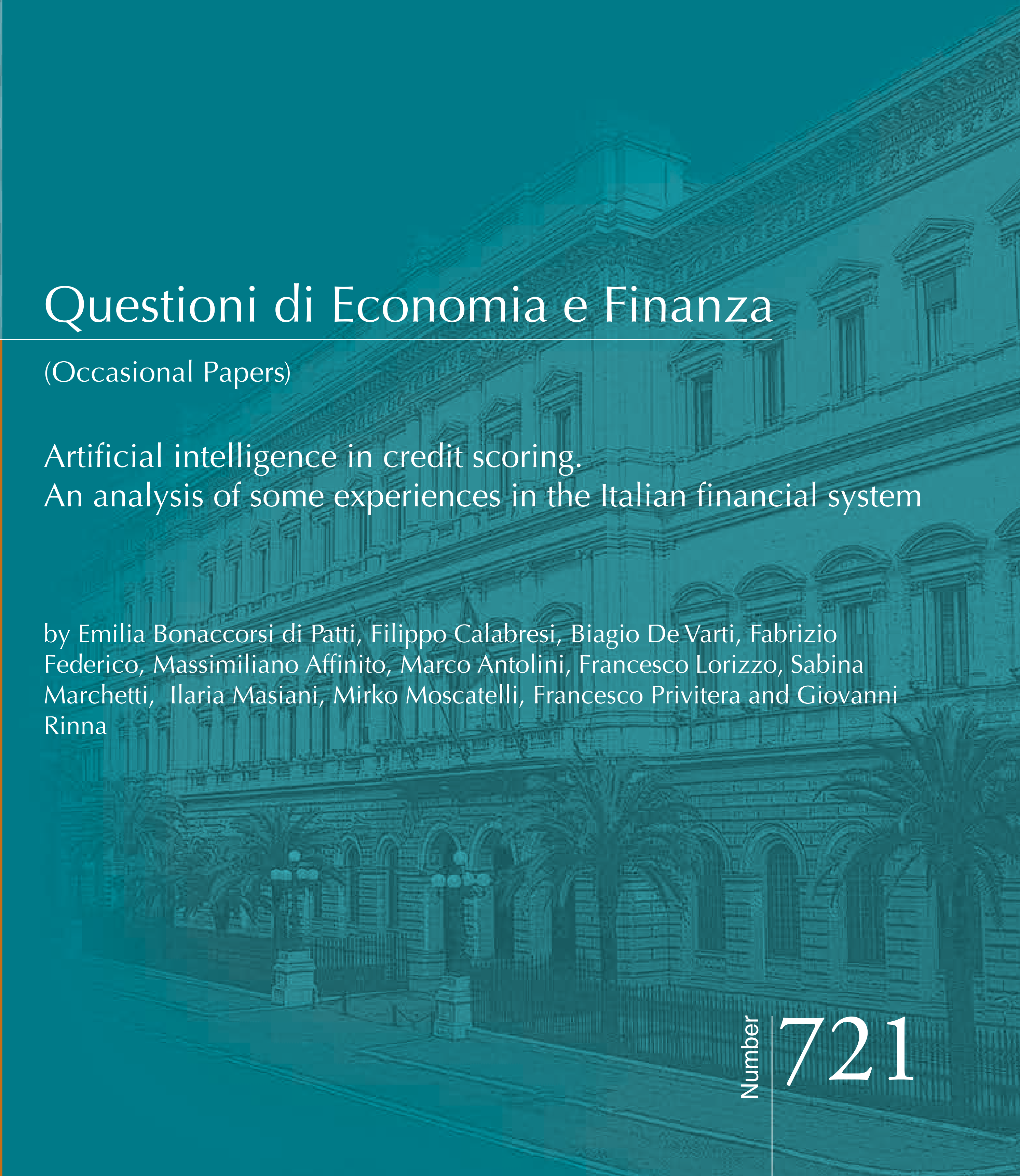
BANCA D'ITALIA
EUROSISTEMA



Questioni di Economia e Finanza

(Occasional Papers)

Artificial intelligence in credit scoring. An analysis of some experiences in the Italian financial system

by Emilia Bonaccorsi di Patti, Filippo Calabresi, Biagio De Varti, Fabrizio Federico, Massimiliano Affinito, Marco Antolini, Francesco Lorizzo, Sabina Marchetti, Ilaria Masiani, Mirko Moscatelli, Francesco Privitera and Giovanni Rinna

October 2022

Number 721

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Artificial intelligence in credit scoring. An analysis of some experiences in the Italian financial system

by Emilia Bonaccorsi di Patti, Filippo Calabresi, Biagio De Varti, Fabrizio Federico, Massimiliano Affinito, Marco Antolini, Francesco Lorizzo, Sabina Marchetti, Ilaria Masiani, Mirko Moscatelli, Francesco Privitera and Giovanni Rinna

Number 721 – October 2022

ISSN 1972-6627 (stampa)
ISSN 1972-6643 (online)

Stampato presso la Divisione Editoria e stampa della Banca d'Italia

ARTIFICIAL INTELLIGENCE IN CREDIT SCORING. AN ANALYSIS OF SOME EXPERIENCES IN THE ITALIAN FINANCIAL SYSTEM

by Emilia Bonaccorsi Di Patti[*], Filippo Calabresi[**], Biagio De Varti[§], Fabrizio Federico[§§], Massimiliano Affinito[**], Marco Antolini[§], Francesco Lorizzo[§§], Sabina Marchetti[*], Ilaria Masiani[§], Mirko Moscatelli[*], Francesco Privitera[**] and Giovanni Rinna[§]

Abstract

This report investigates the use of artificial intelligence and machine learning (AI-ML) techniques used by the Italian financial intermediaries to assess creditworthiness. The analysis aims to evaluate how financial intermediaries use AI-ML techniques for customer selection and management within credit processes, and to gather information on their awareness of specific risks that characterise such methodologies. Stemming from the theoretical analysis of conceptual determinants, techniques and legal/institutional context of AI-ML for credit scoring, the report provides insights on the survey on the adoption of such methodologies by financial intermediaries.

JEL Classification: C52, G21, J15, J16, O32.
Keywords: artificial intelligence, machine learning, credit, credit scoring, algorithmic bias, discrimination.
DOI: 10.32057/0.QEF.2022.0721

Contents

[*] Bank of Italy, Directorate General for Economics, Statistics and Research.
[**] Bank of Italy, Directorate General for Consumer Protection and Financial Education.
[§] Bank of Italy, Directorate General for Financial Supervision and Regulation.
[§§] Bank of Italy, Directorate General for Information Technology.

1. Introduction and main conclusions

The work analyses the use of artificial intelligence and machine learning (AI-ML) techniques to support the assessment of credit risk by Italian financial intermediaries. The objective of the analysis is to investigate the extent to which financial intermediaries are aware of the specific risks that characterize the use of advanced technologies in the sensitive matter of selecting and managing customers based on their creditworthiness. In particular, the aim was to assess to what extent AI-ML models for credit scoring would increase the possibility of biased customer selection with respect to their actual riskiness, including possible forms of discrimination, and how difficult it is for intermediaries to investigate this possibility given the difficulties in retracing the approaches followed by the models.

The work is based on two approaches: the first is a theoretical analysis of both the conceptual and technical key elements of AI-ML applied to credit scoring and of the related regulatory/institutional framework; the second is an in-the-field survey of the experience of Italian financial intermediaries in adopting such models.

The theoretical analysis outlines the technical, regulatory and policy contexts. The main conclusions can be summarised as follows:

- *"machine learning systems and explainability techniques are spreading rapidly*" — Chapter 2 briefly describes the main technological drivers of artificial intelligence that are relevant to the issue under consideration, namely the machine learning landscape, the use of big data and the application of "explainable AI*"* techniques. The upshot is that the increasing availability of data and improved machine learning models imply a level of complexity that requires technological advances to be accompanied by safeguards that ensure the necessary transparency of the approaches being followed;
- "*there is a trade-off between estimation accuracy and explainability*" — Chapter 3 briefly reviews academic studies that compare the use of ML techniques in credit assessment with traditional statistical approaches. In summary, ML leads to more accurate estimates, not least because it is able to exploit information and complex relationships between variables that have no clear economic meaning;
- *"ML systems do not necessarily learn correctly"*— Chapter 4, after outlining the possible biases of credit scoring systems (such as incorrect risk differentiation and discrimination against individuals or social groups), examines in detail the biases that may arise in AI-ML systems applied to the credit sector. As in traditional statistical systems, these biases can occur in the data collection, model specification and output analysis phases. However, in ML systems, biases also tend to generate a dangerous feedback loop in which they can be confirmed and amplified;
- "*existing prudential regulations make it possible to oversee a large part of the risks associated with AI-ML models" —* Section 5.1 presents the relevant prudential rules for financial intermediaries adopting AI-ML systems for credit management. As there are no specific requirements, the rules that apply are the general provisions on the effectiveness of

the risk governance, management and control mechanisms included in the supervisory regulations and, for banks, in the EBA Guidelines on granting and monitoring credit. In the case of ML models used to calculate capital requirements, the relevant specific rules come into play. Even in the absence of specific requirements for AI-ML models, the full application of the general principles included in the rules should ensure that a large part of the specific risks of AI-ML models are mitigated;

- "*non-discrimination in customer relationships is a principle that is not fully set out in the rules"* — Section 5.2.1 shows that there are few general references to the principle of non-discrimination in the customer protection regulations; the search for the right balance between fairness in customer relationships and the freedom of enterprise of financial intermediaries in lending is further stimulated by the adoption of AI-ML techniques;
- "*new challenges posed by the GDPR on disclosure to customers selected using ML"* — Section 5.2.2 illustrates the importance given by European privacy legislation (General Data Protection Regulation, GDPR) to issues relating to profiling, customer consent and the right to enhanced information in the case of automated decision-making; the use of AI-ML techniques makes it more difficult for financial intermediaries to explain to each requesting customer the reasons why the model decided not to grant credit;
- "*international convergence on principles is difficult to translate into rules and practices"* — Section 5.3 summarizes a number of recent assessments by national and international authorities and bodies with regard to the governance of AI. On the one hand, there is a high degree of convergence on principles, such as the need for human control over machine-driven decision-making, transparency and non-discrimination. On the other hand, translating these principles into rules, in order to balance innovation and the protection of consumer rights, seems less straightforward.

The exercise was carried out through a survey and bilateral meetings with the Italian financial intermediaries that use AI-ML techniques for credit scoring, in order to verify the degree of development of the sector, the solutions applied, the expected benefits, and the awareness shown in dealing with the new risks and the measures taken to address them. Chapter 6 details the main findings from the empirical review conducted with the selected financial intermediaries. The main conclusions are as follows:

a. "*growing market"* — the application of ML to credit scoring is still limited but is growing, driven by expectations of higher forecast accuracy;
b. "*concentration in technological choices*" — among a wide range of technological solutions, intermediaries selected a small number of ML models and explanatory techniques that have the lowest implementation costs, thereby becoming market standards;
c. "*clear progress in accuracy*" — the evidence from the literature that ML models produce more accurate estimates than traditional statistical models is confirmed by the experience of the selected intermediaries;

d. "*some steps towards greater financial inclusion*" — intermediaries are starting to exploit the potential in terms of inclusiveness of AI-ML systems thanks to the access to alternative data sources that they provide and the development of models that can better exploit them: there is initial evidence of intermediaries expanding their credit supply to segments of customers that had been traditionally excluded due to their lack of credit history;

e. "*the analysis of discrimination bias is not widespread*" — the adoption of techniques to assess and, possibly, mitigate bias is still limited, as the risk of discrimination is perceived as remote and not currently present;

f. "*perception of many benefits and few risks*" — The feedback from intermediaries is generally optimistic: the use of AI techniques appears to bring many benefits and few additional risks compared with traditional techniques;

g. "*governance and controls are not fully calibrated with respect to the use of AI* techniques" — the governance of ML models appears in many cases not to be focused on the new risks associated with the adoption of AI-ML techniques, including with respect to the adequacy of reporting, staff involvement and training, and the management of outsourcing, which is sometimes used extensively.

2. Artificial Intelligence and main techniques in use

The term "Artificial Intelligence" (henceforth, for brevity, *AI),* originated in the 1950s, generally identifies different theories, methodologies and techniques that enable the design of IT solutions capable of reproducing human intelligence in various ways. This IT branch struggled over the years with some intrinsic technological limitations, in terms of computational capacity, capability of processing large amounts of data and maturity of algorithms. The rapid evolution of technology has gradually made available the necessary processing capabilities to support AI algorithms, which — thanks also to the wide availability of data — have begun to release their capabilities in industrial applications (e.g. the application of *image detection* techniques in production chains for quality check purposes) but also in the provision of services to individuals (e.g. the widespread use of reverse interfaces, such as *chatbots, voicebots or* virtual assistants, also on mobile devices)[1].

These techniques differ significantly from the traditional programming approach, in which a programmer defines an algorithm that is able, by programming a series of operations, to deterministically transform input data into the desired output data. In the AI approach, instead of focusing on formalizing a deterministic algorithm, a model (under an inductive or deductive approach — see *below*) is designed to capture the information needed to automatically derive "knowledge" from the available data. In this context, application development skills are integrated

[1] For an overview of the use of the AI in the banking sector, see CIPA-ABI (2021), "Report on IT in the Italian banking sector".

with those typical of the *data scientist,* a professional figure focused on data analysis and with significant knowledge of the *business* domain.

AI can be conceptually divided into two different approaches, resulting in different algorithmic techniques: the **inductive** approach and the **deductive** approach. In the inductive approach, the machine summarises its knowledge on the basis of the empirical observation of the data, learning from them through a generalisation process; in the deductive approach, the machine generates new knowledge on the input data by an inference process, on the basis of a formal representation of knowledge made using *knowledge representation and reasoning* (KRR) languages. The inductive approach is typically known as **Machine Learning (ML)**, while the deductive approach is known as **Automated Reasoning (AR)**.

The figure and the box below summarize and provide examples of the main differences between the inductive and the deductive approaches.

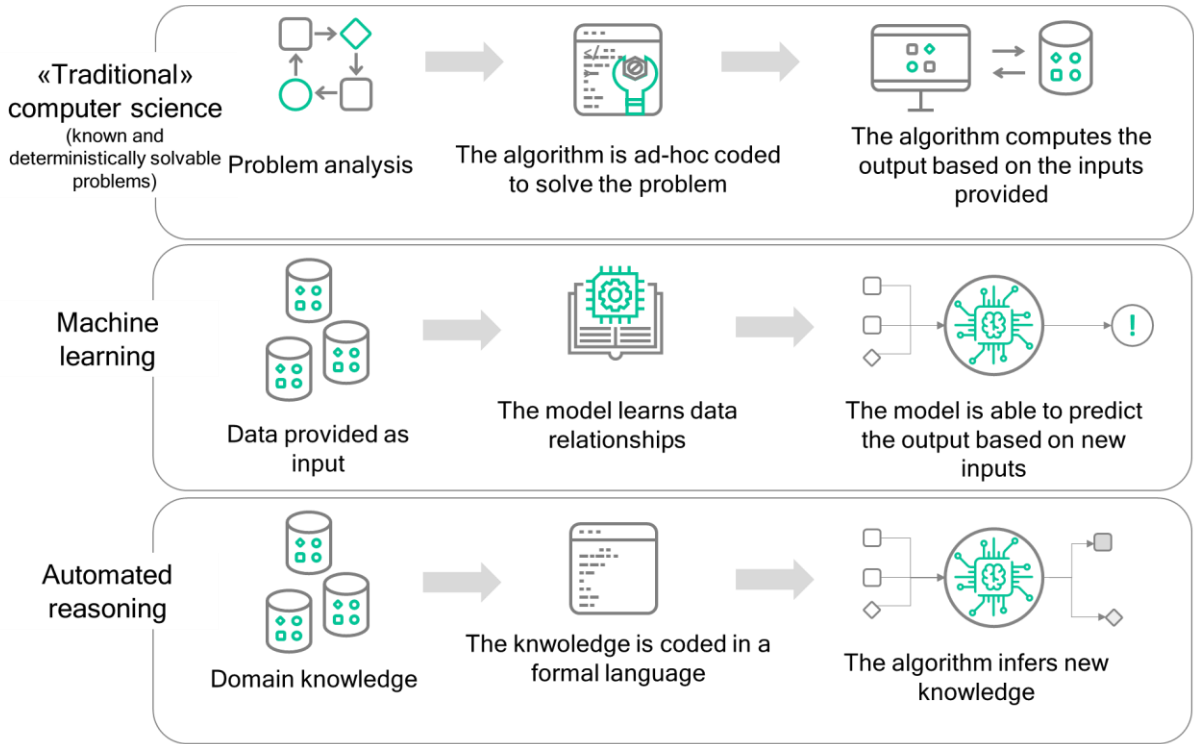

«Traditional» computer science
(known and deterministically solvable problems)
Problem analysis
The algorithm is ad-hoc coded to solve the problem
The algorithm computes the output based on the inputs provided
Machine learning
Data provided as input
The model learns data relationships
The model is able to predict the output based on new inputs
Automated reasoning
Domain knowledge
The knwoledge is coded in a formal language
The algorithm infers new knowledge

Figure1. Comparison between approaches in developing ad hoc deterministic algorithms and the adoption of AI techniques.

An example of the application of the two approaches to the AI

The AI approaches discussed above reflect a different way to address problems. To illustrate the difference, it is convenient to consider a so-called decision problem: in other words, we want to give the machine the responsibility to take a decision in a given situation (e.g. for a car, to decide whether to continue or exit; to grant or refuse a credit line) which is not known a priori.

In the inductive method the algorithm, which was previously trained on a large number of problematic and decision-making pairs (e.g. a set of travel choices or decisions on specific credit applications), will be able to take the decision by abstracting from the concrete cases analysed, for example by statistical models that approximate the dynamics of the decision based on the characteristics of the new situation.

In the deductive method, instead, the algorithm, which was previously trained on the characteristics of the domain of interest (i.e. how to behave while approaching an exit or according to which criteria a credit line is granted or refused), will apply the concepts applied o the specific case, providing the answer.

2.1 Machine learning and different approaches to automatic learning

There are different approaches to ML, which determine the characteristics of the algorithms and their requirements (e.g. in terms of quantity of data required for the training phase). The most known approaches are the so-called supervised, unsupervised, semi-supervised and reinforcement learning, in which:

- ***supervised***: the algorithm produces ("learns") the model of the relationships between inputs and outputs by means of a dataset, previously labeled by a human being (the labeled dataset);
- ***not supervised***: the algorithm produces the model autonomously from the dataset, without the need for it to be labeled;
- ***semi-supervised:*** an intermediate case between the two previous cases, where the dataset is only partially labeled (e.g. because the cost of labeling a dataset is usually high);
- ***reinforcement learning:*** the algorithm performs actions in a way that progressively maximises a profit function, assigning either a positive or a negative value to each round. This approach could be called "*trial and error*".

The supervised approach is often applied to classification problems, where datasets are already classified, so that the algorithm can capture relationships and apply them to new data. These techniques tend to produce more effective models with smaller data sets than the unsupervised approach, which is used for learning the inherent structure of the data without a formalized a-priori knowledge; a typical application is *clustering*, where the algorithm is able to split the data in sets containing similar information (e.g. to distinguish between human faces and cats images).

The following two families of algorithms, which are applicable across the above mentioned approaches, are particularly relevant for their intrinsic characteristics[2]:

- **Ensemble learning***:* for improving forecasting performance, ensemble techniques require the use of more than one model: the final forecast is then generally obtained as either an average or a majority voting of the models forecasts. A common case is ensemble learning applied to decision trees (see example in Figure 2) that operate through a progressive partition of the solution space. These techniques can be divided into two sub-categories: bagging-based (simultaneous learning of independent models, each characterised by the exploitation of portions of the overall information) and boosting-based (model sequences that progressively refine the learning process). Two popular examples for the aforementioned approaches, in the context of supervised learning, are, respectively, random forest and gradient boosting algorithms.

[2] Less popular ML applications based on: (I) Developmental algorithms, dynamic learning processes for the search for optimal solutions within parameter populations; and (ii*) Federated learning*, a set of learning techniques in a context of cooperation between parties, according to a distributed approach.

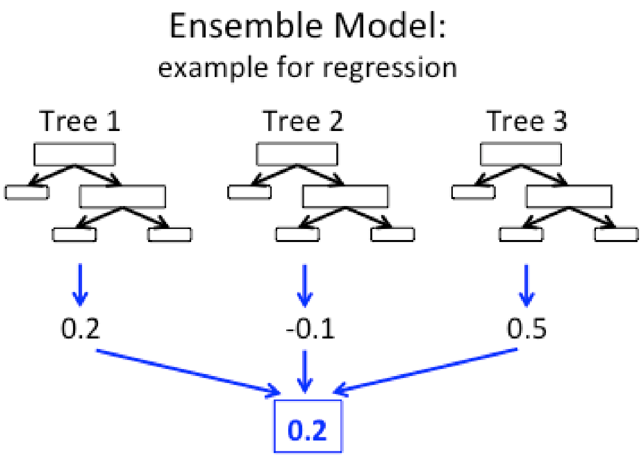


Figure2. Example of ensemble learning with random forest (with three trees and an output-based forecaster)

- **Deep learning**: this is a family of ML algorithms whose learning processes, inspired by the behaviour of the human neurons, are based on networks that interlink nodes organized at successive levels (i.e. *neural networks)*. Each level of the network corresponds to a learning phase of increasingly complex concepts:

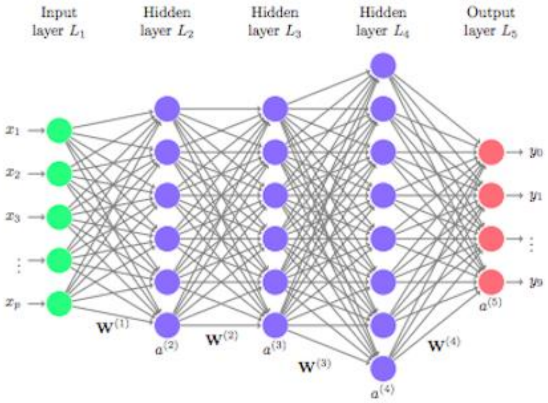


Figure3. Example of neural network.

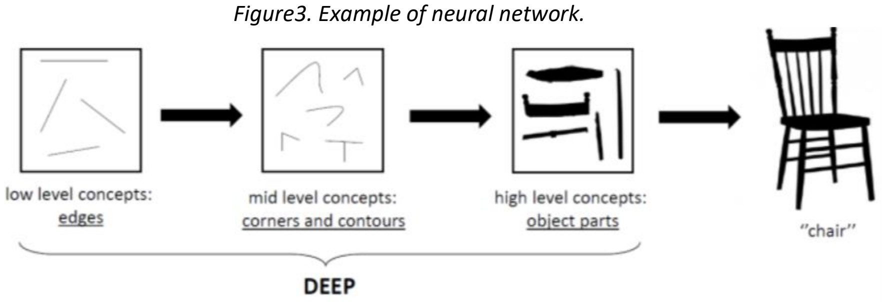

low level concepts: edges
mid level concepts: corners and contours
high level concepts: object parts
"chair"
DEEP

Figure4. Example of progressive learning of concepts in deep learning.

2.2 *Big data, analytics and AI*

In common language, “artificial intelligence” is often linked directly to “big data”; the two concepts are substantially different despite their several points of contact. The term big data usually identifies data with some key features which are described below, as well as the set of algorithms, technologies and IT solutions that support data collection, management and analysis. The paradigm associated with big data is usually referred as “5 V”, which identify three distinctive features and two expected practices of big data:

- Volume: data available for analysis are often in the order of terabytes or above;
- Variety: in addition to the more traditional structured data, this paradigm is complemented by semi-structured data (such as XML files) or unstructured data (such as textual documents or images);
- Velocity: as data are produced at extremely high rates, suitable technologies are needed to process them at an appropriate speed and/or for real-time analysis capabilities;
- Veracity: it should be ensured that data represent as closely as possible the underlying reality, as they can be affected by unreliability due to the nature of the processes used to generate and collect them;
- Value: data must be transformed into information useful for the business.

Therefore, data having at least one of the three features described above are defined as “big data”: high-volume (number of observations and/or number of attributes), high variety (format and/or content) and production/collection speed, that implies usage of non-traditional tools and techniques. As an example, big data definition can comprise large volumes of transactions and payments - characterized by fine-grained data - and textual data, such as the “reason” text field in money transfers as well as social network and internet navigation related data.

VOLUME
Huge amount of data
VARIETY
Different formats of data from various sources
VALUE
Extract useful data
VELOCITY
High speed of accumulation of data
VERACITY
Inconsistencies and uncertainty in data
BIG DATA

Figure 5. The “5 V” of big data.

The link between the big data domain and AI techniques is quite immediate: the availability of huge datasets (and the related technologies to efficiently process them) enables ML algorithms to “learn” better. However, such algorithms can be trained even without big data (assuming, however, that properties of the training datasets - e.g. in terms of dimensions and characteristics - are fit for purpose), so that big data should not be considered as a prerequisite for the application of AI techniques. Moreover, neither the complementary perspective implies a dependency: in fact several techniques (which fall within the commonly referred analytics domain) are applicable to big data without the need of AI algorithms. As a result, the two domains have a complementarity relationship, each benefiting from the progress made by the other.

2.3 *Explainable AI (XAI)*

AI techniques entail some challenges that usually are not present in the development of non-AI IT solutions: among them, it is particularly important the capability to explain, to stakeholders, the results produced by the algorithm, in order to understand the reasons behind decisions taken (or suggested) by the AI algorithm.

The term Explainable AI (XAI) refers to the tools applied by analysts to enrich the output of the ML algorithm with a set of interpretations and explanations about the model behavior.

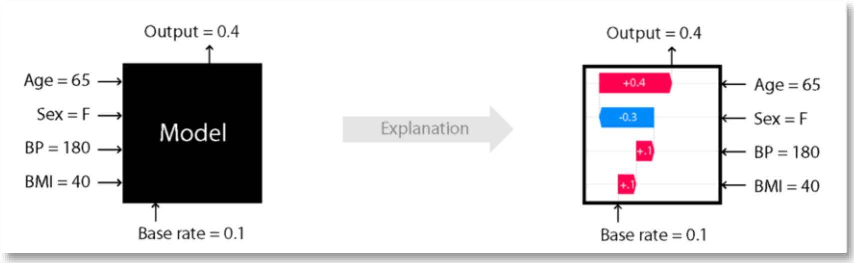
From a conceptual point of view, XAI can be splitted in two main approaches:

- **Interpretability***,* i.e. to track (from a quantitative point of view) the logic that governs the model's behaviour;
- **Explainability***,* i.e. to formulate qualitative assessments (justifications) of the results obtained, with the aim of explaining the model behavior.

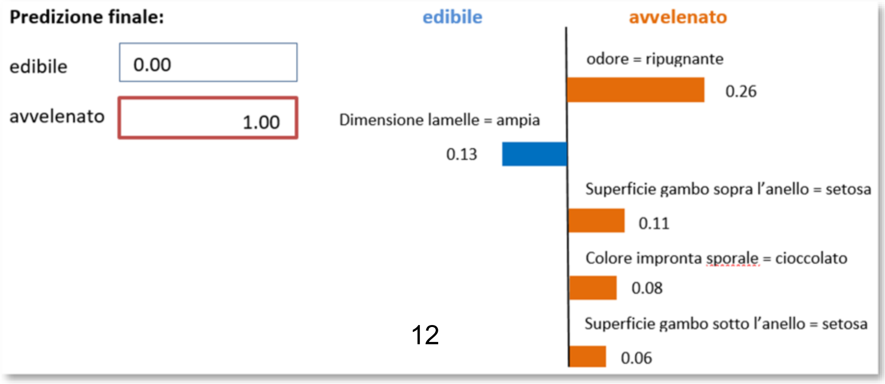
The degree of intrinsic interpretability of the various ML algorithms generally tends to decrease when the forecasting ability increases: for example, neural networks perform very well in terms of accuracy and can only be interpreted, to some extent, using XAI techniques; on the other hand, an inherently interpretable model, such as the decision tree, usually tends to be less performant.

From an implementation point of view, XAI techniques can be distinguished between those that can only be applied to specific classes of models (model-specific explainability), such as the "feature importance" of decision trees, and those applicable to any model regardless of its functional form (model-agnostic explainability), such as SHAP and LIME (see boxes).

Shapley Additive exPlainations (SHAP): model-agnostic explainability technique based on the *Shapley values*, concept originated in the field of game theory to measure how the payout should be fairly distributed among the players in a way that each player receives a share proportional to his contribution. These values represent the contribution of the model features (the individual measurable components) to the model output, as in the figure below:



Lime (Local Interpretable Model-agnostic Explainations): model-agnostic explainabilty technique which aims to explain the model behaviour with changes in the values of *the* feature and observation of the impacts on output (in terms of the contribution each feature makes to the final forecast):



The implementation of XAI techniques can be placed in different model lifecycle phases, namely: prior to model development, for a better understanding of the data used to train the model itself (*pre-modeling explainability*); during model development, to facilitate its interpretation (*explainable modeling*); after model development, to explain the model behavior (*post-modelling/post-hoc explainability)*.

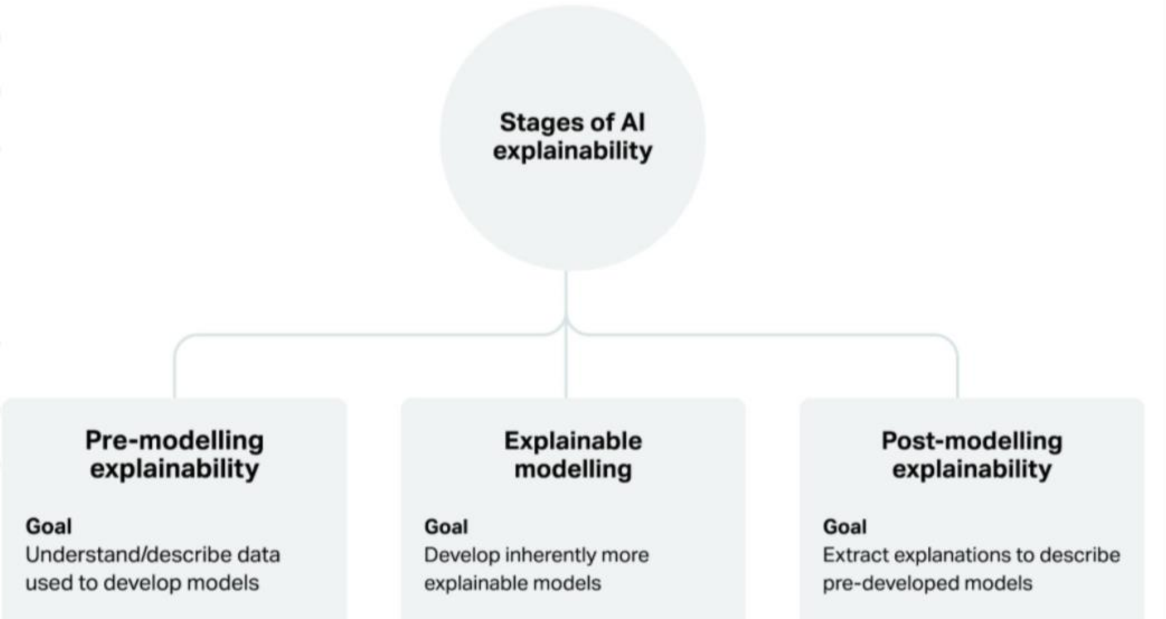

Stages of AI explainability
Pre-modelling explainability
Goal
Understand/describe data used to develop models
Explainable modelling
Goal
Develop inherently more explainable models
Post-modelling explainability
Goal
Extract explanations to describe pre-developed models

Figure 6. Different phases of application of XAI techniques.

3. **AI in the assessment of creditworthiness: benefits and risks, safeguards**

3.1 *The application of ML to credit risk assessment*

One of the areas in which the properties of ML techniques have been explored is credit risk assessment. Compared to the traditional statistical approach, based on econometric estimation of the probability of default (e.g. through logistic models), some key differences are highlighted. In econometric models, the choice of the relevant variables and the specification of the functional forms of relationships are generally guided by economic theory, while ML algorithms automatically select the relevant variables and identify even non-linear and difficult-to-interpret relationships between them.

Several studies have compared the ability of ML techniques to predict the default of firms and individuals with respect to traditional approaches. These studies show that the accuracy of ML models in identifying defaults is generally better than that of econometric models (Fantazzini and Figini, 2009; Khandani et al., 2010; Kruppa et al., 2013; Yuan, 2015; Barboza et al., 2017; Bachman and Zhao, 2017; Fuster et al., 2020; Albanesis and Vamossy, 2019; Moscatelli et al., 2020). The improvement in the accuracy of ML techniques stems primarily from the extension of the range of functional forms and relationships between the different variables evaluated by the model. A recent survey shows that the use of advanced ML techniques leads to an improvement in the accuracy metrics of default forecasts relative to traditional statistical models, mostly between 2 and 10 percentage points but in some cases higher depending on the study and the index employed (Alonso and Carbó, 2020)[3].

Another advantage of ML is the ability to manage and process large amounts of data in terms of volume (number of observations) and richness (number of variables, data types), taking advantage of the increased computing power that has become available in recent years. In general, the types of data sources employed in credit scoring methods have gradually increased (which in fact concerns both econometric models and ML techniques): structured financial data (asset, liability and other economic and financial indicators, indicators on financial flows and payments, market indicators) to structured non-financial data (socio-demographic data also coming from third-party sources), unstructured financial data (analysis of transactional information and information derived from open banking) and unstructured non-financial data (browsing data, digital footprints,[4]information provided on social networks) (Tobback and Martens, 2019; Óskarsdóttir et al., 2019; Berg et al., 2020; Roa et al., 2021).

Thanks to their specific process of training and calibration, ML models can exploit the above-mentioned alternative data sources, making it possible to consider variables that do not have a clear economic interpretation/relationship and would, therefore, not be considered in a traditional econometric model. Exploiting alternative or complementary information to that used in traditional models may increase forecasting accuracy while at the same time enabling the assessment of the creditworthiness of entities otherwise excluded because of lack of standard financial data. In addition, the adoption of ML for credit scoring may increase competition because entities other than traditional financial intermediaries can rely on new data sources to offer financing at lower costs, with shorter approval times, requiring less collateral and documentation (Jagtiani and Lemieux, 2017; Bazarbash, 2019).

From a risk perspective, ML applications are characterized by a trade-off between the increase in accuracy with the cost in terms of lower explainability of the model. Furthermore, the use of ML in customer selection may entail legal and reputational risks arising from the opacity of the processes and mechanisms that allow the better performance to be achieved. The use of XAI methods may weaken the trade-off between accuracy and explainability (Dutchino et al., 2022).

Finally, ML algorithms may produce highly accurate models when forecasting the variable of interest in the sample used, but might have low predictive ability in contexts other than the one in which they were developed (overfitting risk). The reliability of a ML model must be verified on the basis of external validity requirements, i.e. stability of predictive capacity for all individuals in the population (population validity), environments, locations or time periods (ecological validity) other than those considered when developing the model.

[3] A recent study on Italian data (Moscatelli et al., 2020) shows that the magnitude of the improvement in accuracy compared with logistic models varies depending on the type of model and the number of variables considered.

[4] Defined as the "single set of activities, actions, contributions and digital communications that can be found on the internet or digital devices". Source: Wikipedia.

3.2 *Automated reasoning applied to credit assessment*

The application of AR approaches to credit risk assessment are based on formal modelling of the knowledge held by the business entity through a logical-matematical formalism, often able to consider both quantitative and qualitative aspects. The IT systems for AR then automate the application of this knowledge to the individual credit applications.

There are a number of studies that have defined the theory and practice of using AR to assess creditworthiness. The basic technical and theoretical elements were defined with the adoption of the first expert systems (e.g. Zocco, 1985) and have effectively traced the link between the decision to grant a loan and logical coding techniques (Iwasieczko et al., 1986). Theoretical studies were soon complemented by dedicated systems, often designed to optimise performance (*throughput*) and credit accuracy. In the literature there is a well-known case of a leading credit card company, among the first to use an AR-based system for real-time assessment of "unusual" credit applications by cardholders. This approach has produced significant increases in the accuracy of previously human-based decisions (Piketty, 1987).

In more recent contexts, AR techniques are used to grant credit in specific sectors, where knowledge of the business area is highly specialised[5].

The recent introduction of more mature and user-friendly knowledge representations, such as those adopted for Knowledge Graph (Gottlob et al., 2015; Hogan, 2021, Bellomarini et al., 2018) and the adoption of more efficient hardware have led to an increasing interest in applying automatic thinking to the financial field., In particular, Knowledge Graphs are used to assess creditworthiness in a variety of research environments (Beydoun et al., 2020).

4. **Problems of incorrect differentiation and discrimination in the assessment of creditworthiness**

4.1 *Definitions*

A widely discussed topic in the literature on algorithm applications for the classification of units or individuals is the possible presence of bias. Within the assessment of creditworthiness framework, such bias may result in incorrect risk differentiation and discrimination against potential clients.

In general, incorrect risk differentiation is the result of the failure of a model to sort properly customers by their level of creditworthiness, i.e. the inability to achieve an optimal level of allocative efficiency. The use of a model not appropriately differentiating clients by their true riskiness may lead to bias both in the approval of loan applications and in pricing.

Discrimination, in general terms, refers to the presence of prejudice or favoritism towards specific individuals, or social groups, identified by attributes considered to be "sensitive" (Mehrbert et al., 2021). Discrimination is detected when bias affects decision-making processes and leads to

[5] One interesting case in this respect is ALEES (Bryant, 2001), an AR system for granting credit to agricultural operators, developed by Griffith University in Australia.

drawbacks, primarily in terms of personal damage. Discrimination may be direct or indirect. Direct discrimination, or disparate treatment, consists of decision-making processes or conduct that explicitly account for sensitive attributes to identify individuals as vulnerable. Conversely, indirect discrimination, or disparate impact, of one or more individuals does not explicitly identify the latter based on sensitive attributes, even though a *de facto* condition of inequality results from the process - for instance through proxy variables of vulnerable group's membership (Hjpgian et al., 2012)[6].

In order to assess quantitatively and prevent the occurrence of forms of discrimination, definitions of *fairness* available in the literature (Dwork et al., 2012) can be used. The concept of *fairness* makes it possible to measure and evaluate qualitatively the lack of discriminatory mechanisms within a complex system. In this respect, model behavior can be assessed at the aggregate level by adopting a definition of group fairness (e.g. vulnerable communities or groups) or at the individual level. The most common definitions of *group fairness* include: (I*) fairness through unawareness*, which entails preemptive exclusion of sensitive attributes from the database; and (II*) statistical parity*, which defines lack of discrimination as statistical independence between model decisions, conditional on an individual's membership to a group. Definitions of individual fairness include counterfactual fairness, which resorts to probabilistic reasoning to evaluate the performance of the model against counterfactual variations in specific individual characteristics. Once the definition of fairness that is most appropriate for the use case has been set[7], it shall be employed for assessment downstream of the estimation process or within the estimation process itself, as a constraint on the target function optimized by the learning algorithm.

Since the adoption of a specific definition of fairness typically implies a violation of the conditions for the other definitions to hold, the analyst is required to assess which definition could be most suitable in order to eliminate discrimination on a case-by-case basis.

The economic literature has widely investigated the presence of discrimination in the assessment of creditworthiness, mainly in the United States market. Therein, legislation that explicitly prohibits discrimination in lending are in place (Fair Lending). The analyses measure discrimination in terms of the probability of accessing credit and interest rate spreads applied to customers sharing the same characteristics, and find evidence of discrimination related to ethnicity or gender. Some works specifically analyze the link between the use of quantitative methods for credit scoring and the presence of discrimination with mixed results point to different directions. For a review of the literature, see Appendix 2.

[6] A classic example is the redlining, *according* to which individuals belonging to a given social group that can be geographically traced in a popular quartile of a large city suffer forms of discrimination due to the use of the post code in decision-making models.

[7] The adoption of the appropriate *definition of fairness* may depend on a variety of factors. These include the analyst's orientation to pursue an approach that is either of a broad-based or inclusive nature. In the former case, a negative assessment is made of favourable decisions benefitting groups or individuals assessed as not deserving, while the latter rewards the absence of adverse decisions affecting virtuous groups or individuals.

4.2 *Risk of biased assessment of creditworthiness: peculiarities resulting from the use of AI*

When using an AI system to support decisions, the presence of mechanisms leading to incorrect differentiation or discrimination, as well as having an impact on the level of performance, tends to generate a *feedback* cycle in which bias is confirmed and reinforced. Over time, for example, the systematic rejection of credit applications of specific groups within a population caused by an incorrectly specified model contributes to create historical bias in the data. This bias will be present in the samples extracted from the population that will be used to update the model, generating a vicious cycle.

Bias can arise across within each of the different phases of the development cycle of an AI algorithm: data collection, model specification and learning, and output analysis. In the case of ML techniques, it may be present in all the aforementioned phases; in the context of automated reasoning, bias is mainly due to an incorrect specification of the problem or its incomplete formalization[8].

In regard to bias generation during the data collection phase, application of ML techniques to traditional data and to big data ought to be distinguished.

In the first case, the adequate representativeness of the sample is in principle ensured by the adoption of consolidated statistical approaches for sampling from finite populations. Nevertheless, bias may occur due to the fact that data are generated, collected and/or processed by humans (Ntoutsi et al., 2020), resulting, among others, into the over-representation or under-representation of specific groups. In this regard historical bias comes into play, i.e. the circumstance that entities belonging to a specific type, heterogeneous group or social group to whom access to credit has in the past been hampered due to discrimination mechanisms, are under-represented in a sample in favor of categories to which credit was historically granted[9]. When the lack of representation of individuals with specific socio-demographic attributes depends on the presence of explicit mechanisms of systematic discrimination, the estimation process of a model will be affected by the so-called institutional bias[10]. Finally, bias can originate from the omission of relevant features in the data (*omitted variable bias*) or by the use of attributes serving as proxies for unobservable characteristics of individuals, which do not correctly reflect the observed phenomenon (*measurement bias*).

If ML techniques are applied on big data, statistical assumptions on the data generation process and data collection cannot be made or verified, and therefore statistical corrections cannot be applied. In principle, the use of big data should limit bias because of access to information with a high level

[8] Formalisation of the problem required by an AR system can help overcome the risk of bias caused by a deviation in its understanding.

[9] See Appendix 2 for evidence on this issue in the economic literature.

[10] In the context of the credit assessment, an example may be the presence — in the training database — of characteristics systematically identified as favourable or indispensable for the granting of credit (such as the minimum number of years of work experience, an attribute related to the applicant's age). This systematic distortion of the data can result in forms of discrimination, which are "institutional" embedded in the approaches/rules taken by intermediaries in granting credit.

of granularity (detail of information, frequency over time) and detail (number of attributes available for each individual), referring to the whole population and not limited to a potentially unrepresentative sample or to a historically biased sample (Oskarsdóttir et al., 2019). In the real world, however, big data collection processes are still subject to a range of possible sources of bias. In particular, the tendency to self-selection and the natural attitude of individuals to adopt different behaviors[11] and languages depending on the contexts (*Behavioral and Content Production bias*) may seriously undermine the validity of the model, as well as the soundness of the results, and leading to discriminatory mechanisms.

The use of big data also exposes the analysis to self-selection bias resulting from the digital gap between individuals[12], whose access to the internet and new technologies helps to define the intensity of the process of continuous data generation *(digital footprint*). As a result, some individuals generate digital footprints that allow ML systems to profile them, while others are less represented in the digital world (*thin-file)* if not completely invisible (*no-file)*. When ML models are trained on data sources with a significant presence of thin-files and no-files*,* the selection remains exposed to risks of incorrect differentiation of creditworthiness and discrimination. In addition, the use of behavioral data exposes the credit granting process to potentially manipulative dynamics by consumers with high digital skills (Freeman et al., 2017; Calo, 2013).

Even if the data collection process is unbiased, the incomplete specification of the problem can lead to "algorithmic" bias, starting with the type of model used and the related learning process (Baeza-Yates, 2018).

In its different manifestations, algorithmic bias can be generated by multiple factors, including:

i) specification of a functional form that is inadequate to describe the phenomenon being investigated, with consequent underfitting;

ii) failure or inadequate consideration of the different types of error in the model results[13];

iii) omission of attributes critical to understand the phenomenon, to reverse its assessment and for its interpretation *(Simpson's paradox*)[14];

iv) inclusion of significant variables acting as proxies for the identification of an entity as vulnerable (*included variable bias*).

[11] For example, the analysis of spending habits using the set of transactions recorded on a given mobile application will be representative of the population using the smartphone to make payments. The socio-demographic characteristics of this group do not relate to the population as a whole or even less to general spending habits.

[12] It is also known as "*Big Data Exclusion"*.

[13] For example, the cost function in terms of mispricing of defaults must adequately reflect whether false positives are considered to be "more severe" than false negatives or vice versa.

[14] The omission of highly correlated attributes with the variables whose behaviour is sought may prevent the correct assessment of the behaviour. For example, the omission of age information may lead to a biased assessment of the mechanisms described by the model when considering the history of payments from a large older population (D'Alessandro et al., 2017).

Further forms of bias in the use of ML algorithms can be caused by wrong conclusions drawn during output analysis, e.g. due to the lack of explanatory power of the model. The analyst may: (I) draw biased conclusions with reference to groups whose identification is based on incorrectly identified attributes (*aggregation bias)*[15]; (II) formulate causal considerations from distinct cohorts of a cross-sectional sample of a population *(longitudinal data fallacy*)[16]; (III) interpret the relevance of an attribute to the derivation of output as evidence in favor of a causal relationship between the two (*cause-effect bias*).

Appendix 1 reports a brief overview of the techniques that can be applied to identify bias and Appendix 2 a survey of the economic literature investigating discrimination in credit markets, and in particular in connection of the adoption of ML-based credit scoring.

5. Legal/institutional framework

This section explains the institutional and regulatory framework within which the possible use of AI/ML models for credit scoring by intermediaries can be included. In this context, both prudential and customer protection regulations are relevant, including more general provisions protecting data confidentiality. The main guidance emerging in international and national fora on the application of AI/ML techniques to credit scoring is also reported.

5.1 *Inadequate risk differentiation of credit scoring models within the prudential regulatory framework*

There are no specific requirements in prudential regulations relating to the use of AI and ML, or to prevent the application of such techniques. However, the guidance contained therein is applicable and valid irrespective of the model estimation techniques used by intermediaries.

With regard to credit risk, the regulatory framework is primarily represented by the provisions on corporate governance, internal controls and risk management contained in the prudential regulations[17] which set out the principles to be followed to ensure the effectiveness of the relevant processes. With regard to the criteria for granting loans, it is envisaged that, during the preliminary assessment phase, intermediaries will collect all the information necessary to assess creditworthiness using scoring or rating systems. There are no specific requirements regarding the characteristics that such systems need to have, provided that they "provide detailed indications of the level of customer reliability".

[15] For example, the identification of a higher credit deterioration preparedness among firms belonging to a given geographical region may lead to biased assessments where the indication of economic activity that may be significantly related to the output variable of the model is not taken into account, although it is present.

[16] For example, the deterioration in credit across a set of entities can be attributed to their decline in purchasing power over time, rather than being correctly searched for by group characteristics that can be detected transversely.

[17] Circ. 285 and 288 of the Banca d'Italia.

The EBA Guidelines on "*loan origination and monitoring*", recently transposed (20 July 2021) into the national regulatory framework as supervisory guidelines[18], define more specific guidance[19] for the use of "automatized" credit assessment models, including: (I) understanding the assumptions used in the model; (II) internal policies and procedures detecting and preventing bias and ensuring the quality of the input data; (III) the need to assess the adequacy (in terms of traceability, auditability, robustness and resilience) of the inputs and outputs of the models; (IV) the existence *of* policies ensuring that the quality of the model output is regularly assessed; (v) the existence of mechanisms for evaluating the output for the purposes of any *"override*", which incorporates expert judgement; VI) the availability of adequate documentation on the development of the models and their use[20].

By contrast, a more comprehensive and prescriptive regulatory framework is defined for banks seeking authorisation from the supervisory authority to use internal models to measure credit risk for the calculation of capital requirements[21].

The relevant prescriptions, although not binding for models that are not used to calculate capital requirements, may nevertheless provide a useful reference for identifying *best practices* to be followed also for the development and management of AI/ML models.

The main regulatory reference is the *Capital Requirement Regulation* (CRR — Regulation (EU) No 575/2013) accompanied by technical standards and guidelines prepared by EBA. With regard to significant banks, the ECB also published a guide on internal models[22] in which it clarifies how it intends to apply the regulatory framework set out in the CRR and in the *EBA framework*.

The above-mentioned rules include requirements in terms of both more general and process aspects, as well as the quantitative aspects of internal models.

The general aspects relate to the integrity characteristics of the rating assignment process, the effective use of models in business processes, their documentation, the storage of data, the set up of a validation process for internal estimates and the existence of an appropriate internal *governance* and control framework.
By their nature, these requirements tend to be cross-cutting with respect to the techniques used to estimate risk parameters and thus models, applying irrespective of their characteristics and the technologies used for their implementation.

[18] "Intermediaries shall make every effort to comply with the Guidelines" (see footnote 13 of 20 July 2021 — Implementation of the European Banking Authority's Guidelines on loan origination and monitoring (EBA/GL/2020/06))
[19] The guidance listed below is included in a section of the Guidelines only applicable to banks and not to financial intermediaries.
[20] See paragraphs 53, 54 and 55 of the EBA-GL-2020-06 on loan origination and monitoring.
[21] In this respect, the EBA published on 11 November last a Discussion Paper, proposing recommendations for the use of AI systems in the context of internal credit risk models used for the calculation of capital requirements (IRB).
[22] ECB Guide to Internal Models, EGIM.

General requirements for internal models

Among the rules that have a more direct effect in relation to the characteristics of the ML models, it should be noted that the intermediary must have “a process for vetting data inputs into the model, which includes an assessment of the accuracy, completeness and appropriateness of the data” (Article 174 of the CRR). It is worth mentioning also the reference to the need for the intermediary to complement “the statistical model by human judgement and human oversight to review model-based assignments and to ensure that the models are used appropriately”. The intention is that the intermediary is in a position to detect and limit errors arising from model deficiencies.

In relation to the data topic, the requirements regarding documentation of rating systems (Article 175 of the CRR) are relevant: in any case, it must provide “a detailed outline of the theory, assumptions and mathematical and empirical basis of the assignment of estimates to grades, individual obligors, exposures, or pools, and the data source(s) used to estimate the model”.

With regard to the topic of models provided by third parties, it is reaffirmed that the principles relating to the documentation of rating systems should in any case be met even where the vendor “refuses or restricts the access of the institution to information pertaining to the methodology of that rating system or model, or underlying data used to develop that methodology or model, on the basis that such information is proprietary”.

The most important requirements regarding the quantitative aspects of model development, i.e. the determination of the procedure for differentiating risk across borrowers, are based on five main areas: (I) data quality; (II) use of appropriate drivers; (III) rating philosophy; (IV) data representativeness; v) ability to differentiate risk and assess model performance.

Quantitative aspects of internal models.

I) Quality of model input data. — The institution shall have in place a process for vetting data inputs into the model, which includes an assessment of the accuracy, completeness and appropriateness of the data (art 174(b) CRR). This general requirement is particularly relevant for ML models which typically rely on a significant amount of data from both internal sources and external providers.

(II) Use of appropriate risk drivers. — The estimates shall be plausible and intuitive and shall be based on the material drivers of the respective risk parameters. (Article 179(1)(a) CRR). Verifying that the estimated parameters are in line with their expected economic behavior is particularly important for ML models that base their estimates on non-linear relationships by considering a high number of risk factors, thus presenting challenges in terms of explainability.

(III) Rating flexibility. — For each model, institutions should choose an appropriate philosophy *(point-in-time — PiT vs Through-the-cycle - TTC*) underlying the assignment of obligors or exposures to grades or pools and assess the adequacy of the resulting characteristics and dynamics of the rating assignment and risk parameter estimates. Typically, the ML models tend to introduce *PiT* elements into the estimates that could be inconsistent with the rating philosophy objectives identified by the institution; this may occur in an unchecked way owing to the lack of accountability of the ML.

(IV) Data presentation. — The regulation requires that the data used to build the model shall be representative of the population of the institution's actual obligors or exposures; (Article 174(c) of the CRR). In particular, the ability of the model to differentiate the creditworthiness of borrowers should not be hampered by the lack of representativeness of the data used. The characteristics of the ML models, in particular if they include the use of big data, make it difficult to ensure that a lack of representativeness of the data does not result in an incorrect differentiation of borrowers.

v) Risk differentiation and verification of the model's performance. — A rating system shall take into account obligor and transaction risk characteristics and the process of assigning exposures to grades or pools shall provide for a meaningful differentiation of risk l, i.e. it must assign lower levels of risk to debtors with higher creditworthiness (art. 170 CRR). In this context, in the review of estimates process, the institution is required to assess the model's ability to differentiate risk, comparing the performance measured in the model estimation sample with those obtained from more recent portfolios (para. 218 EBA-GL-2017-16). ML systems are exposed to the risk of overfitting that *may* undermine the discriminatory power of the model.

5.2 *Discrimination in European and national legal frameworks and sectoral legislation*

A fundamental principle of equality and non-discrimination is enshrined in Article 3 of the Constitution: "All citizens shall have equal social dignity and shall be equal before the law, regardless of gender, disability, language, education, political views, personal and social conditions". Italian legislation also contains some specific definitions of discrimination, such as gender or gender discrimination or people with disability, which are based on the provisions of the European Directive on equality and employment (2000/78/EC)[23].

5.2.1 Transparency regulation

The Consolidated Banking Law and the transparency provisions contain general references to discrimination that, while limited to issues of access to payment services, introduce a definition of what can be considered discriminatory in the banking and financial context and restate the link between financial inclusion and non-discrimination[24].

With specific reference to the transparency provisions on consumer real estate lending, Article 120-undecies of the Testo Unico Bancario (TUB i.e. Consolidated Law on Banking), on the assessment of creditworthiness, states in paragraph 1 that "...The assessment of the creditworthiness is carried

[23] In our context, the Consolidated Law on migration appears to be relevant, specifying in Article 43(2) that an act of discrimination is committed "When more unfavourable conditions are imposed, or there is a refusal to supply goods or services offered to the public only because of his or her condition as a member of a particular race, religion, ethnic group or nationality".

[24] The TUB (Article 126-noviesdecies) states that "all consumers legally residing in the European Union, without discrimination and irrespective of their place of residence, have the right to open a basic account...". The transparency provisions on payment services (Paragraph 5.2 of Section VI) state: "Any changes in interest or exchange rates shall be applied and calculated in such a way as not to create discrimination between customers. The manner in which these changes are applied and calculated shall be assumed to be non-discriminatory when intermediaries adopt them on the basis of objective and reasoned criteria which apply to all clients on equal terms."

out on the basis of information on the economic and financial situation of the consumer that is necessary, sufficient and proportionate and appropriately verified", thus imposing requirements for the characteristics of the data used by the intermediary for the creditworthiness assessment.

Subsection 5 of that Article also states that "If the credit application is rejected, the lender shall inform the consumer without undue delay of the refusal and, where appropriate, of the fact that the decision is based on automatic data processing", highlighting a right of the consumer to be informed if the refusal of his application for real estate credit was based on the use of algorithmic credit scoring[25].

As in other sectors of the economy, non-discrimination must be combined with the freedom of enterprise of private firms, such as intermediaries, in granting or not credit. The latter principle has also been reiterated on several occasions by the Banking and Financial Ombudsman, who stated that "as a general rule, there is no obligation on intermediaries to grant credit or to review the terms on which credit was granted, without prejudice to the obligation, during the assessment and acknowledgement of any renegotiation requests, to observe the principle of fairness in contractual relationships...the assessment of creditworthiness falls within the scope of the entrepreneurial autonomy of intermediaries"[26].

5.2.2 Privacy

European privacy legislation (GDPR) deals with issues that are of particular relevance for the use of AI-ML systems for credit scoring: profiling, consent to data processing, and the right of access to the logic of an automated decision.

Article 4 defines profiling[27], while Article 9 sets out a general prohibition on the processing of sensitive data[28] for automated decision-making purposes, unless there is an explicit consent of the subject for one or more specific purposes, or if there is a public interest.

While there is a general prohibition of automated decision-making by an individual, Article 22(2) states that, in addition to the case of explicit consent to processing, automated processing is possible if necessary for the conclusion or performance of a contract between the involved parties.

[25] The recent proposal for a revision of the Consumer Credit Directive, published by the European Commission in July 2021, introduces similar forecasts also for consumer credit. The obligation to inform consumers would also apply in the case of personal offers based on profiling techniques.

[26] 2018 ABF Annual Report, pg. 53.

[27] "*Any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements".*

[28] Defined in them as data relating to the place of origin, whether or not; political views; philosophical or philosophical beliefs; trade union membership; (b) (c) (d) (d) medical, health -related; which relates to the gender's lives or the gender orientation of the person; judicial.

A more extensive right to information than that contained in Article 120-undecies of the TUB is granted in Article 15 of the GDPR concerning the right of access of the party concerned, which in paragraph (1)(h) grants the right to information on “the existence of an automated decision-making process, including profiling... and, at least in such cases, significant information about the rationale used, and the importance and expected consequences of such processing for the person concerned”. According to some commentators, the request for “significant information on the rationale used” implies an obligation for intermediaries to provide so-called “local explanations” to the applicants, i.e. including details on the main variables that contributed to the specific score (see Hacker and Passoth, 2021).

5.3 *The position of national and international institutions on AI and the European AI Regulation proposal*

Apart from the guidance provided in the positive legislation outlined above, a number of national and international authorities have expressed positions on the use of the AI.

Among others, the following are worth mentioning with reference to the subject under consideration:

1) the European Commission, which has expressed its assessments in communications - COM(2018) 237 and COM(2018) 795 - and has drawn up two specific documents (the “Ethic guidelines for trustworthy AI” [29] and the “Policy and investment recommendations for trustworthy AI”) [30] and relevant reports (Report of Expert Group on Regulatory Obstacles to Financial Assistance — ROFIEG[31]);
2) the Financial Stability Board, which analysed the possible financial stability implications of the use of AI-ML in financial services in the FSB report, 2017[32];
3) the EBA, which has carried out a more in-depth analysis of the use of the big data and aggregated Analytics (BD &AA) in the banking sector in the context of its FinTech Roadmap[33];
4) the Italian Ministry of Economic Development (MISE), which set out the guiding principles for the introduction of the AI across sectors of the Italian economy in the National Strategy for Artificial Intelligence (MISE, 2020) [34].

[29] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[30] https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence

[31] https://ec.europa.eu/info/files/191113-report-expert-group-regulatory-obstacles-financial-innovation_en

[32] https://www.fsb.org/2017/11/artificial-intelligence-and-machine-learning-in-financial-service/

[33] https://www.eba.europa.eu/eba-publishes-its-roadmap-on-fintech.

[34] https://www.mise.gov.it/index.php/it/strategia-intelligenza-artificiale/contesto

Central banks have also started reflections and surveys about the use of AI by the financial industry over recent years. Among the contributions of different authorities (ACPR[35], Bafin[36], Bank of England[37]) are worth mentioning:

1) the Monetary Authority of Singapore (MAS), which in 2018 set out four principles for the use of AI and data analytics based on an approach called FEAT (fairness, ethics, accountability and transparency)[38]; starting from this logical framework in 2020[39] and 2022[40], a methodology has been developed in order to verify the alignment with these principles;

2) the Dutch Central Bank (De Nederlandsche Bank — DNB) added two additional principles to the scheme proposed by the MAS, developing a six-step approach called SAFEST (soundness, accountability, fairness, ethics, skills and transparency), accompanied by 17 guidelines to make the above principles applicable[41].

3) the Hong Kong Monetary Authority has developed some "supervisory guidelines" for banks that intend to apply AI techniques in their business models[42]. These guidelines are divided into three areas: model risk management[43], consumer protection[44] and cyber security[45].

At the European level, the draft Regulation on the European approach to AI (Artificial Intelligence Act), which, although only at the beginning of its legislative process[46], provides a first regulatory definition of artificial intelligence, outlines the necessary safeguards for the management and verification of AI systems and introduces important points of contact with GDPR (General data protection regulation). Of particular interest for the purposes of this work is the specific reference to AI credit scoring systems and the associated risk of harmful effects.

At the general level, the draft aims to regulate the AI in order to allow for its smooth development and use in the EU, while protecting citizens from possible practices detrimental to their rights[47]. Some applications or practices based on AI technology are considered by the Regulation and are therefore explicitly prohibited (e.g. systems that use manipulative techniques or assign social scores according to people's behaviour), while other uses of the AI are considered to be high-risk and

[35] https://acpr.banque-france.fr/en/governance-artificial-intelligence-finance
https://acpr.banque-france.fr/en/acpr-tech-sprint-explainability-artificial-intelligence

[36] https://www.bafin.de/SharedDocs/Downloads/EN/dl_bdai_studie_en.html
https://www.bafin.de/SharedDocs/Downloads/EN/Aufsichtsrecht/dl_Prinzipienpapier_BDAI_en.html

[37] https://www.bankofengland.co.uk/research/fintech/ai-public-private-forum

[38] https://www.mas.gov.sg/publications/monographs-or-information-paper/2018/FEAT

[39] https://www.mas.gov.sg/schemes-and-initiatives/veritas

[40] https://www.mas.gov.sg/news/media-releases/2022/mas-led-industry-consortium-publishes-assessment-methodologies-for-responsible-use-of-ai-by-financial-institutions

[41] https://www.dnb.nl/media/voffsric/general-principles-for-the-use-of-artificial-intelligence-in-the-financial-sector.pdf

[42] https://www.aof.org.hk/docs/default-source/hkimr/applied-research-report/airep.pdf

[43] https://www.hkma.gov.hk/media/eng/doc/key-information/guidelines-and-circular/2019/20191101e1.pdf

[44] https://www.hkma.gov.hk/media/eng/doc/key-information/guidelines-and-circular/2019/20191105e1.pdf

[45] https://www.hkma.gov.hk/eng/key-functions/international-financial-centre/fintech/research-and-applications/cybersecurity-fortification-initiative-cfi/

[46] The draft presented by the Commission is currently being brought to the attention of the European Parliament and the European Council.

[47] The rules will also apply to non-European companies using AI systems vis-à-vis European users.

therefore subject to a list of safeguards and controls. The AI credit scoring systems of individuals fall into the latter category because of their impact on individuals' lives and the risk of introducing or perpetuating discrimination dynamics in the assessment of the creditworthiness of individuals.

Providers of high-risk AI systems must meet a number of requirements (see below) and establish a control system that guarantees the quality of the service provided and the management and mitigation of risks over time[48]. The vendors of these AI systems must also provide to the national competent authority identified by the Member State, upon request, all the information and documentation necessary to demonstrate the system's compliance with the Regulation.

The requirements for high-risk AI systems include:

1. the use of relevant, representative, complete, free of errors datasets with appropriate statistical properties;
2. the existence of methodologies and practices for data management, model training and validation to ensure the assessment of possible bias;
3. datasets that take into account the characteristics of the specific geographical, behavioural or functional setting within which the AI system is intended to be used;
4. traceability of the AI system's functioning, enabled by the automatic recording of events ("logs");
5. the adequacy of documentation and appropriate forms of transparency for users who must have available concise, complete, accessible and comprehensible information in order to enable them to interpret the system output;
6. appropriate level of accuracy, robustness and cybersecurity;
7. the presence of a level of human control over AI systems, carried out by competent and expert individuals, which should be able to understand, monitor and when necessary to intervene on the systems, possibly deciding to ignore the output they produce.

6. Analysis of survey results

A number of supervised entities have been selected to further analyse the adoption of AI techniques in the context of credit risk management within the Italian banking and financial landscape.

The pool consists of 10 financial intermediaries, both banks and non-banks, of different sizes and adopting business models, which — on the basis of the information available — are testing, developing or using statistical models based on AI-ML techniques in their credit process.

This section reports not only the analysis of responses to the questionnaires, but also some qualitative considerations that emerged during bilateral meetings with the intermediaries.

The main findings of the analysis are as follows:

[48] For supervised intermediaries, this oversight must be integrated with the internal control system provided for in the legislation.

- The use of AI methods in the assessment of credit risk is not yet widespread but is growing: the 10 respondents indicated that they had developed a total of 38 models, of which around 60 per cent were already in use at the time of the survey. Most of the models are aimed at corporate/SME clients.
- In almost all cases, the model scores are provided to support the assessment of creditworthiness by analysts, who are responsible for the final decision. However, some respondents have declared their intention to gradually reduce human intervention in the lending process in the future.
- The main expected benefit that has prompted intermediaries to switch from traditional methods to AI is the improvement in forecasting accuracy.
- Other benefits cited by some intermediaries are the possibility of implementing instant lending processes and exploiting alternative data sources (something made easier by ML models), which would make it possible to select customers with limited credit history more effectively, expanding the range of potential customers.
- In most cases, the models reported in the survey use financial data from internal sources or purchased from analytics service providers; the use of data on current account movements, also from open banking, is widespread. By contrast, the use of web-based and social media data is extremely limited.
- Around 90 per cent of the methods developed are based on combinations of trees (Gradient Boosting Trees, Random Forests). According to the intermediaries surveyed, this choice is due to the greater simplicity of implementation and the optimization of the trade-off between accuracy and explainability afforded by these models.
- Almost all banks have adopted or intend to adopt explainablity techniques to make the model's decision process more transparent. The most common techniques were Shapley Values and Feature Importance, i.e. post-hoc explanation techniques that explain the rationale of an already trained and calibrated model.
- There was no use of explicit bias-reducing techniques, such as balancing the dataset, controlling for historical bias or conducting causal analysis of the results.
- Respondents have adopted or are planning to adopt a definition of fairness in just under half of the models; this share rises to two thirds in the case of models intended for retail customers. All intermediaries that have adopted a definition have opted for fairness through unawareness, which removes attributes explicitly considered sensitive (such as the gender or age of retail customers) from the database, but does not take into account the effects on the model induced by the presence of variables potentially correlated with the sensitive attributes.
- All intermediaries reported that they have set up or plan to set up a model governance framework, supported by specific reports. Nevertheless, the monitoring process covers issues relating to the quality and integrity of input data in just over half of the models analysed.

- Frequently, respondents have indicated that the skills needed for the development, maintenance and risk control of the ML models were unavailabile internally and that they resorted to outsourcing, sometimes in full.

6.1 *Overview of models*

The total number of AI models analysed is 38. In all cases, the AI approach adopted is inductive and encompasses ML techniques, in some cases associated with natural language processing techniques. No deductive automated reasoning (AR) techniques were reported by the sample of intermediaries covered by the survey. The preference for ML techniques can be attributed to their widespread adoption and to the market availability of technological solutions (model training and, in some cases, pre-trained solutions) and professional services (model development skills and tool knowledge).

Algorithms are applied in different phases of the credit risk management process: 58 per cent in the granting phase, 26 per cent in the monitoring of loans phase and the remaining 16 per cent in other phases (pricing, determination of credit risk adjustments, debt recovery or other activities carried out by internal control functions). Six models have been developed for future use in the calculation of prudential requirements.

Intermediaries have already deployed 61 per cent of the models, whereas 13 per cent are currently being tested, and 26 per cent are still under study or under development; almost all of the models being developed will be used for the credit granting phase (see Figure 7).

In most cases, the scores produced by the models are provided to support the assessment of creditworthiness by analysts, who are still ultimately responsible for the final decision. There are two cases that are exceptions to this: in the first case, the human contribution is activated only in the presence of a customer dispute; in the second case, the model is aimed at developing a fully automated instant lending process, albeit for the granting of very small loans. Nonetheless, some lenders report that they intend to gradually reduce human intervention in the lending process in the near future, once the experimentation and first-use phases have been completed. This would facilitate, for instance, the automated granting of a loan in case the model produces a positive assessment (instant lending).

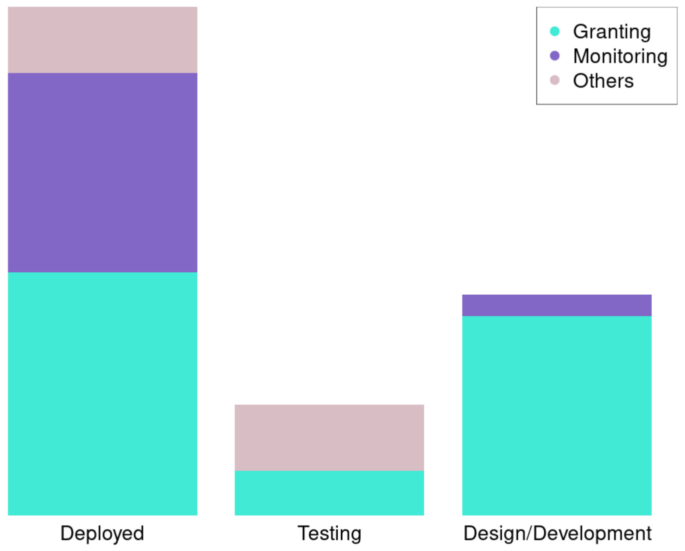


Figure 7. Status of the models by stage of the credit management process

Two thirds of the models serve credit processes targeting corporate or SME customers, whereas one third are meant for retail customers. Focusing only on the most critical phase of the loan granting process, half of the models address customers in the retail segment*,* 5 of which are currently in production, while the other half address the corporate/SME segment (see Figure 8).

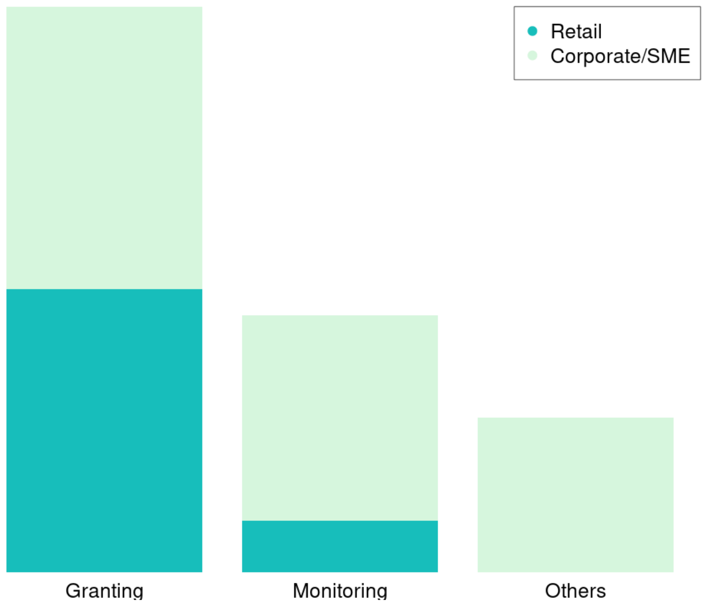


Figure 8. Management phase of the model's credit process by type of customer

6.2 *Model specification and development*

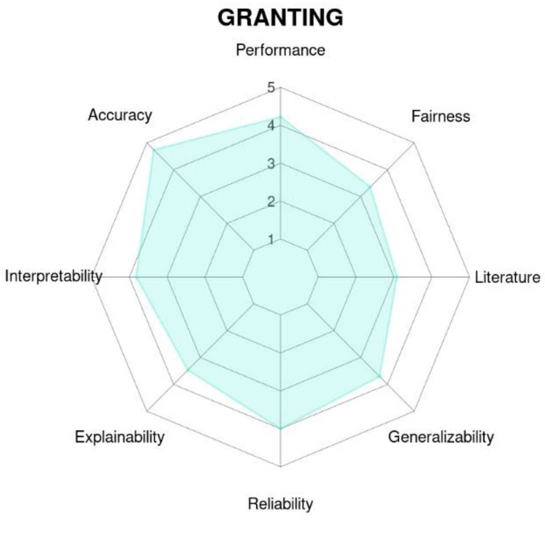
6.2.1 Data sources employed

On average, each model is developed employing four different data sources, with a minimum of one and a maximum of 12; in most cases, the data comprise financial attributes in structured format (e.g. capital and economic or financial indicators). About one third of the models also exploit non-financial data, such as: socio-demographic information on retail customers (available internally or declared by the customer during the credit application process) relating to their business or to relations with other companies, in the case of corporate customers. The use of unstructured data sources, i.e. text, is limited; some intermediaries use ML classification techniques to process the text from payment descriptions.

In most cases, the data are either internal or purchased from analytics service providers that are active in the credit market. Half of the respondents use data on bank account transactions, including those obtained from open banking. Only two intermediaries are testing the use of data from either the web or social media. The first intermediary has used such information in some experimental analysis but will likely no longer do so. The second intermediary developed an ML module to process reviews of companies available on web platforms in order to refine the assessments produced by a rating model; although model predictions currently achieve low levels of accuracy, the intermediary believes that, going forward, the module could provide an increasingly significant contribution, thanks to the growing availability of data.

6.2.2 Model selection and development

The model architecture has been generally chosen from two to five classes of candidate models. The accuracy of the forecasts and the statistical performance criteria have been particularly important in the selection process, especially in the case of lending models. Stable performance over time is also considered essential for the models used in monitoring. The explainability of decisions, the interpretability of the model and its external validity are deemed relevant for the choice among candidate models, regardless of the phase of the credit process to which the model might apply; by contrast, the absence of discriminatory mechanisms and the dissemination of the model in the literature are generally considered less relevant.



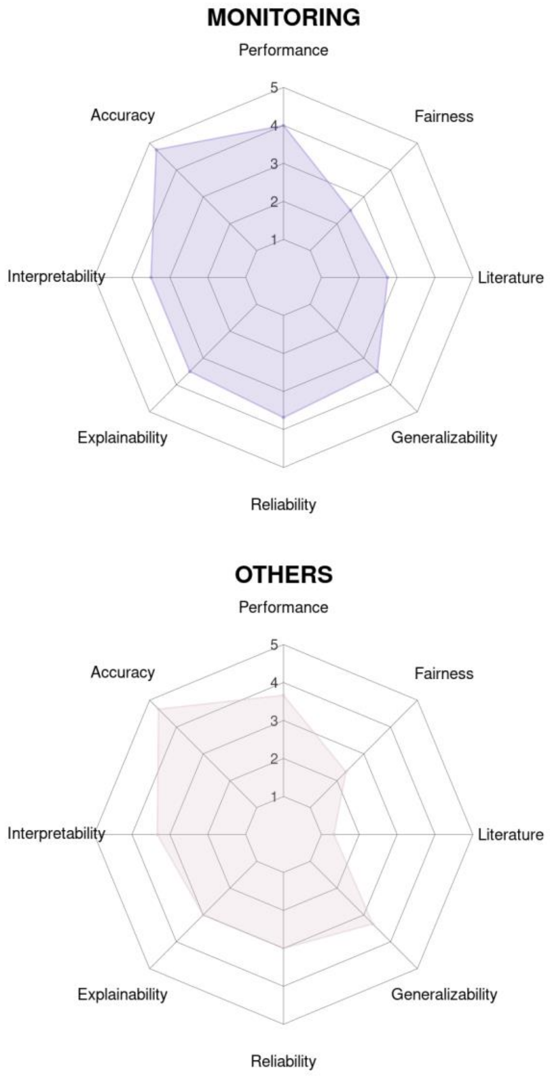


Figure 9. Criteria for the selection of models (scale from 1 — not necessary, to 5 — essential)

With regard to accuracy, the use of ML models has allowed intermediaries to improve their performance significantly compared with the traditional approaches. In some cases, information on the extent of the improvement has been provided by respondents. However, the heterogeneity of the metrics employed prevents direct comparison. The results that were reported are in line with findings from the academic literature.

More than 95 per cent of the models rely on supervised learning techniques, mainly ensemble learning methods using combinations of decision or regression trees (e.g. Gradient Boosting and Random Forest). These are generally available from standard and open source software packages; only two intermediaries developed their own models for the purpose of granting loans, relying on supervised deep learning techniques (artificial neural networks). The remaining models use hybrid or unsupervised approaches, leveraging clustering algorithms (see Figure 10).

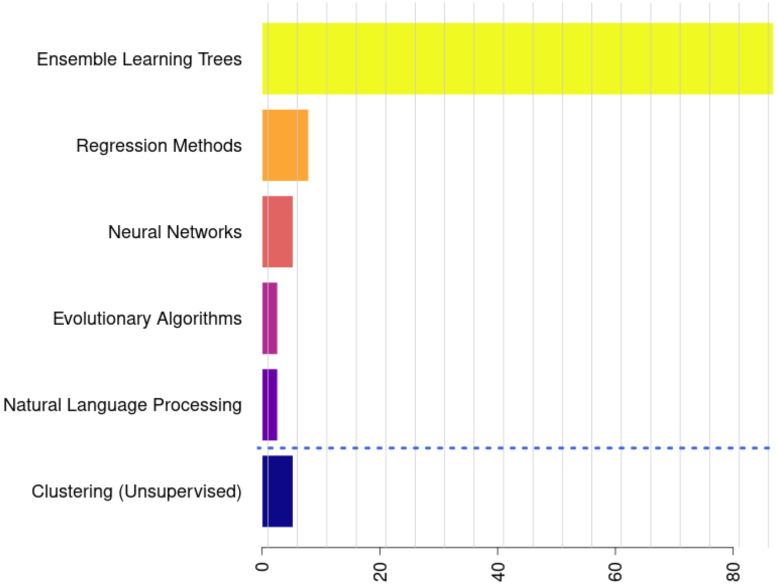


Figure 10. Type of techniques used (percentage shares) — the techniques using a supervised approach are reported above the dashed line; the only technique using an unsupervised approach is reported below the bar.

6.3 *Bias mitigation*

As mentioned above, algorithmic bias can arise at different points throughout the development of an ML model. Intermediaries were asked to indicate the techniques they employed to prevent and mitigate algorithmic bias along the three phases of data collection, model specification and forecast analysis downstream in the estimation process (see Section 4.2 and Appendix 1).

Overall, intermediaries reported that they monitor the data collection and preparation phase for more than 9 out of 10 models. Out of these, the treatment of missing observations (usually carried out in traditional statistical applications as well) has been applied to over 95 per cent of credit granting models (91 per cent when considering those already deployed), and to all the models deployed for monitoring. Techniques to mitigate historical bias have not been adopted for any of the models employed in granting or monitoring credit.

With respect to the specification, training and calibration phase of the models, the most popular techniques to mitigate algorithmic bias reported by the respondents include variables selection processes and the definition of the training process (i.e., choice of the target function), for 55 per cent and 45 per cent of the models, respectively. For the granting phase only, 59 per cent of the respondents reported that, ceteris paribus, the most interpretable model is preferred.

Finally, bias mitigation techniques applied downstream in the estimation process, which are important especially when the ML model contributes to automatic decision-making, are employed in 45 per cent of all models, and 39 per cent of those already in production. The results are shown in Figure 11.

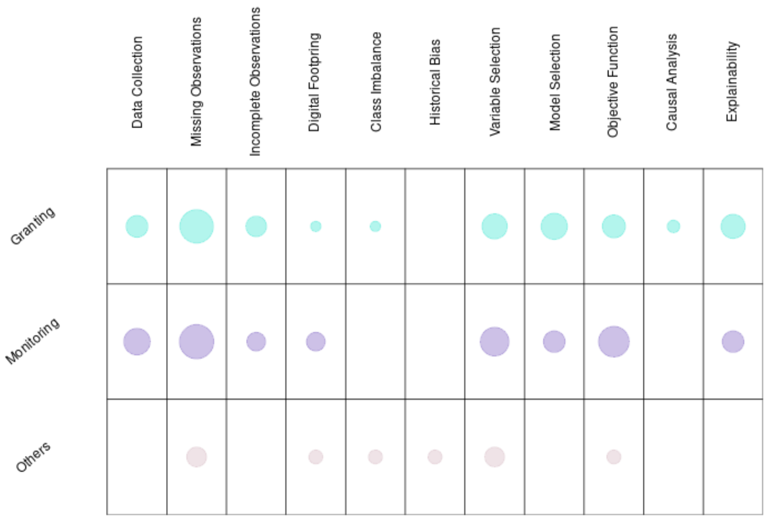


Figure 11. Overview of the safeguards adopted by intermediaries to mitigate the bias. The diameter of the circles reflects the share in the total of the models used for each phase.

During the interviews with the respondents, most intermediaries reported that they have adopted statistical methods for the treatment of the dataset (e.g. imputation of missing values), model specification (e.g. model choice) and performance analysis (e.g. explanatory techniques) for needs other than mitigating the algorithmic bias, e.g. to improve forecast accuracy. This presumably explains the decision not to use typical bias-reducing techniques, such as dataset balancing, controlling for historical bias, and causal analysis of the results. These techniques should be considered as best practices for mitigating significant bias in models for evaluating creditworthiness.

6.4 *Fairness*

Algorithmic bias may produce indirect or involuntary discriminatory mechanisms (see Section 4.1). When evaluating creditworthiness, failure to control for the possible presence of discriminatory dynamics is a particularly significant problem. Overall, respondents reported the adoption of a formal definition of fairness, or the intention to do so in 46 per cent of the models. As illustrated in Figure 12, which breaks down the models by type of customer, the percentage is 61 for retail customer-oriented models and 37 of for corporate/SME customer-oriented models.

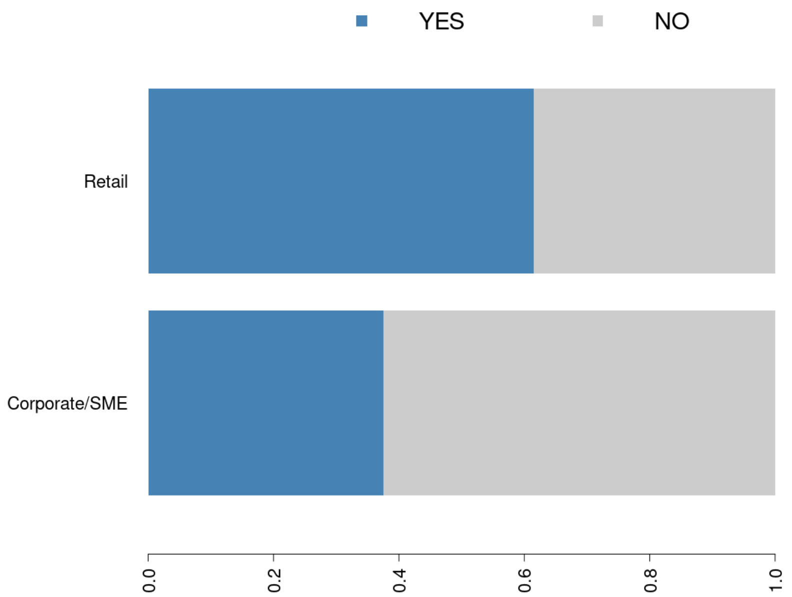


Figure 12. Share of models for which a definition of fairness was reported, broken down by retail and corporate/SME customers.

With respect to the different possible definitions of fairness – individual and group-level – financial intermediaries were asked to report their approach. All respondents reported that they adopted a definition of fairness through unawareness, which provides for the ex ante exclusion of sensitive attributes that enable identification of an individual as belonging to a vulnerable socio-economic group, such as gender for retail customers. In some cases, financial intermediaries decided to include potentially sensitive attributes, although they are not defined as such in the privacy laws (e.g. age, for retail customers), owing to their relevance for traditional models. According to intermediaries, the contribution of these attributes to the assessment of creditworthiness should be considered as a justified "differential treatment" rather than "differential impact" stemming from implicit prejudice.

How intermediaries measure fairness was investigated extensively. Overall, it was found that the above-mentioned focus on the issue of prejudicial discrimination does not translate into control practices established within a corporate policy on the topic. Only one intermediary has an established policy that takes into account the guidelines provided by an internal compliance committee. Another intermediary reported that it performs regular checks on the fairness of the model (specifically via counterfactual analysis). Nevertheless, such activity is informal rather than established within a codified process. Finally, two intermediaries developed models oriented to specific socio-demographic groups, selecting their training sample based on sensitive attributes, upstream of the estimation process. For these models, although this is not made technically explicit, notions of fairness are used to support the extension of credit to categories of people typically excluded due to their poor representation in credit data.

In general, the lack of established processes to address the issue of fairness and the possible presence of discriminatory mechanisms emerged during most interviews. Respondents appear to perceive that the use of (highly granular) non-traditional data and of models that are inherently more opaque than traditional ones is not accompanied by actual risks relating to the presence of discriminatory mechanisms toward customers. The most common arguments supporting this approach are the following: models that do not lead to automatic decision-making or relate to processes other than the granting of loans; the absence of explicit discriminatory assumptions during the development of the model; and the absence of automatic re-training mechanisms. In some cases, what amount to customer profiling techniques were used. In these cases, information on spending patterns (e.g. expenditure on culture, charity and medical expenses) is used to grant retail credit; however, the model assigns limited weight to such features in its decision-making process. Intermediaries considered this information as "financial data" rather than behavioural profiling. In one case, the model uses non-financial attributes (e.g. how, when and for how long the intermediary's web pages are browsed) to define the maximum interest rate that could be charged to customers given their creditworthiness.

6.5 *Model explainability*

The survey asked intermediaries to what extent they used explainability techniques, and who, if any, were the main recipients of the related outputs (e.g. developers, internal validators, external customers, etc.).

No intermediary reported employing explainable modelling techniques while developing models. As a result, intermediaries leverage either pre-modelling explainability or post-hoc explainability techniques (see Section 2.3).

Pre-modeling explainability techniques aim to understand the behaviour of the attributes of the dataset used to train the model. The respondent intermediaries use traditional statistical techniques for this purpose, such as exploratory data analysis or the creation and use of interpretable variables.

By contrast, post-hoc explainability techniques aim to explain ex post the logic of a model, and are particularly useful in the case of complex ML models. These techniques are applied to almost all models (92 per cent). The most frequently used *post-hoc* explainability techniques are Shapley Values and Feature Importance (reported for 71 and 39 per cent of the models respectively), followed by Partial Dependence Plot[49] (19 per cent) and LIME (6 per cent). Such techniques are available off-the-shelf in the main programming languages used to develop ML models (Python, R, etc.). Such feature is likely to have contributed to their adoption by intermediaries. Figure 13 shows the use of model explanation techniques.

[49] The use of the Partial Dependence Plot technique makes it possible to assess the impact of changes in a single attribute, within a given range, on the value of the variable returned as output by the model. The rationale underlying the use of this tool is to investigate and display possible non-linearities in the relationships between the model's input variables and its output.

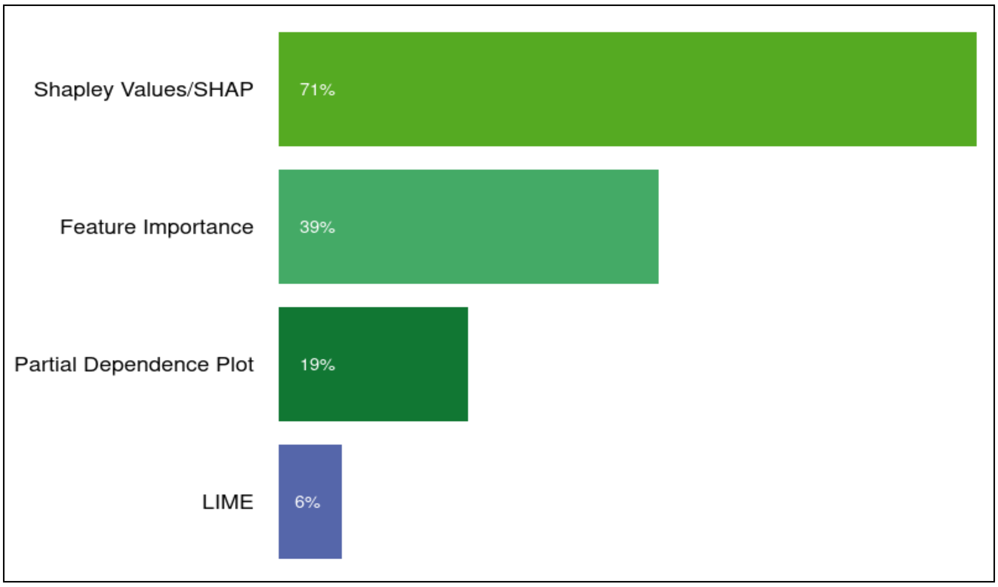


Figure 13. Frequency of use of post-hoc explanation techniques percentage share of all models that have adopted post-hoc explanation techniques).

Figure 14 shows the target recipients for the application of explainability techniques. Developers and users are the main targets (reported for 84 and 79 per cent of the models, respectively), followed by internal validators (68 per cent) and senior management (61 per cent). Although the GDPR provides that, upon request, customers shall receive information on the rationale used by the model employed to determine the evaluation concerning them, no intermediary has so far indicated external customers as recipients of the explanation techniques. In this respect, intermediaries have argued that the model is in many cases a supporting tool, whose outcome needs validation by an analyst, who is ultimately responsible for the final decision, which is also based on additional elements.

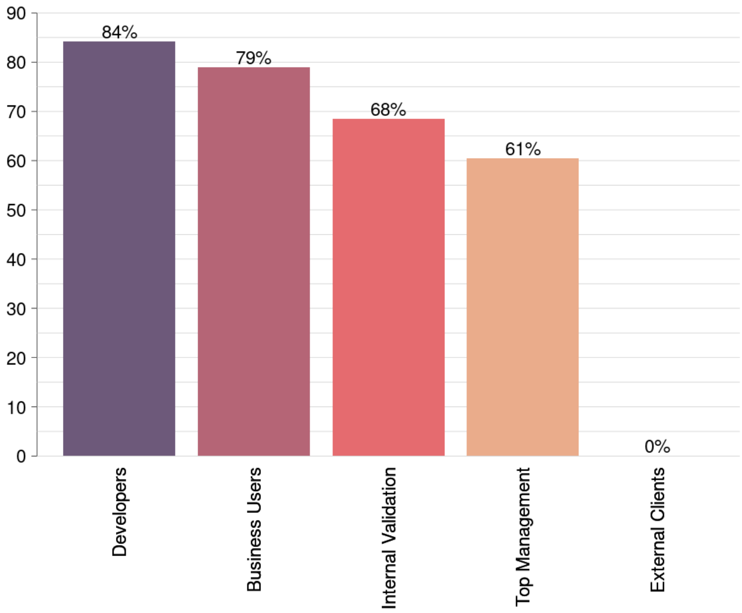


Figure 14. Recipients of post-hoc explanation techniques.

6.6 *Governance*

The governance of model-based processes is of key importance; it is even more important in the case of AI systems owing to the opacity of the internal structure and the decision-making rationale followed. It is therefore necessary for intermediaries to put in place appropriate safeguards to check that AI systems provide actual value added to the processes in which they are used and do not generate incremental risks that are difficult to manage. The necessary controls include verifying the performance of models and the stability of output over time, ensuring the quality and integrity of the data[50] used, and identifying and overseeing the risks associated with the use of these techniques.

With reference to the results of the survey, all intermediaries have set up or intend to establish a monitoring and reporting process for the AI systems in place. The reports focus in almost all cases on the accuracy of the output (97 per cent); for a large share of models, they also includes assessment of the stability of the estimates over time (74 per cent). Less common are metrics related to the quality and integrity of the data used by the models, which are reported for just over the half of the cases.[51]

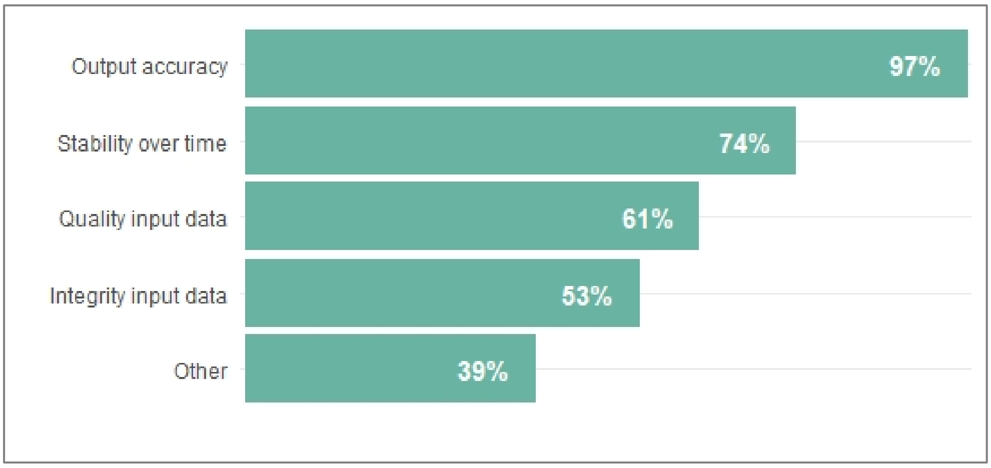


Figure 15. Metrics included in the reporting (data by model).

Regarding the strategy chosen to update AI systems, semi-automatic training systems (i.e. partially supervised by human intervention) have been used in 66 per cent of cases, and manual re-training in 29 per cent of cases. Only one intermediary appears to use, for two models, a fully automated updating procedure. According to the respondents, training the model on a monthly basis improves performance in terms of accuracy; the possible discontinuity between successive versions of the models is not considered an issue, as these models are used for monitoring credit exposures.

[50] Failure to monitor the quality and integrity of the input data could lead to data drift problems (changes in the relations between data assumed or estimated as a result of changes over time in the data provided during the training phase) and therefore require appropriate re-training.

[51] Other metrics have also been declared as being part of the reports (under "Other" in Figure 15), including checks on the consistency between the distributions of current data and those used for training, checks on the distribution of statistical scores for monitoring, and back-testing.

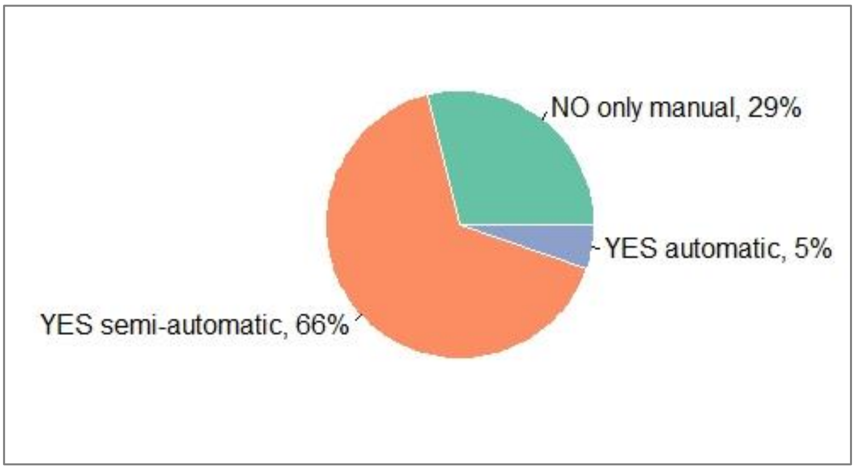


Figure 16. Model re-training system

With regard to the benefits and risks associated with the use of AI techniques, the feedback from intermediaries is generally optimistic: using such techniques appears to have many benefits and few risks compared with traditional techniques. The main benefit is the increase in forecast accuracy, which is considered important by all intermediaries regardless of the purpose of the AI systems. Further benefits considered to be generally important include the possibility of using alternative data sources, greater efficiency in processes, and the fact of providing access to credit to a wider range of customers. The perceived benefit in terms of stability of the model's performance over time is generally considered to be less important.

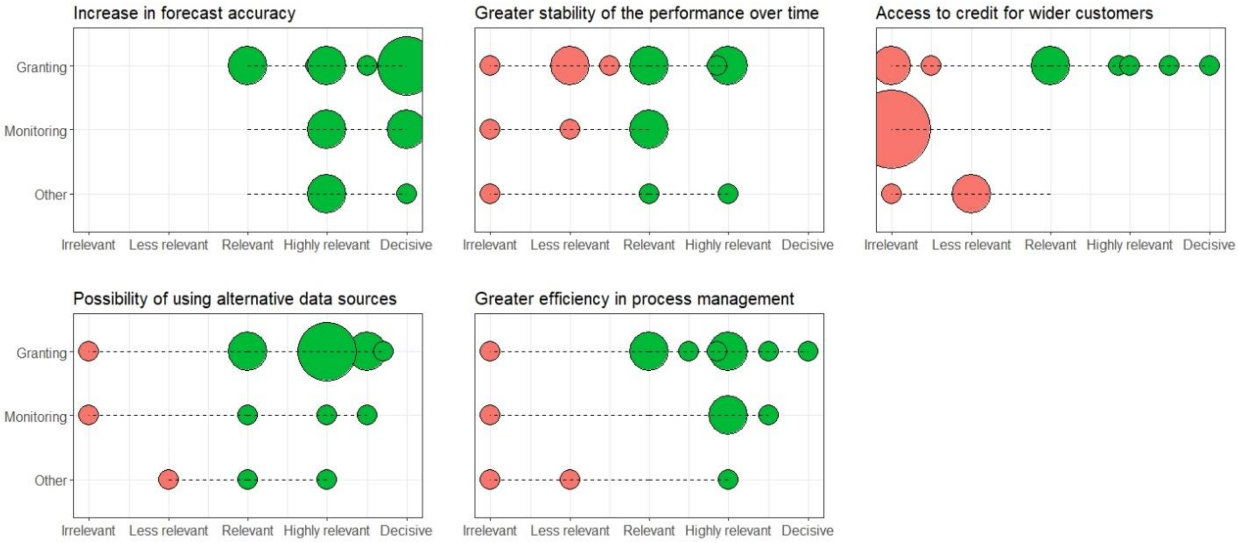


Figure 17. Perceived benefits associated with the use of AI systems (data per intermediary, average of active models for each phase). The diameter of the circles reflects the number of intermediaries reporting the same level of importance.

With regard to the perceived incremental risks with respect to traditional techniques, evidence from the survey shows that the focus of intermediaries is mainly on operational risks: for 5 out of the 9 models used for granting credit, operational risks are assessed as material or very material (2 out of

4 for monitoring models). Other risks, including reputational, legal and IT security risks, are considered in most cases to be insignificant or of limited importance.

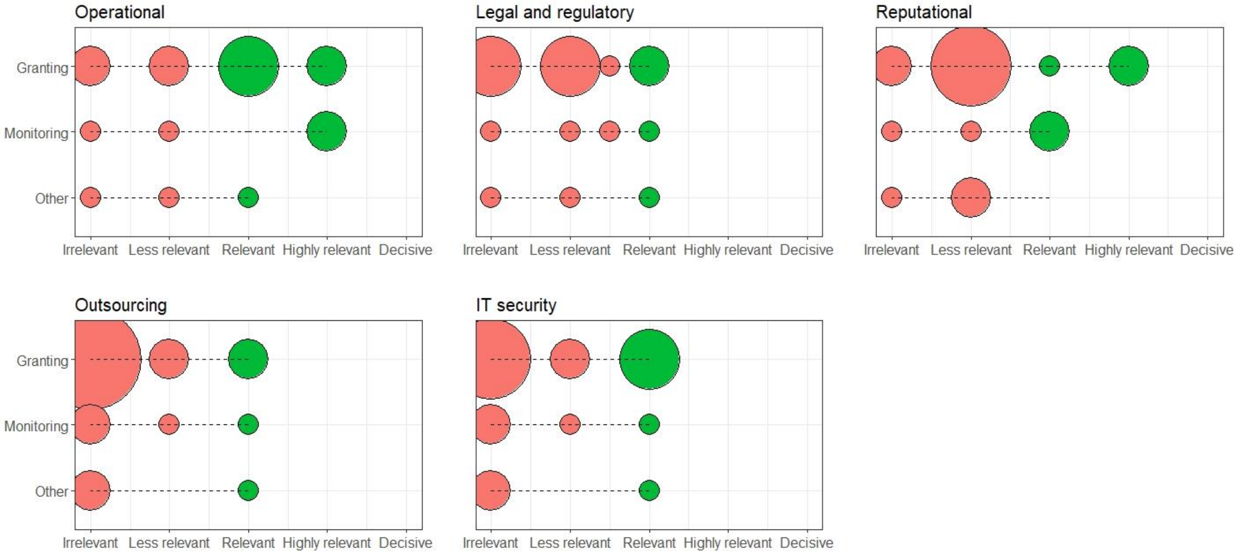


Figure 18. Perceived risks associated with the use of AI systems (data per intermediary, average of active models for each phase). The diameter of the circles reflects the number of intermediaries reporting the same level of importance to the subject.

With regard to the use of outsourcing in model development and management, it is necessary to adopt safeguards that allow intermediaries to maintain full governance of the modelling variable, with particular reference to certain phases of the process (e.g. monitoring of performance and absence of bias in selection, re-training, and transparency vis-à-vis customers).

On this topic, 47 per cent of the models also saw the use of external staff in their development; only 8 per cent of the models were fully developed by external staff. In the latter area, two forms of complete outsourcing were found:

- full outsourcing to an external company that develops the model autonomously and provides the final score for each customer directly to the intermediary;
- “double outsourcing”, where the supplier provides banks and financial companies with the result of a scoring model developed in cooperation with a third party.

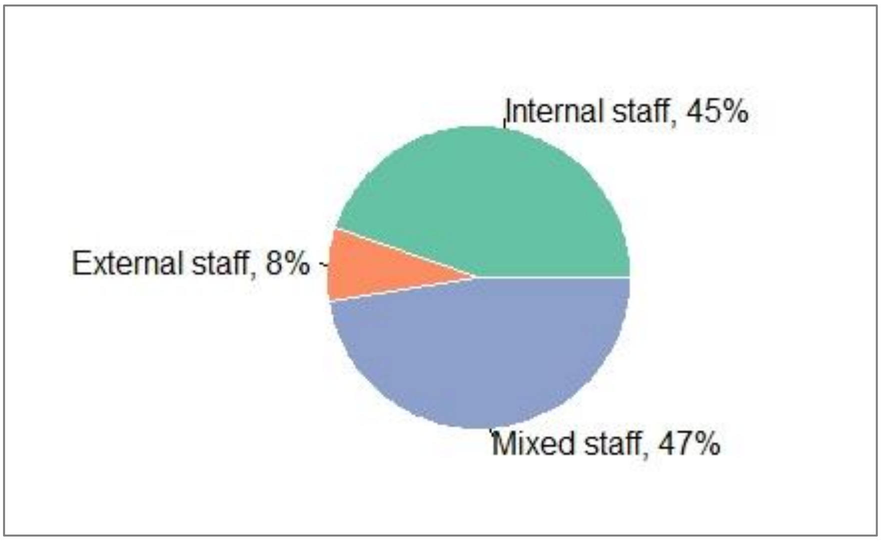


Figure 19. Composition of the working groups (data by model)

Business users of AI models were sufficiently involved in the design and development phase (82 per cent with medium or high involvement), and their understanding of the assumptions underlying the models appears to be high (with 76 per cent reporting a medium-high level). Conversely, the level of general knowledge of AI techniques by the business users is significantly lower, in the range of medium-low.

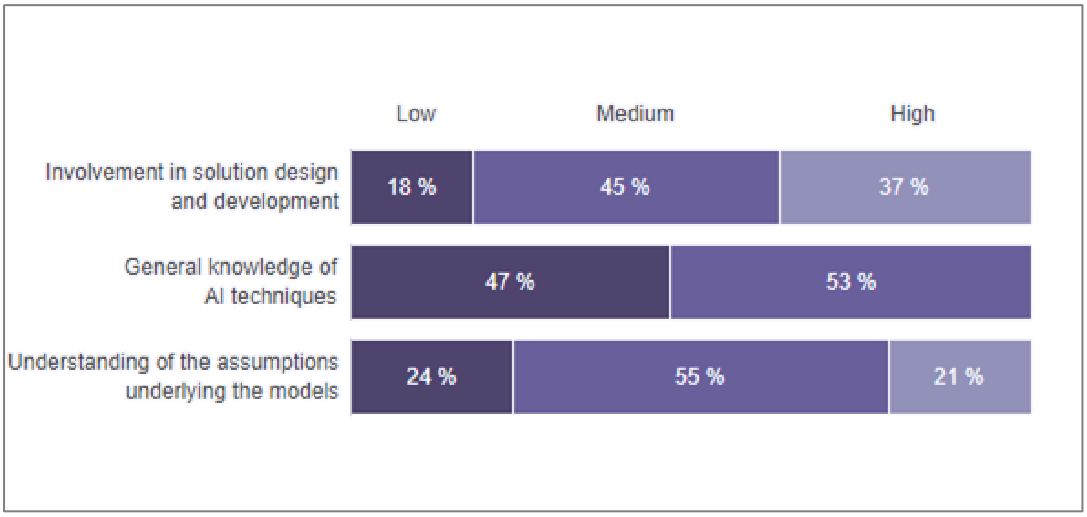


Figure 20. Involvement of business users in relation to AI (data by model)

Finally, the intermediaries were asked to report what kind of information they provide to customers when they receive a request about the rationale behind the decision made using an ML model, in compliance with the requirements of Article 15 of the GDPR. The information provided to customers is generally limited, also given the assumption that the model is normally used to support a credit decision that will be made by an analyst.

Appendix 1 — Bias detection techniques in case of use of AI-ML models

Below are reported some techniques that allow the detection of bias that may arise throughout the main phases of the development of an AI-ML algorithm: data collection, model specification, and output analysis.

Data collection

The types of bias that originate within the data collection and preparation phase are manifold, and are typically detected via in-depth analysis of the data. They include:

(I) sample selection mechanisms, e.g. unbalanced classes for traditional datasets and *self-selection bias* for big data;

(II) non-random *missingness* mechanisms and related treatment of missing data;

(III) the presence of incomplete observations with respect to the target variable, including *thin-file and no-file subjects* as well as invalid *digital footprints*;

IV) historical bias in data and institutional bias, exogenous to the modelling process.

Bias is typically detected at this stage through elicitation processes[52] (Manzi et al., 2019) and collective development approaches (e.g., *crowdsourcing*).

Bias can be mitigated, depending on the applications and surrounding regulatory framework, via targeted selection of subsets of observations, or the allocation of weights to ensure adequate representativeness of all vulnerable subpopulations in the training set.

Model specification and learning

While defining the model, any type of bias translates into deviations from the ontological formalisms for reasoning. These are defined by the analyst to explain the set of assumptions underlying the model formulation, and the implications deriving from the relationships between them, following an approach similar to the aforementioned bias elicitation.

Critical steps for bias detection are the monitoring of omitted variable bias and/or included variable bias, the choice of the model (e.g. by favoring more interpretable models over black boxes) and of

[52] Elicitation of the distortion is a process that requires an expert group to technically assess a given model in its different assumptions and components, in order to identify the different critical points. The use of such an exercise is typically associated with the use of traditional econometric models and has subsequently been extended to the ML. Operationally, it consists of assessing the benefits, and the risks, of including an attribute, captured on a given scale, or of individual observation units.

the objective function used for training (for example, including of regularization terms or penalties targeting algorithm bias)[53].

Moreover, once the optimal trade-off between the external and internal validity of a model is assessed, it is key to monitor the stability of the model over time or across different application contexts. These controls typically aim to prevent and mitigate performance decreases due to drifting dynamics, caused by changes in the phenomenon modelled or in part of its characteristics. Among others, schemes of periodic updating (re-training) or continuous updating (online learning) of the model, as well as dynamic selection of features used (feature dropping), intended to prevent and mitigate drift mechanisms, shall be implemented.

Output analysis

Bias can finally occur downstream of the estimation and calibration process of the model, due to the lack of an adequate understanding of the mechanisms that led to the results. To this end, evaluation criteria are critical for the analysis of:

(I) the causality mechanisms and dependency relationships among features, in order to contain cause-effect bias or aggregation bias;

(II) the justification of decisions determined by sensitive attributes;

(III) the explainability of the model, using Explainable AI techniques.

[53] The inclusion of constraints that pursue *fairness* may in some cases lead to a reduction in accuracy; e.g. (Hardt et al., 2016; Hickey et al., 2020).

Appendix 2 — Economic literature on discrimination in the credit market and the contribution of quantitative methods and ML models

The economic literature on discrimination in the credit market has focused mainly on ascertaining whether or not there is discrimination of population subgroups, particularly in the North American and UK markets, and gender discrimination.

There are a number of papers suggesting that firms owned by minority groups are discriminated in the form of rationing of the amount of credit (e.g. Cavalluzzo and Cavalluzzo, 1998; Cavalluzzo et al. 2002 and Blanchflower et al. 2003; Fraser 2009a).

Some studies find that there is discrimination in the household loan market. In particular, the rejection rate for a loan application is higher for some ethnic groups than for the population of Caucasian origin, holding constant characteristics of the borrower and the loan (Munnell et al. (1996); Ross and Yinger, 2002; Ross and Tootell (2004)). This result is confirmed when taking into account the characteristics of the area of residence of applicants, including the local prevalence of specific ethnic groups (Tootell, 1996).

Other analyses provide evidence on interest rates charged: Edelberg (2007) shows that there is a significant degree of unexplained heterogeneity in the rates charged on consumer credit and mortgages, especially before 1995, which is potentially attributable to discrimination. Direct evidence of a higher cost of credit for some ethnic groups is also found in Bayer et al. (2018)", Ghent et al. (2014) and Cheng et al. (2015). However, in the United States, the cost of credit consists of several components, so it may not be sufficient to compare interest rates to demonstrate discrimination. A recent paper shown that some ethnic groups pay higher interest rates but tend to select mortgages with lower initial fixed costs (Bhutta and Hizmo, 2021). The study does not elaborate on the reasons for the observed differences, but suggests that they may reflect different preferences or differences in cash holdings. Moreover, inefficiencies in the mortgage market, which prevent customers from searching for the most advantageous contract, could have different effects across household categories owing to their different financial expertise (Woodward, 2008; Woodward and Hall, 2012).

Discrimination between groups could arise not only in prices and rejection rates but also in the quality of the services provided. Hanson et al. (2018) show, through an experiment conducted in the United States, that the response rate for potential borrowers asking for information about the contracts offered is lower for some ethnic groups c and that these groups are provided less information by bank staff.

Many studies document the presence of gender discrimination. Discrimination would be *taste-based*, i.e. based on prejudice, rather than statistical discrimination, i.e. resulting from the correlation between gender and characteristics relevant for the identification of creditworthiness but not observable (Hisrich and Brush, 1984; Buttner and Rosen 1988, 1989). Analyses show that female entrepreneurs have greater difficulties in accessing the credit market than male entrepreneurs (Fay and Williams, 1993; Cavalluzzo et al., 2002; Fraser, 2009b), especially in the start-up phase (Orser et al., 2000); they also need to provide more collateral and pay higher interest rates (Coleman, 2000). Cross-country evidence suggests, on the other hand, that gender differences in access to credit tend to be more pronounced in environments where the financial system is less developed (Muravyev et al., 2009).

Evidence supporting the presence of discrimination, by ethnic origin and by gender, is also available for Italy, with regard to corporate credit. In general, the Italian studies make use of very detailed databases to include in the analysis a wide range of characteristics of the company, contract type and intermediary; this makes it more possible to verify the presence of systematic differences in access to credit across groups of borrowers, all other things being equal. Albareto and Mistrulli (2011) show that, between 2004 and 2008, the interest rates applied by Italian banks on loans to micro-firms owned by immigrants are about 70 basis points higher than those paid by Italian entrepreneurs, controlling for firm characteristics, but that the differential narrows with the lengthening of the business's credit history. According to the authors, on the one hand the knowledge acquired by banks in the course of the credit relationship makes it possible to overcome the largest ex ante information asymmetries; on the other hand, foreign entrepreneurs improve their knowledge of the Italian banking market and manage to obtain better conditions over time.

Bellucci et al. (2010), analysing the credit lines granted by a large Italian bank to sole proprietorships between 2004 and 2006, highlight that female borrowers have greater difficulties in accessing credit and, even when they do not pay higher interest rates, are significantly more likely to have to pledge collateral. Similar evidence results from the study by Calcagnini et al. (2015), based on data on credit lines granted by three Italian banks over the period 2005-2008; in the sample analysed female-owned firms are significantly smaller and younger than male-owned firms and tend to rely on smaller loans that are typically less collateralized. Nevertheless, the reduced access to credit is only partly explained by these characteristics since, even considering these factors, female-owned firms need to provide more frequently collateral to access credit.

Evidence of gender discrimination is also present in the work of Alesina, Lotti and Mistrulli (2013), which uses data from the Credit Register on loans granted to micro firms in the period 2004-07. The analysis, which takes into account a very broad set of firm and local credit market characteristics, shows that firms owned by women pay higher interest rates, even when they are not riskier than those owned by men. The spread is small but is not explained by differences in credit history. Another study also employing bank-firm-level data from the Central Credit Register, by Cesaroni et al. (2013), shows that, controlling for all available observable characteristics, female-owned firms faced a more pronounced tightening in credit supply conditions over the period 2007-09 than other firms.

The above-mentioned analyses compare the conditions applied across customer categories but do not analyze whether the adoption of credit scoring methods contributes or not to discrimination. Some analyses have tried to shed light on the impact of credit scoring in the United States, where the diffusion of statistical methods of credit scoring has spurred a debate on its impact on access to credit (see Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit, 2007). The results are mixed and depend on the data used and on the methods of investigation. Avery et al. (2009 and 2012) conclude that quantitative methods appear to reduce discrimination, while other analyses suggest that some characteristics, such as age, may lead to differences in customer classification that are not justified by genuine differences in creditworthiness.

More recently, some studies analyze the effect on discrimination of the adoption of credit assessment techniques based on AI algorithms in comparison with traditional credit scoring models. Bartlett, Morse, Stanton and Wallace (2021) estimate the different effects on discrimination resulting from the use of traditional and innovative credit assessment techniques in the US mortgage market. Their analysis shows that Latin American and Afro-American borrowers, with equivalent characteristics to other clients, pay rates which are around 8 basis points higher (4 for refinancing of outstanding mortgages); the rate disparities are about one-third lower for loans granted by FinTech firms and these intermediaries do not exhibit differences in rejection rates across ethnic groups. With regard to US real estate mortgages, Fuster, Goldsmith-Pinkham, Ramadorai and Walther (2020) compare in a simulation exercise the ability to predict default of traditional models and ML models, as well as the interest rates that would be applied on the basis of credit risk. Their evidence suggests that ML models are more accurate in predicting the default than less sophisticated technologies and tend to increase market access, reducing the dispersion between acceptance rates across different subgroups of the population. The ML models also result in greater differentiation in the cost of credit between customers; considering all the customers who would borrow on the basis of both traditional and innovative technologies, the ML model generates higher rates for most risk-averse clients within the groups of Afro-Americans and Hyspanics.

Another simulation, by Bono, Croxon and Giles (2021), uses a large set of very detailed data for the United Kingdom to test the effects of a hypothetical switch from a traditional credit scoring model to an ML model. The analysis confirms the higher accuracy of the ML model and suggests that the ML model does not exacerbate or eliminate potential distortions vis-à-vis groups with sensitive attributes, such as ethnicity and customer gender.

Overall, the evidence available so far does not suggest that ML has the potential to amplify discrimination, if present, compared with traditional statistical methods.

References

For text and appendix 1

Albanesi, S., & Vamossy, D. F. (2019). Predicting consumer default: A deep learning approach (No. w26165). National Bureau of Economic Research.

Alonso, A., & Carbó, J.M. (2020). Machine Learning in Credit Risk: Measuring the Dilemma between Prediction and Supervisory Cost, Banco de España WP 2032.

Bacham, D., & Zhao, J. (2017). Machine learning: challenges, lessons, and opportunities in credit risk modeling. Moody's Analytics Risk Perspectives, 9, 30-35.

Baeza-Yates, R. (2018). Bias on the web. Communications of the ACM, 61(6), 54-61.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. Expert Systems with Applications, 83, 405-417.

Bazarbash, M. (2019). Fintech in financial inclusion: machine learning applications in assessing credit risk. International Monetary Fund.

Bellomarini, L., Gottlob, G., & Sallinger, E. (2018). The vadalog system: Datalog-based reasoning for knowledge graphs. arXiv preprint arXiv:1807.08709.

Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of fintechs: Credit scoring using digital footprints. The Review of Financial Studies, 33(7), 2845-2897.

Beydoun, G., Suryanto, H., Guan, C., Guan, A., & Sugumaran, V. Unlocking Knowledge Graphs (KG) Potentials in Support of Credit Risk Assessment.

Bryant, K. (2001). ALEES: an agricultural loan evaluation expert system. Expert systems with applications, 21(2), 75-85.

Calo, R. (2013). Digital market manipulation. Geo. Wash. L. Rev., 82, 995.

Cascarino, G., Moscatelli, M., & Parlapiano, F. (2022). Explainable Artificial Intelligence: interpreting default forecasting models based on Machine Learning. Bank of Italy Occasional Paper, (674).

d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. Big data, 5(2), 120-134.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference (pp. 214-226).

Fantazzini, D., & Figini, S. (2009). Random survival forests models for SME credit risk measurement. Methodology and computing in applied probability, 11(1), 29-45.

Freedman, S., & Jin, G. Z. (2017). The information value of online social networks: lessons from peer-to-peer lending. International Journal of Industrial Organization, 51, 185-222.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2020). Predictably unequal? the effects of machine learning on credit markets. (October 1, 2020).

Gottlob, G., & Pieris, A. (2015, June). Beyond SPARQL under OWL 2 QL entailment regime: Rules to the rescue. In Twenty-Fourth International Joint Conference on Artificial Intelligence.

Hacker, P. & Passoth J. H. (2021). Varieties of AI Explanations Under The Law. From the GDPR to the AIA, and Beyond (August 25, 2021).

Hajian, S., & Domingo-Ferrer, J. (2012). A methodology for direct and indirect discrimination prevention in data mining. IEEE transactions on knowledge and data engineering, 25(7), 1445-1459.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in neural information processing systems, 29, 3315-3323.

Hickey, J. M., Di Stefano, P. G., & Vasileiou, V. (2020). Fairness by Explicability and Adversarial SHAP Learning. arXiv preprint arXiv:2003.05330.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., & Zimmermann, A. (2021). Knowledge graphs. Synthesis Lectures on Data, Semantics, and Knowledge, 12(2), 1-257.

Iwasieczko, B., Korczak, J., Kwiecień, M., & Muszyńska, J. (1986). Expert system in financial analysis. IFAC Proceedings Volumes, 19(17), 113-120.

Jagtiani, J., & Lemieux, C. (2017). Fintech lending: Financial inclusion, risk pricing, and alternative information.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), 2767-2787.

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.

Manzi, G., & Forster, M. (2019). Biases in bias elicitation. Communications in Statistics-Theory and Methods, 48(18), 4656-4674.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.

Moscatelli, M., Parlapiano, F., Narizzano, S., & Viggiano, G. (2020). Corporate default forecasting with machine learning. Expert Systems with Applications, 161, 113567.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... & Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1356.

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. Applied Soft Computing, 74, 26-39.

Piketty, L. (1987, April). The authorizer's assistant: a large commercial expert system application. In Proceedings of the AI and advanced computer technology conference. Long Beach, CA.

Roa, L., Rodríguez-Rey, A., Correa-Bahnsen, A., & Valencia, C. (2021). Supporting Financial Inclusion with Graph Machine Learning and Super-App Alternative Data. arXiv preprint arXiv:2102.09974.

Tobback, E., & Martens, D. (2019). Retail credit scoring using fine-grained payment data. Journal of the Royal Statistical Society: Series A (Statistics in Society), 182(4), 1227-1246.

Yuan, D. (2015). Applications of machine learning: consumer credit risk analysis (Doctoral dissertation, Massachusetts Institute of Technology).

Zocco, D. P. (1985). A framework for expert systems in bank loan management. Journal of Commercial Bank Lending, 67(2), 47-55.

For appendix 2

Albareto, G., & Mistrulli, P. E. (2011). Bridging the gap between migrants and the banking system. Bank of Italy Temi di Discussione (Working Paper) No, 794.

Alesina, A. F., Lotti, F., & Mistrulli, P. E. (2013). Do women pay more for credit? Evidence from Italy. Journal of the European Economic Association, 11(suppl_1), 45-66.

Avery, R. B., Brevoort, K. P., & Canner, G. B. (2009). Credit scoring and its effects on the availability and affordability of credit. Journal of Consumer Affairs, 43(3), 516-537.

Avery, R. B., Brevoort, K. P., & Canner, G. (2012). Does Credit Scoring Produce a Disparate Impact? Real Estate Economics, 40, S65-S114.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2021). Consumer-lending discrimination in the FinTech era. Journal of Financial Economics.

Bayer, P., Ferreira, F., & Ross, S. L. (2018). What drives racial and ethnic differences in high-cost mortgages? The role of high-risk lenders. The Review of Financial Studies, 31(1), 175-205.

Bellucci, A., Borisov, A., & Zazzaro, A. (2010). Does gender matter in bank–firm relationships? Evidence from small business lending. Journal of Banking & Finance, 34(12), 2968-2984.

Bhutta, N., & Hizmo, A. (2021). Do minorities pay more for mortgages? The Review of Financial Studies, 34(2), 763-789.

Blanchflower, D. G., Levine, P. B., & Zimmerman, D. J. (2003). Discrimination in the small-business credit market. Review of Economics and Statistics, 85(4), 930-943.

Bono, T., Croxson, K., & Giles, A. (2021). Algorithmic fairness in credit scoring. Oxford Review of Economic Policy, 37(3), 585-617.

Buttner, E. H., & Rosen, B. (1988). Bank loan officers' perceptions of the characteristics of men, women, and successful entrepreneurs. Journal of Business venturing, 3(3), 249-258.

Buttner, E. H., & Rosen, B. (1989). Funding new business ventures: Are decision makers biased against women entrepreneurs? Journal of Business Venturing, 4(4), 249-261.

Calcagnini, G., Giombini, G., & Lenti, E. (2015). Gender differences in bank loan access: An empirical analysis. Italian Economic Journal, 1(2), 193-217.

Cavalluzzo, K. S., & Cavalluzzo, L. C. (1998). Market structure and discrimination: The case of small businesses. Journal of Money, Credit and Banking, 771-792.

Cavalluzzo, K. S., Cavalluzzo, L. C., & Wolken, J. D. (2002). Competition, small business financing, and discrimination: Evidence from a new survey. The Journal of Business, 75(4), 641-679.

Cesaroni, F. M., Lotti, F., & Mistrulli, P. E. (2013). Female Firms and Banks' Lending Behaviour: What Happened during the Great Recession? Bank of Italy Occasional Paper, (177).

Cheng, P., Lin, Z., & Liu, Y. (2015). Racial discrepancy in mortgage interest rates. The Journal of Real Estate Finance and Economics, 51(1), 101-120.

Coleman, S. (2000). Access to capital and terms of credit: A comparison of men-and women-owned small businesses. Journal of small business management, 38(3), 37.

Edelberg, W. (2007). Racial dispersion in consumer credit interest rates.

Fay, M., & Williams, L. (1993). Gender bias and the availability of business loans. Journal of Business Venturing, 8(4), 363-376.

Fraser, S. (2009a). Is there ethnic discrimination in the UK market for small business credit?. International Small Business Journal, 27(5), 583-607.

Fraser, S. (2009b). Small firms in the credit crisis: Evidence from the UK survey of SME finances. Warwick Business School, University of Warwick.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2020). Predictably unequal? the effects of machine learning on credit markets. The Effects of Machine Learning on Credit Markets (October 1, 2020).

Ghent, A. C., Hernandez-Murillo, R., & Owyang, M. T. (2014). Differences in subprime loan pricing across races and neighborhoods. Regional Science and Urban Economics, 48, 199-215.

Hanson, A., Hawley, Z., Martin, H., & Liu, B. (2016). Discrimination in mortgage lending: Evidence from a correspondence experiment. Journal of Urban Economics, 92, 48-65.

Hisrich, R., & Brush, C. (1984). The woman entrepreneur: Management skills and business problems. University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship.

Munnell, A. H., Tootell, G. M., Browne, L. E., & McEneaney, J. (1996). Mortgage lending in Boston: Interpreting HMDA data. The American Economic Review, 25-53.

Muravyev, A., Talavera, O., & Schäfer, D. (2009). Entrepreneurs' gender and financial constraints: Evidence from international data. Journal of comparative economics, 37(2), 270-286.

Orser, B. J., Hogarth-Scott, S., & Riding, A. L. (2000). Performance, firm size, and management problem solving. Journal of small business management, 38(4), 42.

Ross, S. L., & Tootell, G. M. (2004). Redlining, the Community Reinvestment Act, and private mortgage insurance. Journal of Urban Economics, 55(2), 278-297.

Ross, S. L., & Yinger, J. (2002). The color of credit: Mortgage discrimination, research methodology, and fair-lending enforcement. MIT press.

Tootell, G. M. (1996). Redlining in Boston: Do mortgage lenders discriminate against neighborhoods? The Quarterly Journal of Economics, 111(4), 1049-1079.

Woodward, S. E., & Hall, R. E. (2012). Diagnosing consumer confusion and sub-optimal shopping effort: Theory and mortgage-market evidence. American Economic Review, 102(7), 3249-76.

Woodward, S. E. (2008). A study of closing costs for FHA mortgages. US Department of Housing and Urban Development, Office of Policy Development and Research.