Rossella Locatelli · Giovanni Pepe · Fabio Salis

Artificial Intelligence and Credit Risk

The Use of Alternative Data and Methods in Internal Credit Rating

palgrave
macmillan

Artificial Intelligence and Credit Risk

Rossella Locatelli · Giovanni Pepe · Fabio Salis

Artificial Intelligence and Credit Risk

The Use of Alternative Data and Methods in Internal Credit Rating

palgrave
macmillan

Rossella Locatelli
University of Insubria
Varese, Italy

Giovanni Pepe
KPMG Advisory
Milano, Italy

Fabio Salis
Credito Valtellinese
Sondrio, Italy

ISBN 978-3-031-10235-6 ISBN 978-3-031-10236-3 (eBook)
https://doi.org/10.1007/978-3-031-10236-3

Translation from the Italian language edition 'Artificial Intelligence e credit risk' by Rossella Locatelli et. al. © Aifirm 2022 Published by Aifirm. All Rights Reserved.

This Palgrave Macmillan imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

About This Book

The present work is the translation of the AIFIRM Position Paper n. 33 “Artificial Intelligence e Credit Risk. Possibili utilizzi di metodologie e dati alternativi nei sistemi interni di rating”.
Aifirm is the Italian Association of Financial Industry Risk Managers.

The contents of the manuscript were developed with a working group methodology. The composition of the working groups is presented in the last section of this volume. All the listed members have contributed to the writing of the single paragraphs as specified in the last section of the volume.

The working groups were coordinated by Corrado Meglio, Aifirm.

Executive Summary (English)

During the last decade the increase in computational capacity, the consolidation of new data processing methodologies and the availability of and access to new information concerning individuals and organisations, aided by widespread internet usage, has set the stage for the development and application of artificial intelligence (AI) in companies in general and financial institutions in particular.

The application of AI in the banking world is even more significant thanks to the use of larger and larger datasets for credit risk modelling. The Revised Payment Services Directive (PSD2), which, since 2019, enables companies and individuals to make their spending behaviour information available to third parties, has increased the importance of this trend.

Credit risk scoring has largely been based on customer data modelling for many years now. These modelling techniques (linear regression, logistic regression, decision trees, etc.) and the datasets exploited (financial, behavioural, social, geographic, sectoral information etc.) are referred to as "traditional" and have been the *de facto* standards in the banking industry.

The challenge on the horizon for credit risk managers is to find ways to leverage the new AI toolbox and new data to enhance the models' predictive power, without overlooking problems associated with the interpretability of results and focusing on the ethical issues that AI raises. AI techniques could become crucial when dealing with one of the most

important issues in credit risk modelling: credit decisions for individuals or companies that are not yet customers of the bank, i.e., whose behaviour history is not available to the bank.

Moreover, the disruption caused by the COVID-19 economic crisis has revealed how the use of alternative datasets, like transactional datasets, could improve the models' predictive power during regime changes, which traditional models fail to track on a timely basis.

Leading Italian banks have already advanced significantly in this direction, similarly to their peers in other developed economies, focusing on how to include alternative datasets in their credit risk models and on which alternative modelling techniques yield added value.

With the publication of this position paper, AIFIRM's objective is to examine the alternative techniques and data leveraged for credit risk management, describing and analysing the array of methodological approaches for the use of alternative techniques and/or data for supervisory and management credit rating models. This paper thus provides an overview of the steps already taken by banks and discusses the state of the art of AI and alternative data within banks. The premise is that, by using alternative datasets and AI techniques, the increase in the predictive power of rating models may enable greater risk discrimination, without losing control on the scoring techniques.

Traditional data and modelling techniques can be combined with alternative data and techniques and, by doing so, several solutions can be considered:

- the possibility of expanding the starting dataset, adding alternative data to the traditional data and processing both using alternative techniques only;
- the possibility of harnessing the informational power of the alternative data, using such data in models developed with both traditional and alternative techniques.

In the case of the former, using AI techniques alone makes it possible to better harness the informational power of the available data, increasing the risk discrimination capacity. In the second, the combination of traditional and alternative techniques makes it possible to balance issues relating to performance and interpretability, a circumstance that is relevant given the fact that the implementation of alternative techniques undoubtedly leads to higher accuracy in rating models but makes them more difficult to interpret than traditional models.

Alternatively, the use of alternative data on a stand-alone basis, processed with AI techniques may be considered, as this typically results in better performance than traditional data processed with traditional techniques. By the same token, a widespread use of AI techniques may pose ethical problems, in that a reduced human control on non-traditional models can end up in undesirable discriminations of credit applicants.

This paper is organised in five chapters.

The introduction describes the use of AI credit risk models in Italy.

Chapter 2 covers the methodologies used to develop AI components or models. In particular:

- Section 2.1 analyses the categories of alternative data that can be used to enlarge the dataset in rating models—i.e., transactional data, social data and data from utilities companies—and describes the necessary processing of such data using natural language processing (NPL) techniques to be implemented in AI models;
- Section 2.2 presents several machine learning (ML) algorithms that can be used to develop credit risk models—i.e., decision trees, random forests, gradient boosting and neural networks—analysing the strengths and weaknesses of each.

The third chapter provides a few examples of AI credit risk models implemented by certain banks for development and validation and one applied to asset management. In particular:

- Section 3.1 covers two business cases based on the inclusion of innovative data in Probability of Default (PD) models, in which new types of information are implemented in the models through one or more AI modules and are developed individually and then integrated with the other modules.
- Section 3.2 discusses the methods of integrating traditional modelling techniques with alternative methods and describes a two-step approach whereby the PD model is enhanced with AI.
- Section 3.3 is dedicated to a specific application of AI models for asset management to corroborate investment decisions.

- Section 3.4 explores AI techniques that can be used as challenges in the validation of credit risk models by presenting business cases in which AI is used to validate models for both supervisory and management reporting purposes.

Chapter 4 analyses techniques for the comparison and validation of AI models. In particular:

- Section 4.1 describes a series of key trustworthy indicators (KTIs), the aim of which is to render operational the principle of AI model trustworthiness. As established by the European Commission in the regulatory framework for AI proposed on 21 April 2021, trustworthiness is based on elements like accuracy, robustness, equity, efficiency and interpretability.
- Section 4.2 focuses on the two main aspects of AI: interpretability and stability.

Chapter 5, which concludes this paper, addresses the evolution of AI models. In particular:

- Section 5.1 outlines possible developments in the use AI models for credit risk scoring.
- Section 5.2 explores the main aspects related to the ethics and transparency of the results of AI models, namely privacy, transparency, discrimination and inclusion, in order to trace the objective of delineating the boundaries within which the use of AI methodologies do not put the security of users at risk.

Contents

1 **Introduction** 1

2 **How AI Models Are Built** 9

2.1 Processing of Unstructured Data in AI Models 10

2.2 Stand-Alone AI Models 19

References 27

3 **AI Tools in Credit Risk** 29

3.1 Use of Alternative Techniques and Data in Probability of Default Models 30

3.2 How to Improve Traditional Models Using AI Techniques 44

3.3 Applying an AI Model to Asset Management 48

3.4 Use of AI Models for the Validation/Benchmarking of Traditional Models 56

References 64

4 **The Validation of AI Techniques** 65

4.1 Possible Comparison Criteria Between Traditional Models and AI Models 66

4.2 Interpretability and Stability of the Models' Outcomes 71

References 78

5 **Possible Evolutions in AI Models** 81

5.1 The Role of AI Models in the Credit Risk of Tomorrow 81

5.2 Ethics and Transparency of Results 84
References 91

Appendix 93

Glossary 97

Bibliography 101

Index 103

About the Authors

Rossella Locatelli is Full Professor of Banking at the University of Insubria, Italy. She graduated in Economics and Banking Science at the Università Cattolica del Sacro Cuore, Italy, where she was a researcher until 1998. She is also the co-manager of CreaRes, the Business Ethics and Social Responsibility Research Centre, and manager of Criel, the Research Center on Internationalization of Local Economies. She serves and has served as board member of some listed companies, banks, insurance companies and other financial companies.

Giovanni Pepe is KPMG Partner since May 2015 where he leads the quantitative advisory practice within the Financial Risk Management Line of Service. As Director of the Bank of Italy, first, and as Adviser of the European Central Bank, later, he led many on-site inspections at several of the largest Italian and European banks on a variety of topics, including the validation of internal credit, market and counterparty credit risk models. He graduated from Federico II University of Naples and specialized in finance at the London School of Economics.

Fabio Salis is Chief Risk Officer of Creval since 2018. Formerly, he was Head of Risk Management at Banco Popolare since 2012, where he led important projects such as validation of credit and operational risk models and EBA stress test. He graduated from University of Pavia (Economics Department), specializing in quantitative methods.

List of Figures

Chart 2.1 Frequency of words present in the reasons for payment of transactions (*Source* Own processing) 16

Chart 2.2 Frequency of bigrams present in the reasons for payment of transactions (*Source* Own processing) 16

Chart 2.3 Results of the topic modelling analysis using LDA (*Source* Own processing) 18

Chart 2.4 Cluster analysis of transactional data (*Source* Own processing) 19

Chart 2.5 An artificial neural network (*Source* Own processing) 27

Chart 3.1 SHAP summary plot (*Source* own processing) 35

Chart 3.2 Customer with irregular payments of salaries (*Source* own processing) 37

Chart 3.3 Customer with regular payments of salaries (*Source* own processing) 37

Chart 3.4 Customer with seasonal payments of salaries (*Source* own processing) 37

Chart 3.5 Performance of the transactional score during COVID-19 (*Source* own processing) 39

Chart 3.6 Transactional score trend (*Source* own processing) 40

Chart 3.7 Importance plot (*Source* own processing) 42

Chart 3.8 ROC curve (*Source* own processing) 42

Chart 3.9 Performance comparison of ML techniques (*Source* own processing) 43

Chart 3.10 Relationships between risk drivers and the scores of the linear and random forest models (*Source* own processing) 49
Chart 3.11 The process (*Source* own processing) 51
Chart 3.12 The decision-making process (*Source* own processing) 54
Chart 3.13 Error convergence as the number of iterations increases (*Source* own processing) 62
Chart 5.1 Trend in the number of working papers published according to an analysis by The Economist 83

List of Tables

Table 3.1 Performance of the linear model compared to the random forest model 48
Table 3.2 Performance of the internal model compared to the neural network 62
Table 3.3 Performance of the internal model compared to the neural network 64

CHAPTER 1

Introduction

Abstract Nowadays that new masses of data have become available, and the artificial intelligence (AI) techniques have become more interpretable, the financial service industry is investing on the development of AI models. In particular, alternative types of information have become accessible due to the business relationships of banks with their customers, the progressive digitalisation of the economy, the availability of information of the websites, newspaper articles and social media, the COVID-19 pandemics. Those types of data can be used with different purposes to enhance several aspects of the credit risk modelling: early warning, provisioning, benchmarking, loan granting and risk discrimination.

Keywords Digitalised data · Artificial intelligence · Machine learning · Processing capabilities · Alternative information · PSD2 · Scoring · Early warning systems · High-frequency data

The growing availability of increasingly digitalised data flows is changing the way we interpret the lives of people and companies. It is now clear that companies stand to gain real competitive edge by harnessing these data and seizing upon sophisticated processing techniques to cut costs and implement decision-making processes that are as data driven and automated as possible and also better suited to the characteristics of their counterparties.

R. Locatelli et al., *Artificial Intelligence and Credit Risk*,
https://doi.org/10.1007/978-3-031-10236-3_1

Now that an extensive, unstructured mass of data is available, the ability to harness developments in artificial intelligence (AI) has become a critical success factor for the banking and financial service industry, which is, by its very nature, specialised in interpreting and generating information. One of the challenges for banks, therefore, lies in its ability to get a leg up by investing in new data processing capabilities and using new processing techniques, as they become available.

Due to the type of activity that they perform in the economic system, banks have always needed to gather and process information to be used in business activities like lending and pricing decisions, in monitoring exposures for risk measurement and management and in performance assessment.

In addition to the information that banks acquire as part of their business relationships with customers, for years now, intermediaries have been able to acquire additional information externally (macroeconomic data, the credit reports in feedback flows from the Bank of Italy's central credit register and information generated by other credit bureaus).

The field of "additional" or "alternative" information has substantially expanded in recent years as society and human behaviour evolve and regulations change.

Specifically, the progressive digitalisation of the economy means that businesses and individuals now generate vast quantities of information, most of which is unstructured and not necessarily financial, that can be used for, among other things, customer profiling, a practice widely embraced by Big Tech for commercial purposes and, particularly, to tailor their commercial proposition to customers.

For example, this information includes news published on websites, newspaper articles and social media posts and comments—which taken together help define a person's or a company's online reputation—and the extensive data describing the habits of various categories of operators, starting with the day-to-day financial transactions they perform.

In respect of the latter, it is important to emphasise that, in recent years, the use of debit and credit cards has skyrocketed in Italy, partly due to regulations designed to thwart the underground economy. The number of active POS terminals rose from 1.8 million in 2014 to 3.4

million in 2020.[1] This trend has obviously driven a large number of transactions that were previously settled in cash onto national and international e-payment circuits, generating information hitherto unavailable.

A major regulatory development, the second European Payment Services Directive (PSD2), then made this information more usable for credit scoring purposes.

The new Directive has made information arising from financial transactions much easier to use as, on the one hand, it has required financial intermediaries that stored the transactions to give third parties access to information that was, until then, exclusively the prerogative of such intermediaries while, on the other, it has made it possible for the same banks to use third-party information when the potential new customers allow to do so.

In this context of greater availability of "alternative" data, the COVID-19 pandemic, which triggered an increase in the number of electronic transactions but also affected the overall uncertainty about macroeconomic conditions, reinforced the need for banks to expand their traditional datasets to include information updated in real time or gathered from less consolidated sources. This was specifically in response to the need to compensate for the loss of sensitivity demonstrated by "traditional" rating instruments during the pandemic, an actual regime change in everyone life, that in the economic sector was characterised by unprecedented series of relief measures in the shape of banking loan moratoriums and new financing secured by government guarantees, fiscal stimulus to support businesses and workers, bans on employee dismissals, extraordinary changes in monetary policy and supervisory measures.

As they impacted the "traditional" information typically used in credit rating systems (e.g., loan moratoriums make it more difficult to assess the evolution of a borrower's financial position and ability to repay), these measures crippled the rating systems' ability to correctly gauge the level of risk associated with customers, risk that could abruptly materialise when the various relief measures were lifted (i.e., the cliff edge effect).

This context, therefore, reinforced the need for intermediaries to expand their datasets and deepen their analyses of the data they already had in order to identify, far enough in advance, any "zombie" borrowers surviving solely thanks to the relief measures and for which the banks needed to determine the most appropriate steps to take in order to

[1] Bank of Italy Payment System Statistics—March 2021.

prevent potential defaults and to adopt the most appropriate strategies to mitigate credit risks.[2]

The continuous enrichment of the information available to banks beyond that traditionally available to them has given their risk management departments the opportunity to revise some of the structural characteristics of credit risk models, adjusting various aspects for efficient credit monitoring.

First, the availability of additional data beyond a company's financial statements, bank performance data (i.e., internal and external performance data) and the information released by credit bureaus has made it possible to raise the degree of the models' precision, thereby improving their contribution to pinpointing opportunities for development, as well as in the measurement of risks.

Second, the fact that the new information is "high frequency", which is to say available in real time or close to it, makes it possible to construct models that are more sensitive than traditional models to changes in the characteristics of borrowers, making them more forward-looking or at least less backward-looking.

To move forward on this developmental path, it is essential to use machine learning (ML) techniques that, enabled by the astonishing progress in the computational capacity of banks' IT systems, may be decisive in capturing the non-linear relationships between explanatory variables and the target variable. Moreover, with these techniques, the modelling is "less constrained" than traditional modelling because it does not require prior assumptions regarding the expected relationships between variables, the distribution of those variables, the number of estimable parameters and other elements that make up the estimation model.

Nevertheless, the true challenge lies not only in building models that are methodologically and technologically sophisticated but foremost in building them so that they produce results that can be more or less easily

[2] Moreover, the priorities of the ECB Banking Supervision of 2021 include: "[…] banks' capacity to identify any deterioration in asset quality at an early stage and make timely and adequate provisions accordingly, but also on their capacity to continue taking the necessary actions to appropriately manage loan arrears and non-performing loans", as reported at www.bankingsupervision.europa.eu.

explained to the various users of the outputs of the credit risk management models in the various fields where they are used.[3] These fields include: business strategies—where the output of the models must be explained to customers (to whom it must be possible to justify the credit decisions made) and to credit managers (who must be able to understand the forces underlying a decision to accept, deny or revise a loan already in portfolio); the quantification of credit losses on loans in portfolio; and the parameters used to calculate capital requirements. In this latter case, the need to explain models' outcomes is definitely high as banking regulators investigate carefully the process followed by banks to compute capital requirements.[4]

The outputs of the models must, therefore, be explainable to multiple parties, both inside and outside the company.

A recent survey by the Institute for International Finance (IIF)[5] confirms the fact that one of the main areas of application of ML techniques is credit scoring. Specifically, the survey shows that while the use of these models for supervisory reporting purposes is somewhat limited by the need to implement simple and easily interpretable models, these limitations apply to a lesser extent when the same techniques are used for management purposes.

One important aspect that concerns both the "supervisory reporting" and "management reporting" applications equally is the integration of the traditional data used for credit risk management (e.g., credit bureaus) with data that are, by their nature, "innovative" (e.g., transactional data). Their integration can significantly contribute to making estimates more precise and timelier.

For example, while credit bureau data based on prevalently historic information can produce accurate estimates, these data update slowly and

[3] Institute of International Finance (IIF) Briefing Note (June 2021)—Explainability is a trust issue. Regulators are concerned with levels of complexity that offer limited or no insight into how the model works (i.e., so-called black boxes). They want to know how models, including AI and ML, reach decisions on extending or denying credit, whether FIs have appropriate risk controls in place, and the like. A challenge of using AI/ML models is often the lack of transparency, which is imperative to building trust with customers and other stakeholders. Banks have long had to explain their decision processes to improve confidence in the robustness of a model.

[4] Par. 179 CRR 575/2013: "The estimates shall be plausible and intuitive".

[5] Institute of International Finance (2019). Machine Learning in Credit Risk.

do not immediately reflect current market conditions. This was particularly clear in recent times with the public health emergency caused by COVID-19 and the resulting macroeconomic uncertainty, when traditional information which, as noted earlier, already suffers a physiological time lag, became an unreliable expression of the creditworthiness and the real potential resilience of companies.

In this context, the use of transactional data—which are, by definition, point-in-time—marks a considerable step forward in improving the predictive accuracy of the models and in overcoming the lower sensitivity of traditional models in singular circumstances like those created by COVID-19.

As highlighted, for the purposes of management reporting, there are many potential applications for ML techniques applied to innovative, high-frequency data.

First, they may be used in advanced early warning systems (EWS) for the early identification of borrowers whose behaviour "tends" to reflect, with respect to a target event (i.e., default), a potential risk on a forward-looking basis. While traditional systems usually require indicators based on expert opinions, ML techniques are well suited to handling large quantities of high-frequency data/updates and make it possible to generate effective and efficient early warnings.

Again in this case, the COVID-19 pandemic rendered more pressing the need to integrate traditional EWS models with additional components that are better at "interpreting" extraordinary conditions, for instance by recognising, through the most recent transactional trends, borrowers' actual resilience, and thereby limiting false alarms.

In some ways, a similar case is the IFRS 9 model of allocating financial assets to one of three stages among which obligors must be bucketed, in which ML techniques make it possible to capture farther in advance a borrower's inclination to migrate to another stage as per the IFRS staging allocation rules.

Loan granting processes provide another context in which the potential of AI techniques may be harnessed. The added competitive edge gained from the in-depth analysis of data and high-speed processing and decision-making comes from the ability to gather core information (e.g., income, consumption, etc.) from non-traditional data sources and to extract patterns that are relevant for the purposes of monitoring credit performance over a period of time that is usually more extended in comparison to, for instance, the one year adopted for monitoring purposes.

Additional areas in which artificial intelligence techniques may be used relate to non-performing loans (NPL) and particularly the clustering of customers that have defaulted in order to select the most appropriate credit recovery methods and to identify the portions of the distressed portfolio to be sold to specialized operators.

In the various examples of applications given here, the underlying idea is the search to combine algorithmic intelligence, which handles vast quantities of quantitative data with ease, with human intelligence, which places the risk assessments in the overall context and management processes. This is the underlying idea that guides the position paper.

CHAPTER 2

How AI Models Are Built

Abstract This chapter describes the various kinds of data that are mostly in use today in AI models, differentiating between "structured", "semi-structured" and "unstructured" data. Text analysis and Natural Language Processing are illustrated as the main structuring techniques for unstructured data. Some examples of alternative credit data are described, including among others transactional data, data extracted from telephones and other utilities, data extracted from social profiles, data extracted from the world wide web and data gathered through surveys/questionnaires. Also, the chapter describes the opportunity of estimating a model only by means of machine learning techniques, detailing the characteristics of the most used ML algorithms: decision trees, random forests, gradient boosting and neural networks. The application of a special type of neural network is detailed: the autoencoder.

Keywords Structured data · Unstructured data · Text mining · NLP—Natural Language Processing · Topic modelling · Part-of-speech tagging · Sentiment analysis · Data cleansing · Dictionary · Alternative credit data · Transactional data · Machine Learning—ML · Artificial Intelligence—AI · Decision tree · Random forest · Gradient boosting · Neural network · Autoencoder

R. Locatelli et al., *Artificial Intelligence and Credit Risk*,
https://doi.org/10.1007/978-3-031-10236-3_2

2.1 Processing of Unstructured Data in AI Models

The best practice for the optimal use of one's data in the development of cutting-edge credit risk management models lies in the ability to set up methods for the integration of the various types of information (structured, semi-structured and unstructured) available on the counterparties to be evaluated in the various business processes.

"Structured" data (typically used in "traditional" credit risk models) are data that meet a set of predetermined rules or that can be defined in terms of type (date, name, number, characters and address) and mutual relationships.[1]

In addition to "structured" data, there are "semi-structured" data, which contain semantic tags without the typical structure associated with relational databases. These data are not schematic and not suitable for a relational database. They are represented using labels, charts and tree structure. A few examples of semi-structured data are e-mails, HTML files and XML files used mainly to transmit data between a server and a web application.

"Unstructured" data do not have a predefined model and cannot be organised in rows and columns. Examples of unstructured data are images, audio recordings, videos, e-mails, spreadsheets and objects stored as files. Unstructured data can originate from a wide variety of sources. For example, they may be extracted from human language with natural language processing (NLP), gathered through sensors, extracted from social media or collected from NoSQL databases. This makes them difficult to understand and ambiguous to classify.

Since most of the information available is unstructured, it is easy to see why, especially in data-driven companies, it has become crucial to analyse unstructured data in order to identify buying habits, capture new trends, shape commercial offers and provide information on how to improve a specific service or the company as a whole.

[1] Structured data are based on a scheme and may be represented by rows and columns and filed in a central repository, typically a relational database, where the pieces of data may be retrieved separately or in a variety of combinations for processing and analysis.

2.1.1 The Main Structuring Techniques for "Unstructured" Data Are Text Analysis and Natural Language Processing

The extraction of information from data in a text format is one of the most complex data analysis processes. There are many reasons for this. On the one hand, as we mentioned earlier, when text is extracted, the data are not structured. There are no observations or clearly defined variables (rows and columns). This means that to perform any type of analysis, the unstructured data must first be converted into a structured dataset before moving on to the normal modelling framework.

On the other hand, and more generally, the field of natural language processing is certainly one of the most difficult aspects in the modern disciplines of artificial intelligence. Human language is rarely precise or written without ambiguity. Understanding human language means understanding more than just the words, but the concepts and how they are interconnected to create meaning. Although language is one of the easiest things for the human mind to learn the ambiguity of language is what makes natural language processing so difficult for a computer to master.

NLP techniques entail the application of algorithms to identify and extract natural language rules so that the unstructured data in the language can be converted into a format that is comprehensible to computers. Considering the topics addressed in this paper, the most important procedures include, but are not limited to:

- *topic modelling*: this consists of creating tags for specific topics using key words from a document through, for example, LDA (Latent Dirichlet Allocation) which finds the key words to extract;
- *part-of-speech tagging*: this entails identifying the type of entity extracted, e.g., a person, a place or an organisation through named-entity recognition (nouns, adjectives, attributes, etc.);
- *sentiment analysis*: this is an extremely vast field under continuous development. Sentiment analysis may be used to classify the sentiment, opinion or conviction of a statement, from very negative to neutral to very positive. Developers often use an algorithm to identify the sentiment of a word in a sentence or they use sentiment analysis to analyse social media.

Cleansing the dataset is a crucial step in any type of data mining. Nevertheless, data cleansing is even more important when the data in the set are unstructured. The main activities in the preparation and cleansing of a text include, inter alia:

- elimination of special characters (e.g., #);
- standardisation of the text (e.g., converting all the letters to lower case);
- elimination of the "stop words",[2] which are those that are most often used and vary from one language to the next. In Italian, stop words are articles, conjunctions and prepositions. Other stop words may be, for example, commonly used verbs (to seem or to have) but only in certain conjugations (it seemed, they have);
- elimination of specific words that are not considered significant and elimination of numbers when they are not used or are unusable;
- *stemming* or *lemmatization*: this is a fairly complex process that consists of reducing an inflected word to its root form, or "stem". In stemming, the stem need not be identical to the root of the word (e.g., "go", "went" and "going" would be mapped on the stem "go"). With lemmatization (which is more complex), the words would be properly mapped on the verb "to go".

Just from this shortlist is it easy to see how the cleansing and preprocessing of data are one of the most time-consuming steps in any text mining process.

A dictionary makes it easier to convert "unstructured" data into a "structured" format. Various free dictionaries are available online for sentiment analysis (especially in English). However, for certain specific analyses, an ad hoc dictionary must be created with the elimination of everything that is not strictly necessary for the analysis, identifying the most frequent words, those that are "equivalents" or "related". For example, if we were to analyse a set of transactional data to see how many of these transactions relate to online purchases, many texts would probably include the word "amazon" but others would undoubtedly also include abbreviations like "amzn" or words that may, to some extent, be associated with Amazon "concepts", like "kindle" or "prime subscription". If the dictionary used does not contain the word "amzn", the transactions containing the word "amzn" will not be tagged as online

[2] Stop words are considered such because they create stops in the processing.

purchases and the analysis will consequently be incomplete. In this case, the imprecise results will be due to an incomplete dictionary.

In general, the files containing "unstructured" data tend to be much larger than those containing structured data and occupy far greater volumes than "structured" data, as far as into the Petabytes. In response, in recent years, new tools (like Hadoop, NoSQL or MongoDB) have been developed to store large datasets.

2.1.2 What Does "Alternative Credit Data" Mean?

Alternative credit data, also referred to as "big data", are all data not directly related to a customer's credit history. Alternative data on a customer may be taken from a series of non-traditional data sources (e.g., digital platforms that can provide information on consumer activity for credit risk assessments). They are, in general, a combination of information gathered from several sources, including payment chronologies, data extracted from social media, information about online purchases and much more. Below is a list of examples of the main sources of data that can be used in the customer analysis process to measure credit risk:

- *transactional data*: these are generally qualitative data not typically used in "traditional" models, for instance in reference to customers' use of credit or debit cards or bank transfers performed. These data may be used to generate a wide range of predictive features like the "ratio of the account balance to total spending in the last X weeks" or "expenditure ratios over various time horizons" or features may be developed based on the number, frequency and value of transactions for various types of purchases and/or retailers. This processing can be extremely time consuming because the data are not generally very "clean", in particular, to analyse the transactional data on bank transfers, the natural language in the reason for payment must be processed;
- *data extracted from telephones and other utilities*: these are based on the customer's credit history and they are considered alternative because they do not actually appear in most credit reports. One critical factor relates to privacy protection issues that arise when these data are used;
- *data extracted from social profiles:* data may be extracted from Facebook, LinkedIn, Twitter, Instagram, Snapchat and other social media

sites, but with considerable regulatory obstacles in the wake of recent European Union privacy protection laws. Theoretically, a frequently used way of overcoming these obstacles is to focus less on the actual data and more on the metadata, e.g., the number of posts and their frequency or the extension of the social network to which they belong. Moreover, these data can be *biased*, in that consumers may edit the underlying information, or insufficiently predictive, there being the significant issue of the textual data's actual predictive capacity, i.e., whether the indicator can effectively be used as predictor of, for instance, the borrower's creditworthiness;

- *data extracted from the world wide web*: the online reputation of a user, person or company may be monitored by gathering online data, reviews, comments or news to gain a real-time picture of the "web consensus" about a person, a company or product.
- *data gathered through surveys/questionnaires*: an innovative way of assessing credit risk when the borrower does not have a credit history is through psychometric tests.[3]

Looking at transactional data, the chronology of a banking customer's transactions contains an extensive set of information on the type, trend and level of their inflows and outflows. This information is obviously differentiated depending on the type of counterparty (consumers, small and medium businesses and larger, more structured companies). The transactions performed by an individual customer, for instance, may reveal a lot about the level and volatility of her/his inflows (e.g., a detailed analysis of the information in bank transfers) or about the level and type of his/her consumption, going so far as to arrive at information concerning the time of day he/she does shopping and the geographical location of his/her purchases.

Furthermore, these data provide an understanding of whether the consumer is mostly focused on e-commerce or if the spending is typically in-store.

[3] Liberati C. et al. "Personal values and credit scoring: new insights in the financial prediction", Journal of the Operational Research Society – February 2018. The word "psychometrics" refers to the set of psychological analysis methods for quantitative behavioural assessments. The leader in this field, the Entrepreneurial Finance Lab (EFL), bases its scores on 10 years of research at Harvard.

As mentioned earlier, this is a very broad set of information whose potential for credit profiling has been, to date, mostly untapped by banks, which, however, risk missing out on this opportunity in that, because of PSD2, which established open banking in the European Union, even the smallest fintech startup can develop products and processes using these data as long as they have access to the mass of data needed to train the models on which the products they offer rely on.

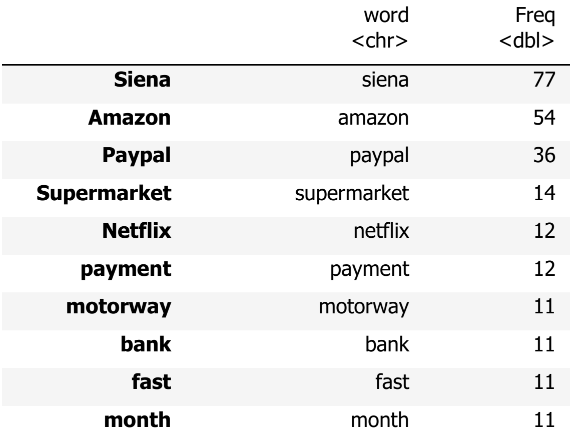
Information on transactions is often abbreviated using ambiguous conventions that frequently make it difficult for even the customers themselves to recognise their own transactions. Typically, raw data contain duplicates, long strings of complicated text and many numbers with very little explanatory power in the relevant context. First and foremost, it is, therefore, necessary to cut out the noise and translate the data into something that anyone can easily understand. The textual data standardisation and cleansing techniques that we mentioned earlier are extremely useful in this sense. For example, an unstructured string like the following:

‘date 23/06/21 time 00 00 loc 800–279-6620 operator: amazon it *283vi02 × 4 amnt orig curr 21.98 card no: 12345678’
can be easily transformed into a structured text with fields using elementary text analysis:

```
[1] “—————————”
   [1] “Transaction no. 1”
   [1] “date          23/06/21”
[1] “operator          amazon”
[1] “amount            21.98”
[1] “card_no.        12345678”
[1] “—————————”
```

A simple analysis of the frequency of the words that most frequently appear in transactions can give us much information on the characteristics of the payments. For instance, looking at a series of roughly 250 transactions by one of the authors (who volunteered them), we can, with just a few lines of code, create the following frequency chart (Chart 2.1).

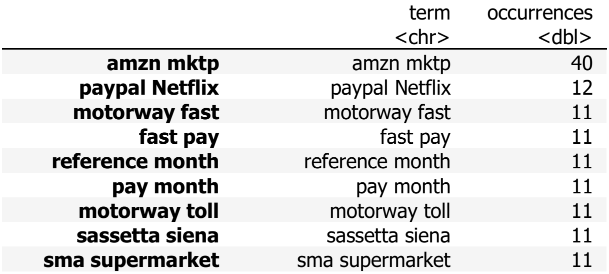
This enables us to easily gather information on the location (the payments were mainly made in Siena) and on the main purchases (mostly online and food). A steady number of online purchases over time could be used, for example, as features in the creditworthiness analysis. The regularity of streaming service payments might also be indicative for this purpose.



	word <chr>	Freq <dbl>
Siena	siena	77
Amazon	amazon	54
Paypal	paypal	36
Supermarket	supermarket	14
Netflix	netflix	12
payment	payment	12
motorway	motorway	11
bank	bank	11
fast	fast	11
month	month	11

Chart 2.1 Frequency of words present in the reasons for payment of transactions (*Source* Own processing)

Once the string of "cleansed" text is available, we can begin transactional data enrichment (TDE). The originally existing data may be enriched with a search for geographical information or information on the economic sector to associate with the information already present in the transaction, for example. A simple bigram frequency distribution analysis—bigrams being the words that appear in pairs—on the aforementioned data results in the following outcome (Chart 2.2).



	term <chr>	occurrences <dbl>
amzn mktp	amzn mktp	40
paypal Netflix	paypal Netflix	12
motorway fast	motorway fast	11
fast pay	fast pay	11
reference month	reference month	11
pay month	pay month	11
motorway toll	motorway toll	11
sassetta siena	sassetta siena	11
sma supermarket	sma supermarket	11

Chart 2.2 Frequency of bigrams present in the reasons for payment of transactions (*Source* Own processing)

This shows that the use of a payment service, for example, is used regularly for a subscription to a popular streaming service, that motorway tolls are mostly paid by credit card and even which supermarket is most frequently used for purchases.

Lastly, once the transactional data have been cleansed and are clear, we can extract even more information, organising them into categories. One objective of clustering algorithms is, for example, understanding that a given transaction is an "online" purchase, that the money that the customer spent at "Pizzeria XXX" was spent in the "restaurants" category and that purchase xyz was very probably a "food" purchase.

This categorization may be automated (or semi-automated) using AI techniques like LDA or unsupervised clustering processes (reference should be made to the bibliography for a description of these). Chart 2.3 shows the results of a typical topic modelling analysis using LDA:

As shown, three expenditure categories have been identified, reasonably attributed to online purchases and payments, ATM withdrawals and payments and, lastly, purchases on premises. Even a classic cluster analysis shows mainly these three spending categories (Chart 2.4).

The business model that many fintech companies specialised in consumer credit automation[4] have adopted is based on a similar clustering process. The most interesting aspect, in some ways, (at least from the standpoint of a bank) is that these companies use open banking (PSD2) as a data source because they do not have, or do not yet have, the same volume of information that banks have on their customers. After this data gathering and processing process, they typically use ML and AI to construct their risk models.

Lastly, it is useful to bear in mind that even the European Central Bank promotes the use of transactional data (at least for the purposes of early warning indicators), as stated in a recent letter sent to the CEOs of major European banks in the aftermath of the COVID-19 pandemic.[5]

[4] For example, from Faire.ai, fintech B2B (https://www.faire.ai/).

[5] SSM-2020–0744, "Identification and measurement of credit risk in the context of the coronavirus (COVID-19) pandemic", Frankfurt am Main, 4 December 2020.

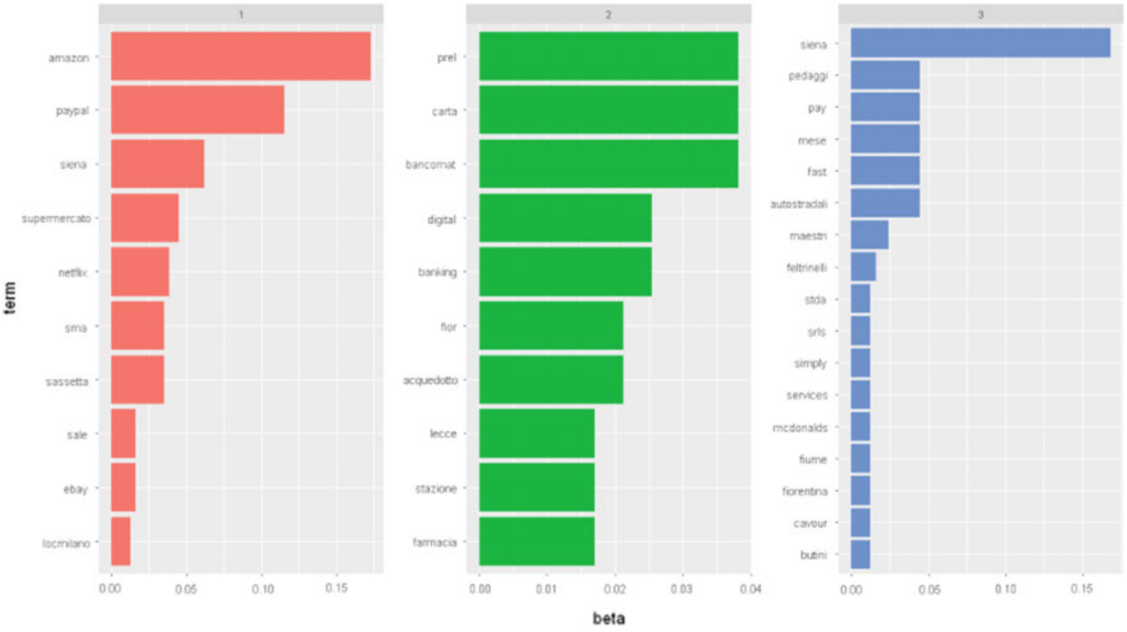
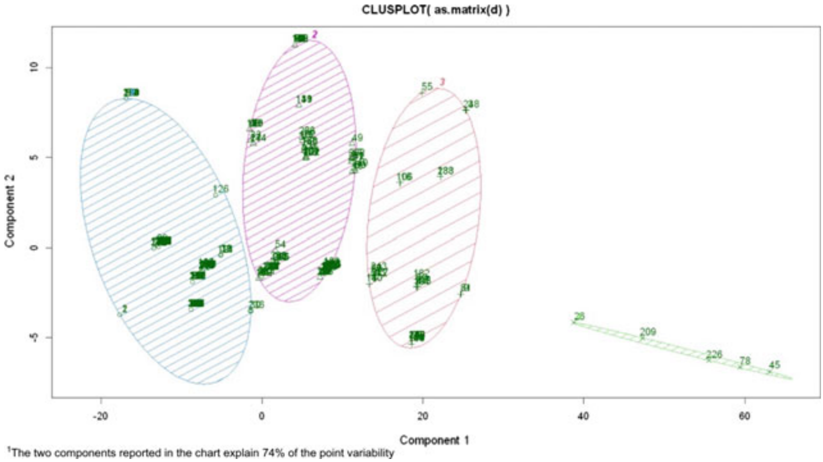


Chart 2.3 Results of the topic modelling analysis using LDA (*Source* Own processing)



¹The two components reported in the chart explain 74% of the point variability

Chart 2.4 Cluster analysis of transactional data (*Source* Own processing)

2.2 Stand-Alone AI Models

In the scope of the ongoing updates to the Basel 2 framework (as we write, the European Commission has published its proposed amendment to Directive 2013/575/EU, the Capital Requirements Regulation, also known as "CRR3"), the modelling of the risk parameters used to calculate capital requirements for supervisory reporting purposes has been mainly geared towards statistical approaches that make the models fully interpretable (art.174 CRR—"*Use of models*"[6]).

This requirement is met—in prevalent practices for estimating credit ratings—using regressive models (i.e., logit functions), that enable a direct parametric association (through the β estimated by the regression) between each variable used in the model and the default event. This is

[6] Art. 174: "a) [...] The input variables shall form a reasonable and effective basis for the resulting predictions [...]; e) the institution shall complement the statistical model by human judgement and human oversight to review model-based assignments and to ensure that the models are used appropriately. Review procedures shall aim at finding and limiting errors associated with model weaknesses. Human judgements shall take into account all relevant information not considered by the model. The institution shall document how human judgement and model results are to be combined".

true in terms of both "directional" impacts between the individual variable and the riskiness (the coefficient may be positive or negative) and "magnitude" impacts (the coefficient is more or less large as a number).

Recently, the traditional system of rating models has undergone a series of potential changes due to the combined effect of, inter alia, the following:

- the need for ever more accurate risk estimates, even for reasons not directly related to calculating the supervisory capital requirement (e.g., early warning models, models to support pre-approval and fast lending processes);
- the chance to include forward-looking information in estimates, whose assessment requires the use of information that, until now, had not been fully exploited, such as information that can be gathered from the aforementioned transactional data or that relating to the borrowers' economic micro-sectors;
- the availability of new sources of information, including unconventional information, which was previously unavailable or, in any case not available and not collected or historicised so as to be used;
- the significant advancement of computational capacity thanks to the most recent technologies, now capable of fully enabling AI techniques.

These challenges have compelled banks' risk management departments to begin reflecting on the feasibility of gradually evolving from traditional methods to more sophisticated techniques and these reflections—often conducted in concert with the organisational units responsible for transforming banks into data-driven organisations—have led players to consider the methodological approaches known in the past, like decision trees, random forests and gradient boosting.

Approaches like these—considered grey boxes in terms of their interpretability, as the link between the input data and the risk measurement output is less clear than with traditional techniques—are essentially based on iterative decision-making algorithms. Neural networks are at the cutting edge of AI technologies and here, once again, we find methodologies that have been around for some time without having been widely applied in credit risk measurement because they were considered black boxes in terms of interpretability of their results.

The various methodologies described above are summarised below, with indication of their strengths and weaknesses.

2.2.1 Decision Trees

These models are based on an algorithm that has been widely documented in the literature. The structure of a decision tree is simply a series of nodes and branches. The nodes represent a macro-class of input variables, in which each node corresponds to a specific linear test function, which establishes the appropriate partitions of the input variables. With binary classification, the simplest type, the most frequently used algorithm in the literature is based on the Gini coefficient or entropics (Kamber et al. 1997).

The branches or, to be more precise, the arches, represent all the properties, or the splitting rules, which determine the path within the tree and, finally, the classifications. These properties/rules are defined in relation to the specific values assumed by the attribute identifying the parent node.

Decision trees owe most of their popularity to their simplicity and the ease with which they can be interpreted and translated into "if–then" scenarios.

There are substantially two preliminary stages in building a decision-making algorithm:

- building, in order to expand the size of the tree (more arches and nodes) and to define the properties/splitting rules (arches) capable of classifying the outputs (nodes). The natural conclusion of this stage is a particularly "bushy" tree, which could lead to overfitting issues[7];
- pruning, to reduce the size of the tree, but without reducing the algorithm's predictive capacity.

Alternatively, an early-stopping strategy may be used, in which the pruning is carried out earlier. Specifically, this approach avoids the risk of overfitting the tree, as a minimum "improvement" value is established

[7] Overfitting is when a statistical model contains an excessive number of parameters compared to the number of observations and therefore achieves an excellent performance on the training set, but a weak performance on the validation sets.

(i.e., information gain and/or reduction in the impurity index) when going from one node to the next one in the building algorithm: when, at the nth node, the improvement is less than the early-stopping criterion, the building algorithm will end.

The decision trees described earlier are very efficient when they work with categorical variables, the classification is expressed through explicit conditions, the algorithm itself determines the significant variables and the presence of outliers does not excessively affect the result (typically the outlier data will be isolated in peripheral nodes).

Conversely, this estimation approach taken by the model precludes incremental learning,[8] typically requires many training sets to avoid overfitting and entails an unstable classification (a completely different tree is created if just a few lines of the training set are removed) with an algorithm that only breaks down one variable at a time. Moreover, the conditions can only be inferred from the algorithm and not set as hypotheses, even where they are known and, finally, the results of the regression never fall outside the range of the training data, so the algorithm can only predict target variables in the range observed in the training set.

2.2.2 Random Forests

One widespread type of model based on decision trees entails replicating many times (even over 1,000) the estimate of the tree using only one subset of the available variables each time.

This classification method is called random forest and is based on a regression and classification algorithm that creates a forest of decision trees constructed out of multiple datasets, extracted through bootstrapping (random sampling that organises the various nodes and divides them randomly). The more trees there are in the forest, the better the results of the model will be. It is important to have a low correlation between the models entered: in this way, each decision tree created takes independent decisions and also reduces individual errors, in that a sort of protective barrier is formed between one tree and another. Certain trees may generate distorted results, whereas others will generate correct results, leading the model in the right direction.

[8] Incremental learning is when statistical models learn as they acquire new data.

Two methods may be used to avoid an undesired correlation between the behaviour of one tree and that of the others:

- the first, i.e., bootstrap aggregation (or bagging), reduces the variance and improves the accuracy of the prediction. Bootstrapping is usually based on sampling with replacement to make the selection procedure completely random. The model's predictions are then aggregated so that the final prediction considers all possible results. If the model provides for a class, the predictions can be aggregated by selecting the most frequently estimated class or, if the model provides for probability, by calculating the average of the estimated probabilities and therefore choosing the class with the highest probability;
- the second, based on the "randomness", considers every possible characteristic when the node is divided and chooses the characteristic that produces the greatest separation between the observations in the left node and those in the right node (differentiated on the basis of the classification driver). On the other hand, every tree in a random forest can only choose one random subset of characteristics. This requires an even greater variation between the trees in the model and translates into lower correlation between the trees.

The random forest algorithm is extremely efficient when working with categorical variables as it itself determines the significant variables. Any outliers will not excessively influence the outcome (they are typically isolated in peripheral nodes), and it does not present the overfitting problem present in "traditional" decision trees.

On the other hand, the main limits of this methodology are that the classification only allows for one ranking of the impact of the various attributes on the outcome, incremental learning is not possible and the result of the regression, as in the case of decision trees, can never fall outside the range of the training data.

2.2.3 Gradient Boosting

Gradient Boosting are some of the most well-known AI models. The main objective of gradient boosting is to minimise the residual error of the model through a sequence of decision trees that, at each step, learns from its previous errors. There are various variations of gradient boosting methods, such as light gradient boosting, XGBoost, CatBoost, etc.

The gradient boosting algorithm begins with a calculation of the residual error in the base model, often a linear model. In the next step, the error in the base model is predicted using a decision tree with the same variables as the base model. Then, in the next step, the residual error of the base model combined with the first decision tree is calculated and another decision tree is used to calculate and keep track of the new residual error using the same list of variables. These steps are repeated until the early-stopping criteria are reached. This algorithm is efficient when working with categorical variables, as it itself determines the significant variables, and its advantage is that outliers do not excessively influence the outcome beyond creating additional "versions" of it.

On the other hand, substantial computational efforts are required to estimate and calibrate the model and incremental learning is not possible.

2.2.4 Neural Networks

Neural Networks models constitute the last frontier in AI techniques applied to data modelling, be they conventional or non-conventional data. These methodologies are often thought of as a "black box" in terms of the association between the input parameters and the output classification.

Neural networks are based on a model that puts the explanatory variables in communication with the target variables through various layers of latent variables, referred to as "hidden layers", consisting of combinations of transformed input variables. A neural network is indeed an "adaptive" system capable of modifying its own structure (the nodes and interconnections) based on both external data and internal information that interconnect and move through the neural network during the learning and reasoning phase.

A biological neural network receives data and external signals (humans and animals perceive them through their senses through complex organisations of nerve cells with different tasks like perceiving the surrounding environment and recognising stimulus), which are then processed into information through a massive number of interconnected neurons (which constitute calculation capacity) in a non-linear, variable structure in response to the data and external stimuli.

Similarly, artificial neural networks are non-linear structures of statistical data organised as modelling tools. They receive external signals on a layer of nodes (i.e., the processor) and each of these “entry nodes” is connected to various internal nodes in the network which are typically organised into several levels so that each individual node can process the signals received, transmitting to the next levels the result of its processing (i.e., more sophisticated, detailed information). An artificial neural network typically has at least three layers. First is the input layer, which accepts the incoming data (each neuron in this layer represents a feature of the dataset), another is the output layer, which provides the result (i.e., the output) of the neural network’s processing, plus one or more hidden layers between the input layer and the output layer.

It is the hidden layers’ task to use the features of the dataset to learn new features. A neural network is defined as “deep” when it contains two or more hidden layers. In such cases, the network will use the features it has learned in a hidden layer to learn additional new features that are even more significant in the next layer, and so on until they are used for classification, regression or clustering in the output layer.

Learning from experience occurs inside the networks, which can make conclusions based on a complex and seemingly uncorrelated set of information. In a neural network, changing the weight of any connection has an effect on the results processed in all the other neurons in the next levels.

One of the most important positive elements of neural networks is that they work extremely well for both classification and regression in deterministic problems. Furthermore, the classifications that they produce are very robust for noisy data and they are also capable of exploiting relationships between features that are difficult to identify using other modelling approaches.

On the other hand, neural network techniques require a particularly large number of training sets. Features that are irrelevant to the classification may weaken the outcome and, as noted earlier, neural networks are black boxes that do not show the real random relationships between the variables.

Given the trade-off between the possibility of simultaneously processing an extremely large amount of data and the need to have a particularly large number of robust training sets to prevent overfitting, in credit risk modelling, these techniques offer particular added value in estimates of retail portfolios, which are highly standardised and present homogeneous information sets, normally not requiring any judgement and expert-based decisions in the creditworthiness assessment process.

In the corporate segment, especially for larger companies presenting more layers of operational complexity, such as multi-industry businesses and/or multinationals or when company operations are vulnerable to geopolitical forces that are difficult to model (e.g., the presence of incentives to adopt ESG and green policies or the effects of the implementation of Italy's national recovery and resilience plan in the wake of the pandemic), the exclusive use of AI models without human intervention would lead to less immediate results.

The next section describes a few ways to internalise the soft information of business experts in AI algorithms aimed at assessing corporate borrowers.

2.2.5 Autoencoder, a Special Type of Neural Network

Autoencoder neural networks are developed to generate new data by compressing the input in a latent space and then reconstructing the output based on the information gathered. They may, therefore, be used to detect outliers as well. The great advantage to this is that it identifies observations that would otherwise lead to large estimation errors.

This type of artificial neural network consists of two parts (Chart 2.5).

- the encoder is the part of the network that compresses the input in a latent space and that can be represented by the coding function h = f(x);
- the decoder is the part of the network that reconstructs the input based on the previously gathered information. It is represented by the decoding function r = g(h).

The autoencoder can, therefore, be represented as the function d(f(x)) = r, where r is the closest to the original input x.

The objective of training the autoencoder is to create a latent space h with features that are helpful for the analysis. This objective is achieved by setting limits on the coding, forcing the space to be smaller than x. In this case, the autoencoder is referred to as undercomplete. By training the undercomplete space, the autoencoder recognises the most relevant features of the training data.

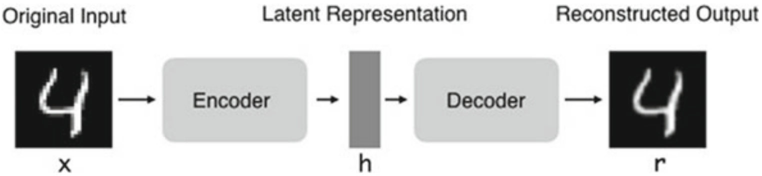


Chart 2.5 An artificial neural network (*Source* Own processing)

If the autoencoder is overcomplete, the size of the latent space is greater than the input. In this case, the input can be copied to the output without learning anything about the data, even with a simple linear encoder and decoder.

The application of this algorithm is described further on in the business case "Unsupervised algorithms – autoencoder neural networks used to detect outliers" (Chapter 3).

References

Kamber et al., "Generalization and Decision Tree Induction: Efficient Classification in Data Mining", 1997.

Liberati C. et al., "Personal values and credit scoring: new insights in the financial prediction", Journal of the Operational Research Society, February 2018.

SSM-2020–0744, "Identification and measurement of credit risk in the context of the coronavirus (COVID-19) pandemic", Frankfurt am Main, 4 December 2020.

CHAPTER 3

AI Tools in Credit Risk

Abstract This chapter describes four types of application of AI into Credit Risk modelling. The use of alternative transactional data together with the application of machine learning techniques in the context of the Probability of Default (PD) parameter estimation leads to enhancements of the PD models, able to capture phenomena that were not properly explained by the traditional models. Some examples are described in this paragraph: risk discrimination for borrowers with seasonal business, identification of counterparty risk during the COVID-19 crisis, early warnings and advanced analytics in loan approval Several combinations of traditional modelling techniques and AI techniques can be used to enhance the outcome of the credit risk models. In particular, the business case "two-step approach" is described, detailing the intervention of the AI techniques in a second phase of the model estimation, when the traditional techniques already produced a result. The third part of the chapter describes the application of an AI model to asset management. The model is aimed at supporting an asset manager's investment decisions. The last section of the chapter describes how to implement machine learning techniques with benchmarking purposes in the context of the validation of credit risk models used for the estimation of the regulatory capital.

Keywords Probability of default—PD · Machine learning—ML · Structured information · Unstructured information · Transactional data · Current account transactional data · POS transactional data · Payment cards · Social network data · Interpretability of results · SHAP summary

R. Locatelli et al., *Artificial Intelligence and Credit Risk*,
https://doi.org/10.1007/978-3-031-10236-3_3

plot · Early warning · COVID-19 · Advanced analytics · Loan approval · Risk discrimination · Seasonal business · Counterparty risk · Business case · Use case · Artificial intelligence—AI · Traditional modelling techniques · Traditional techniques · Logistic regression · Logistic model · Linear model · Linear regression · Two-step approach · Combination of techniques · Business case · Random forest · Asset management · Semi-liquid investment grade · Illiquid investment grade · Unstructured data · Neural network · Deep neural network · Validation · Credit risk models · Credit risk · Benchmarking · Credit risk assessment · Traditional model · Traditional modelling techniques · Managerial model · Regulatory capital

3.1 Use of Alternative Techniques and Data in Probability of Default Models

The extraordinary increase in the availability of data, new and sophisticated analytical techniques offering added value and customers' rising expectations of a complete digital experience are key factors that are increasingly crucial to market success of nowadays financial firms.

Against this backdrop, the models traditionally adopted by banks to assess the creditworthiness of their clients, or to estimate the probability of default (PD) of their counterparties, in the financial risk managers jargon, are also evolving, to make the most of these opportunities.

More in detail, within their PD models banks are adding up to the traditional modules aimed at exploiting "traditional information", new modules that that will exploit:

– internal data sources that are available and proprietary to the bank but not yet used;
– new data sources, including those outside the bank;
– new algorithms (i.e., ML, which we will explore in the paragraphs further on).

All this makes it possible to obtain online ratings, i.e., calculated in real time, suitable for more efficient and accurate automated processes and capable of improving both performance and the customer journey and keeping track of the impacts of extraordinary events (e.g., the pandemic).

The ML component can be introduced in the various credit risk models in a variety of ways, considering different types of alternative and innovative data and various algorithms capable of modelling a more granular type of data than that typically used to quantify borrowers' creditworthiness.

The business cases described in this paragraph reflect the inclusion of innovative data in the internal rating models that estimate PD ("PD models").

All this is possible thanks to new types of information included in the models through one or more specific ML modules, developed individually and then integrated with other modules (also developed individually) using typical integration techniques.

The traditional modules generally cover certain types of information areas, which may vary depending on the type of borrower (e.g., socio-demographic, financial, performance, quality).

The ML modules are meant to introduce a new type of information that, in the business cases considered, corresponds with innovative data, such as data on current account transactions (the bank's internal information or information acquired from other banks) and data extracted from social networks. This innovative version of PD models can be used for both supervisory and management reporting purposes.

In particular, the business cases of two large Italian banks are described in this paragraph.

The first application ("Model 1") refers to the inclusion of 4 innovative data modules in the PD model of the bank's SME Retail segment. The model was validated by the ECB and is currently used to calculate the supervisory capital requirement for the credit risk of borrowers to whom the model applies.

The second business case ("Model 2") refers to the inclusion of an internal transactional data module in the PD model for management purposes of the bank's Small Business segment (i.e., companies with turnover of less than €5 million).

The two business cases will be described together as they share many common elements. Where necessary, the characteristics of each will be stated and analysed individually to appropriately highlight any aspects specific to one.

3.1.1 The Type of Data Analysed and How They Are Managed

This advanced version of the model is developed starting from the typical base of PD models for corporate borrowers: first, the internal behavior (i.e. the behavior of the borrowers with the bank), external behavior (i.e. the behavior of the same borrowers with other banks collected through the information made available by credit bureaus, public or private), financial statements and socio-demographic modules are developed using logistic regression to find the relationship between the explanatory variables and the target variable, which is the borrower's default within one year from the observation date.

This information is then supplemented with the modules developed using alternative data, such as:

- internal current account transactional data
- external current account transactional data, available under PSD2
- POS transactional data
- payment cards
- social network data acquired from the user's digital footprint.

The input data in Model 2 are transactional data, i.e., internal data pertaining to transactions involving the current accounts held with the bank. Specifically, the following data are gathered and used to construct the indicators: transaction date, transaction amount, sign (inflow/outflow) and transaction type.

Certain transactions are automatically categorised so that the purpose of the payment can be easily read, whereas others need to be processed in order to identify the purpose (i.e., typically the stated reason for bank transfers needs to be used in order to determine the purpose of the transaction).

The information is divided into two types:

- *structured information*: these are the transaction codes that can be used to determine the purpose of the payment (e.g., "electronic tax payment")
- *unstructured information*: these are the stated reasons for bank transfers (e.g., "March rent") which must be processed using NLP techniques before they can be used.

For additional details on NLP techniques, reference should be made to Sect. 2.2 (“Stand-alone AI Models”).

After the raw data have been processed, each transaction is categorised according to the purpose of the payment (both for structured information and unstructured information). The categories usually include, but are not limited to: Business, Taxes, Utilities, Salaries and Rent expenses. Indicators are constructed based on these categories using more or less advanced mathematical functions.

The indicators are then compared and selected based on their performance and considering their intercorrelation. The final list of selected indicators is then implemented in the ML model.

In Model 1, in addition to the internal transactional data, the following information was analysed in separate ML modules: daily current account transactions (e.g., volatility, liquidity and growth), sales flows and the volatility/growth of revenues generated on the POS channel, the customer’s transactions and the nature of expenses paid for using debit and credit cards (e.g., cash withdrawals) and the social network data of SME Retail customers in the hotel sector or, more generally, in the tourism industry, used to determine their web sentiment, i.e., the user’s satisfaction and the company’s reputation.

As for AI and ML techniques, in the two businesses cases, the algorithms that were selected to implement the ML component in the PD models were: decision tree, random forest and gradient boosting. Clearly, these different algorithms are tested one at a time in order to identify the one that leads to the most robust results. These innovative techniques are always compared with the traditional estimation technique that is based on logistic regression and used as an internal benchmark.

More complex algorithms like deep learning algorithms (i.e. neural networks) were not chosen because the intention was to apply these new ML techniques in the banking industry gradually. Deep learning algorithms have been considered too complex and difficult to interpret, and, therefore, less suited to encouraging discussion with the Regulator and business units. Whether to increase the complexity of the modelling techniques used will be evaluated in the future.

For a description of the characteristics of the various ML algorithms, reference should be made to Sect. 2.2 “Stand-alone AI Models”.

3.1.2 *The Interpretability of Results: An Important Factor*

Indeed, the results of ML models must be interpretable. The interpretability of the models used to calculate capital requirements is a legal obligation (Par. 179 of CRR—Reg EU 575/2013):

> "1. In quantifying the risk parameters to be associated with rating grades or pools, institutions shall apply the following requirements:
>
> > (a) an institution's own estimates of the risk parameters PD, LGD, conversion factor and EL shall incorporate all relevant data, information and methods. The estimates shall be derived using both historical experience and empirical evidence, and not based purely on judgemental considerations. The estimates shall be plausible and intuitive and shall be based on the material drivers of the respective risk parameters. The less data an institution has, the more conservative it shall be in its estimation".

A detailed analysis of the tools that may be used to interpret the results of ML algorithms is laid out in Sect. 4.2 "Interpretability and Stability of the Models' Outcomes". Below is a description of the outcomes that were used in these two business cases to manage the global interpretability aspects (relating to the explanation of the relationships between the initial inputs and the target variable overall) and the local interpretability (relating to the explanation of the relationships between the individual clusters and the individual indicators):

- *traditional benchmark model;*
- *partial dependence plot (PDP);*
- *individual conditional expectation (ICE);*
- *local interpretable model-agnostic explanation (LIME);*
- *Shapley additive explanation (SHAP).*

For illustrative purposes, the methodology considered most effective in interpreting the results of PD models, SHAP, is shown here. It assigns a marginal contribution to each variable of the model considering the possible interactions with the other variables. The PD variation is observed for each combination and, based on the PD variation, the relative weight of the variable is calculated. The SHAP summary plot in Chart 3.1 shows the range and impact of each variable in order of importance (feature value).

Looking at the empirical results, in the case of Model 1, the implementation of alternative data and techniques has led to a noticeable improvement in performance: the introduction of alternative triggered a 5% increase in the accuracy ratio of the final result. Performance is further improved by another 5% points when the alternative data are modelled with ML techniques.

In Model 2, the transactional module's performance is excellent on the out-of-time set (75% accuracy ratio) and contributes significantly to the performance of the overall model.

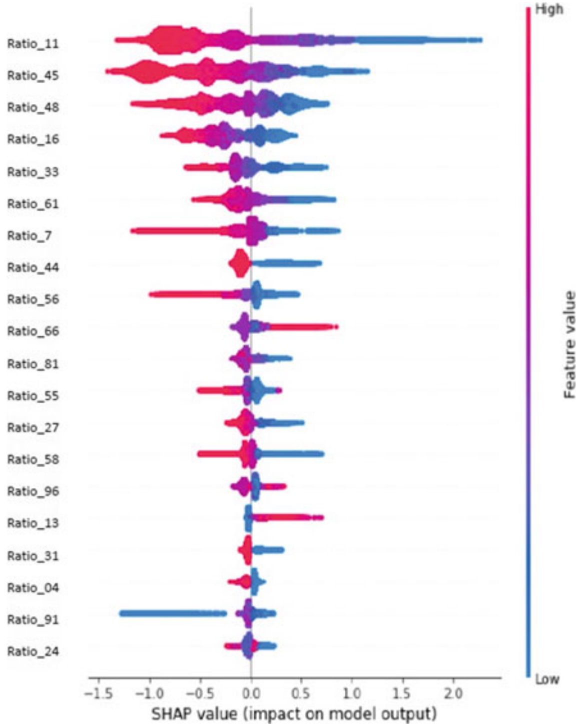


Chart 3.1 SHAP summary plot (*Source* own processing)

However, the transactional module has proved to be highly correlated (80%) with the traditional internal behavioural module, which processes information associated with the relationship of a borrower that is already a bank customer, analysing their credit behaviour (drawdowns, regularity of payments) and processing the information on the number of current account transactions and their amount. Even if the indicators used in the performance module to analyse the transactions are constructed out of structured information (e.g., number of payments and their amount) while those in the transactional module are based on unstructured information processed using NLP techniques, it is obvious that the type of information being used is the same and that the information that can be gathered is similar. This is why the transactional module was not included in the PD model of existing customers.

However, the transactional module has been used for approving credit, i.e., for processing borrowers for which no internal behavioural information is available, and this use offers significant added value. Indeed, combining transactional data with ML techniques has led to the development of a high-performance model that can rapidly capture phenomena and signals that traditional models find difficult to intercept.

A few examples of practical applications of innovative models are given below.

3.1.3 A Practical Case: Risk Discrimination for Borrowers with Seasonal Businesses

Once the transactional module was implemented, it was possible to discriminate borrowers' riskiness by looking at the trend in the payment of salaries. In particular, the analysis differentiated companies with seasonal businesses from those that operate all year round. The Figures below (Charts 3.2, 3.3, and 3.4) show the trend in payment of salaries on the left and the variable's contribution to the calculation of the transactional score on the right, where the variable is constructed based on the Fourier series of transactions relating to the payment of salaries.

The three cases analysed refer to three different companies: the first company presents irregular payments, the second regular payments and the third seasonal payments. In the first case (Chart 3.2), the model has signalled a risk related to the irregular payment of salaries. Payment of salaries is the third largest contributor to the final score for this customer

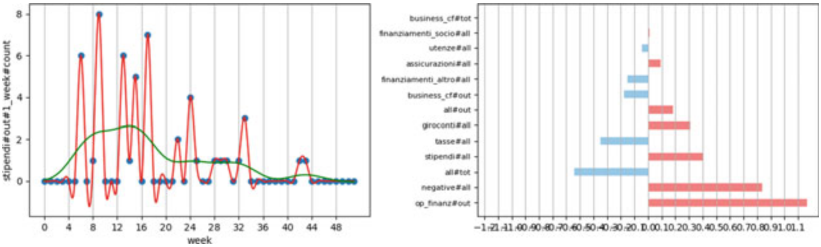


Chart 3.2 Customer with irregular payments of salaries (*Source* own processing)

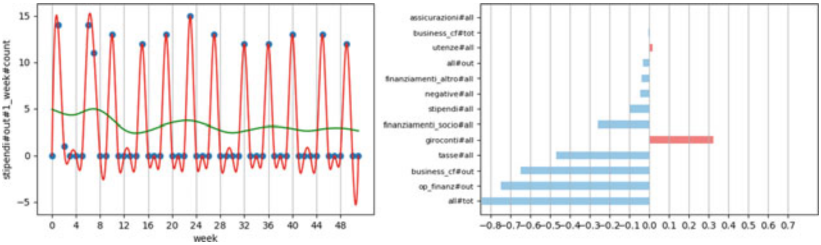


Chart 3.3 Customer with regular payments of salaries (*Source* own processing)

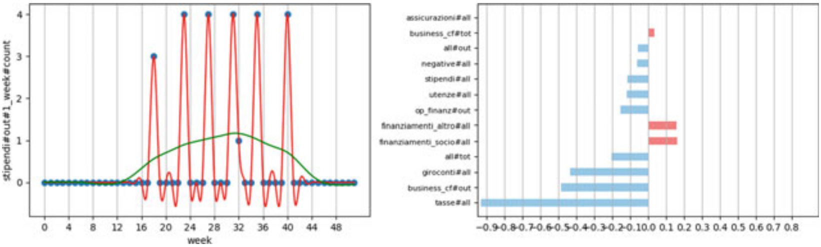


Chart 3.4 Customer with seasonal payments of salaries (*Source* own processing)

and its contribution to the final score is negative. The model has assigned the customer a negative score, and the customer defaulted within 12 months of the valuation date.

In the second case, shown in Chart 3.2, the time set of payments of salaries clearly shows a series of regular transactions. The model has positively scored the regularity of payments of salaries, which has, together with other factors, contributed to keeping the company's final score very low.

In the third case (Chart 3.3), the time set of payments of salaries clearly shows regular transactions only from May to October. Although the business is clearly seasonal, the model can positively score the regularity of payments of salaries without penalising the customer due to its seasonal trend, an error traditional models would have easily incurred.

3.1.4 A Practical Case: Identification of Counterparty Risk During the COVID-19 Crisis

The transactional score proved to be extremely reactive in identifying changes in the riskiness of borrowers during the COVID-19 crisis. Indeed, the transactional score remained stable in 2019, then rapidly grew after February 2020. When the restrictions were eased (June 2020), the transactional score began falling, indicating a partial return to the business standard.

Furthermore, the transactional score makes it possible to discriminate between customers that requested moratoria (for which risk rose during the first lockdown) from others (that demonstrated more resilience with a corresponding improvement in their score starting in May). Indeed, the median risk of customers with and without moratoria diverges at the time of the lockdown. Once the lockdown ended, the curves show a reduction in risk, although they in any case remain spaced apart (Chart 3.5).

3.1.5 A Practical Case: Early Warnings

The transactional score has proved to be potentially usable for early warnings as well. It is possible to start with, for instance, the observation of customers that defaulted in 2019 to examine their performance in the three previous years.

The evidence shows that the average score for customers that defaulted within a 12-month time horizon is also systematically higher in the years preceding the default, demonstrating that the transactional information has a long prediction horizon. Furthermore, the transactional score increases as it nears the default event.

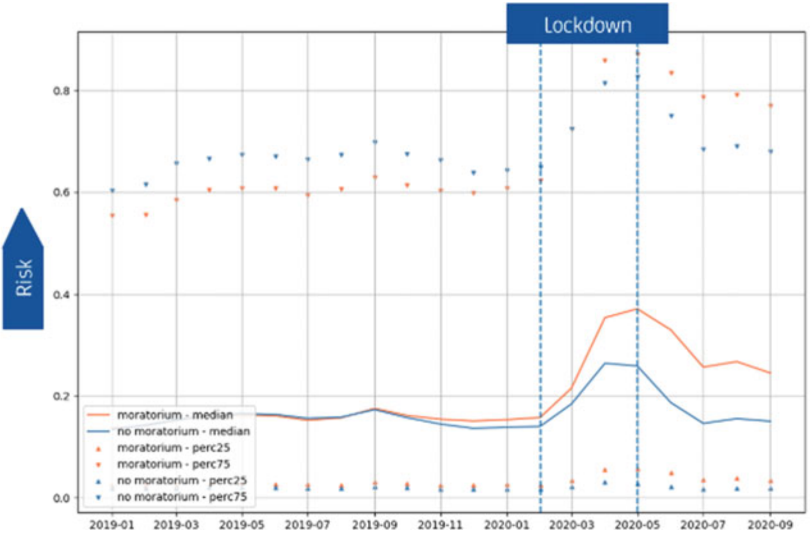


Chart 3.5 Performance of the transactional score during COVID-19 (*Source* own processing)

The model can also pick up signs of credit quality deterioration that had not occurred before and intercept these signs with an increase in the predicted risk. What is reported is, therefore, a form of back-testing for the transactional score and affirms that the transactional score could potentially be used even earlier for early warnings (Chart 3.6).

3.1.6 A Practical Case: Advanced Analytics in Loan Approval[1]

Loan approval is a crucial process in banking and, consequently, is at the centre of bank managers' and the regulators' focus. The banking regulator's interest in these issues was recently proved when the European Banking Authority decided to issue Guidelines on Loan Origination and Monitoring.

[1] This paragraph was written by Working Group 4.2.

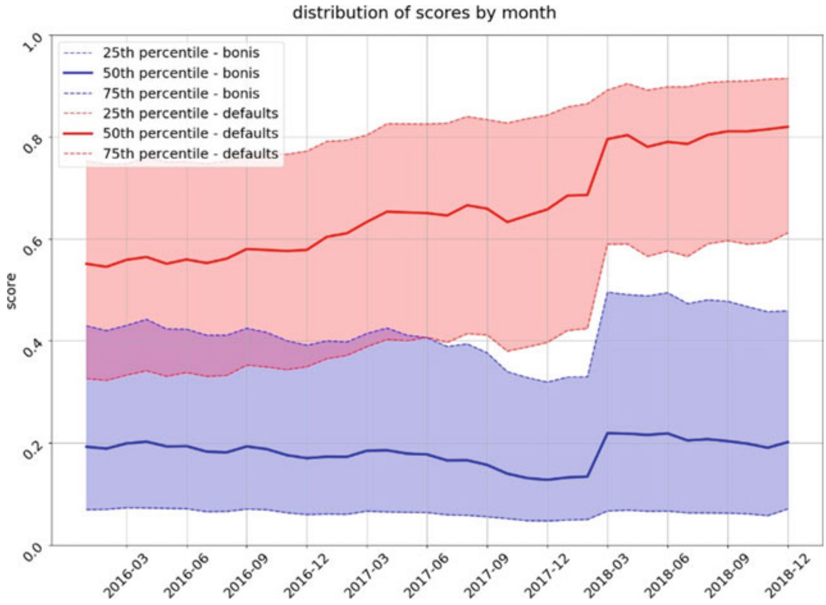


Chart 3.6 Transactional score trend (*Source* own processing)

These guidelines explicitly refer to the possibility of using advanced ML-based creditworthiness assessment techniques, reflecting the advancements currently underway in the financial sector for more accurate credit approval and monitoring.

In many cases, ML techniques are implemented in traditional credit processes to optimise them. However, with this implementation comes the dilemma of the trade-off between improving the model's performance and the traceability of the reasons for their predictions. In the case studies described below, the assessment techniques of the random forest ML model are presented and compared with a traditional model in order to be able to verify the interpretability of the ML techniques applied to creditworthiness assessment processes.

This case study deals with the addition of an ML component, specifically using a random forest approach, to a pre-existing PD estimation model based on logit/probit methodologies. We are most concerned with the aspects relating to the assessment of the gain in terms of accuracy and with the interpretability of the outputs.

The following hypotheses were developed:

- target variable: default within 12 months of origination*;*
- ML application using the random forest model constructed on a training set with an assessment of the key risk indicators in terms of the importance of the explanatory variables;
- validation of the set used to assess the model's performance with different cut-offs (using the ROC curve and the confusion matrix);
- construction of different classification models using various random forest parameters.

After the estimation process, to support the interpretability analyses, the contributions were extrapolated, both at overall level and at the level of the individual risk drivers produced by the random forest methodology.

Specifically, two different analyses were carried out: the first related to the importance plot, i.e., a diagram comparing the contributions of the variables implemented in the random forest model with the contributions of the same variables in the probit/logit model. The second analysis provides the overall performance based on the area under the RoC curve (AUROC).

The importance plot shown below lists the variables from most significant to least significant.

To identify the most significant variables, a specific test is illustrated in the mean decrease accuracy chart, which expresses how much accuracy the model losses when each variable is excluded (Chart 3.7).

Graphically, the superior variables contribute more the model than the inferior variables and have more predictive power in the classification of customers.

Another way to compare the various models applied uses the ROC curve (where ROC stands for receiver operating characteristic), described below. This methodology helps compare the performance of different models with multiple cut-off values, indicating the relationship between the true positive rate (TPR) and the false positive rate (FPR).

The chart shows the performance of the random forest (black curve), the logistic distribution of the same sample compared to the RF (red curve) and the PD calculated by the rating engine without the new indicators in the RF (blue curve) (Chart 3.8).

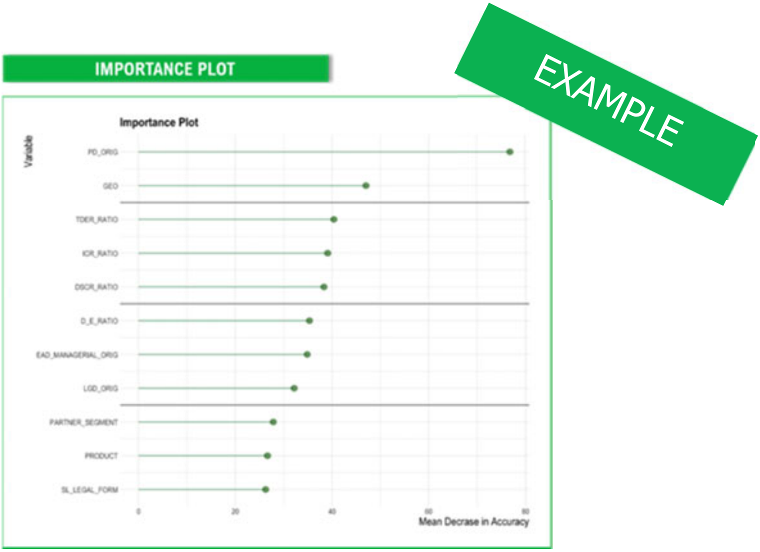


Chart 3.7 Importance plot (*Source* own processing)

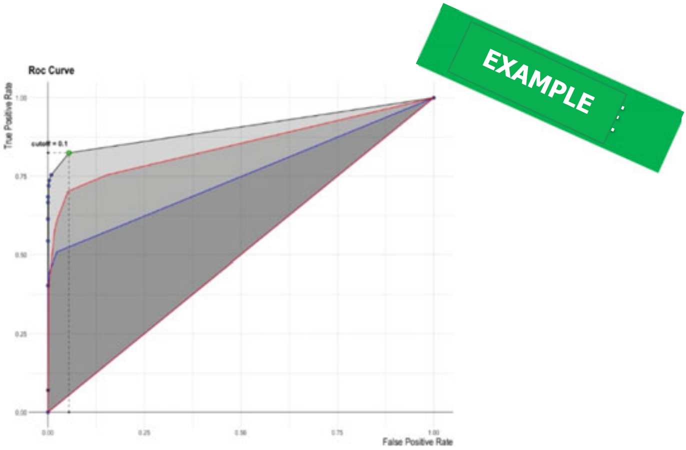


Chart 3.8 ROC curve (*Source* own processing)

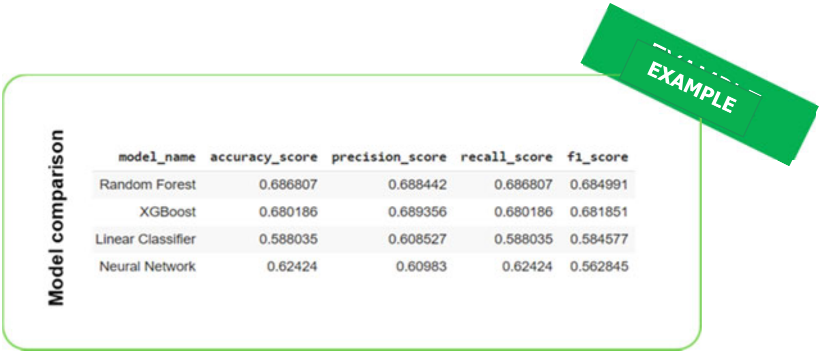
The analyses show that the random forest methodology leads to more predictive results than the logit model, despite possible problems in terms of the interpretability of the model. These problems should be adequately considered, especially if the model is used in credit approval processes.

The logit model accurately determines the estimation coefficients for each variable included in the model, from which the correlation between the variables and the target can be inferred. With the random forest model, it is possible to identify which variables are most predictive of the modelled event, thanks to the importance plot, but without revealing the link with the target variable.

This interpretability weakness could be overcome by adding a correlation analysis (either a diagram or a matrix) to the results of the random forest, which would provide a more complete interpretation of the variables used in the random forest model.

The case study also shows how, in the comparative interpretation of the models, accuracy is a preferred and more appropriate criterion when selecting the model with the best performance, since it provides the ratio of correct predictions to total predictions, even compared to other possible ML techniques (see Chart 3.9 as an example).

The case study described above showed how, using ML techniques, the estimation methodology can be supplemented to include additional indicators, increasing the model's predictive capacity.


EXAMPLE
Model comparison

model_name	accuracy_score	precision_score	recall_score	f1_score
Random Forest	0.686807	0.688442	0.686807	0.684991
XGBoost	0.680186	0.689356	0.680186	0.681851
Linear Classifier	0.588035	0.608527	0.588035	0.584577
Neural Network	0.62424	0.60983	0.62424	0.562845

Chart 3.9 Performance comparison of ML techniques (*Source* own processing)

3.2 How to Improve Traditional Models Using AI Techniques

Stand-alone AI models may be used to supplement more traditional modelling techniques (e.g., logistic regression or linear models).

The first example of how a model may be integrated consists of using traditional models to select[2] the relevant drivers, which may then be reprocessed and combined in the final model using advanced algorithms. In this way, the traditional models are the first "filter" that excludes highly correlated variables or variables that are not meaningful from a statistics or credit perspective, whereas alternative methods are used to develop the model.

Another way of using two techniques jointly could be the creation of an actual function integrating the output of a traditional model with the output of one or more ML models. With this approach, it is assumed that various algorithms can more accurately model the various aspects or sub-areas of the phenomenon in question and that their interaction might bring out the "best" of the various approaches.

A third possibility consists of integrating traditional and innovative models with an expert, rather than statistical, approach. For example, this could happen using a notching matrix, adjustments or overrides of the outputs of the traditional model, particular in those sub-areas where the predictive power is weakest (and, therefore, would most benefit from the integration of ML algorithms). At the same time, there may be instances in which traditional models and alternative models lead to convergent predictions and other instances in which the predictions are divergent or inconclusive, enabling the analyst to focus the expert valuation on the alternative models.

A fourth methodology, which could be considered a variation of the second, entails using AI models to correct prediction errors committed with traditional models. This approach is detailed in the business case presented in the next section.

Integrating traditional models with alternative AI-based methodologies in one of the ways described above may offer a series of advantages over using a stand-alone AI model.

[2] See also Lu Han, Liyan Han, Hongwei Zhao (2013) "Credit Scoring Model Hybridizing Artificial Intelligence With Logistic Regression". *Journal of Networks*, 8, 253–261.

First, this choice may be a gradual transition towards adopting more advanced methodologies, combining modelling techniques that are now consolidated in market practice with more innovative solutions, thereby preserving the interpretability of the model's outputs, especially for internal users of rating models like loan managers and credit analysts.

The integration approach might also be the easiest solution in the medium term for banks that use internal models not only for management reporting purposes but also for supervisory reporting since, in this case, the models must be formally approved by the regulator.

To this end, combining traditional models with AI models can guarantee more comparable results or make it possible to quantify and compare the extra performance of the AI modules with the model's overall discriminative capacity.

Similarly, this approach leaves a certain degree of flexibility in its design, in that it does not preclude the possibility of "dismantling" the AI component to apply, if necessary for the individual use cases, only traditional models.

However, the decision to integrate these types of models may give rise to unknown variables. Certainly, this choice implies the need to estimate several models (both traditional and alternative) and entails the identification of the most appropriate integration methodologies, requiring inevitable extra efforts for risk management units, not to mention the potential complexities of the deployment phase when they are taken from the test environment and implemented in the bank's legacy systems.

Finally, the decision to apply an integration approach may, in certain cases, reflect a compromise that, for extremely non-linear phenomena that need to be modelled, could lead to a dull performance compared to a purely AI-based model.

3.2.1 A Practical Application: The Two-Step Approach

This paragraph presents a possible solution for the integration of traditional and AI models, which an Italian bank used for its PD model for large corporate borrowers.

In particular, it entails the development of a model in two sequential steps. In step one, the model is developed taking a traditional approach and in step two, AI-based approaches are implemented in the traditional model.

The objective is to improve the model's overall performance while maintaining a methodological foundation with a level of interpretability typical of a traditional model.

In the business case considered, the two-step approach is applied to the quantitative component of the Large Corporate PD model which uses the shadow rating technique, where the aim is to replicate the credit rating determined by an ECAI (external credit assessment institution) rather than estimate the borrower's actual PD directly.

3.2.2 The Estimation Methodology Adopted

The first step in the development of the model is based on a practice frequently used in shadow rating approaches in order to rate large corporate borrowers. It entails associating with each observation the rating assigned by the ECAI as the target variable. Next, following the typical phases of the modelling, a long list of drivers is compiled based on the aggregates, financial statements ratios and attributes of the borrower, to obtain a shortlist of drivers and, consequently, to identify a relationship (e.g., through an OLS regression or other similar linear method), which maximises the alignment between the target variable and the score resulting from this relationship, using the selected drivers as the inputs).

Step two is carried out to improve the module's performance. An ML algorithm is applied to the linear model's errors, using the same drivers identified in step one[3] as the inputs. The output of the algorithm is added to the initial score. This technique improves the module's overall performing, minimising the disadvantages created by the complexity and interpretability of the complete model. The training set consists of the annual financial statements data of thousands of borrowers over a period of more than ten years, each borrower's general information, information on its region and industry and the complete history of its ratings assigned by the ECAI over the same time horizon.

The two-step approach is described in more detail further on.

[3] The approach of limiting the analysis to the identified drivers in the first step is not irrevocable. In this specific case, the choice was dictated by the express intention of maintaining the same drivers previously agreed with the analysts responsible for assigning the rating, and the interpretability within the traditional model had already been analysed. However, the boundary of the analysed drivers can always be broadened in step two.

Given the observation *i—th* in the training set and $X_{1,} - X_{N,}$ the value of N drivers in the shortlist for the observation *i—th*:

- an OLS regression or similar linear method is applied $LM(X)$ to the training set in order to calculate the model's score and obtain the error (difference from the target variable):

$$\varepsilon_{\mathrm{LM},i} = \mathrm{Target}_i - \mathrm{LM}\left(X_{1,i}, X_{2,i}, \ldots, X_{N,i}\right)$$

- the ML algorithm is calibrated. For example, if a random forest is used (RF(X)) to predict the error compared to the linear model, using the same training set but only considering the drivers in the shortlist:

$$\varepsilon_{\mathrm{LM},i} = \mathrm{RF}\left(X_{1,i}, X_{2,i}, \ldots, X_{N,i}\right) + \varepsilon_{\mathrm{RF},i}$$

- the final score is calculated as the sum of the score deriving from the linear model and the score resulting from the random forest algorithm:

$$\mathrm{Score}_i = \mathrm{LM}\left(X_{1,i}, X_{2,i} \ldots, X_{N,i}\right) + \mathrm{RF}\left(X_{1,i}, X_{2,i}, \ldots, X_{N,i}\right)$$

Common techniques are used to define the training and test sets and to calibrate the hyperparameters of the algorithm.

The results obtained with the two-step approach are analysed in two main ways: prediction stability and performance improvement. The stability of the prediction is evaluated using the usual metrics, i.e., checking that the model's performance remains stable by applying the estimated parameters to a set that is independent of the training set. The final model's performance in terms of its accuracy ratio (AR) is compared with its performance when it was exclusively linear (step one, using the traditional approach), to ascertain whether the addition of the ML score leads to a meaningful improvement in the model. The table below shows the ARs of the two-step model considering training, test and total sets, in order to compare those obtained with the ones resulting from the linear model only (Table 3.1). It shows that the two-step model leads to a roughly 11% improvement in performance.

Another area to explore is the interpretability of the final model and, in particular, the entity and method by which each model driver affects the final score, both individually and collectively with the others. Various methods of interpreting AI models are described in Sect. 4.2 "Interpretability and Stability of the Outcomes of the Models".

Table 3.1 Performance of the linear model compared to the random forest model

		Accuracy (%)
Linear model	Total set	51.6
RF Model	Total set	62.7
	Training set	63.0
	Test set	62.1

One of the most frequently used techniques to this end consists of partial dependence plots (PDPs), which give us a picture of the marginal effect of one or more drivers with respect to the predicted target variable. The underlying idea is to show the relationship between one of the drivers and the target variable, all other drivers remaining constant at their average value.

For instance, in this specific business case, it is interesting to see whether there is a relationship between one or more drivers and the score integrated with RF (X), which deviates significantly from the linear model. For example, Chart 3.10 shows the difference between the linear relationship and the RF relationship for one of the model's drivers: the red line represents the relationship for the classic, purely linear model, whereas the blue line shows the behaviour with the addition of the RF score.

The chart illustrates how the relationship of the two-step approach is essentially a disruption of the linear relationship. In particular, for the driver's extreme values, the score resulting from the two-step model is much lower than the score calculated using the linear model, while there is a slight upward adjustment for the values in the middle. Reference should be made to the specific section of this paper for a more detailed discussion of the interpretability of results.

3.3 Applying an AI Model to Asset Management

The business cases described thus far all share a common banking theme: they refer to credit scoring systems used to assess customers with different characteristics (micro-businesses, SMEs and large corporate customers) but nevertheless in the scope of banking relationships. Although the borrowers' characteristics significantly affect the nature and frequency of the updating of the data used in the assessment systems, because there is a banking relationship in place, "private" information can be acquired, and

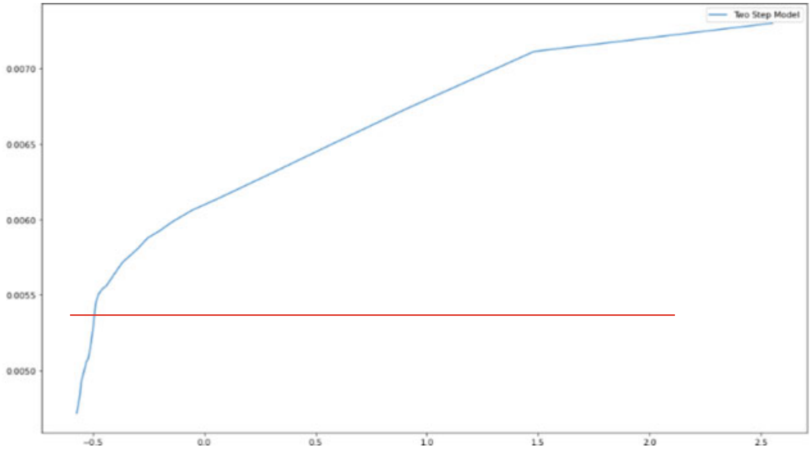


Chart 3.10 Relationships between risk drivers and the scores of the linear and random forest models (*Source* own processing)

it is on this "private" information that the AI models are based—such as, for example, the models that use transactional data—or that can be used to correct the models' outputs, as in the two-step models used to assess large corporate borrowers.

However, the AI models are also well suited for use in contexts where the lack of a previous banking relationship means there is no private information or that such information takes a different form from that described up until now, consisting of, for instance, the findings of a financial or commercial due diligence review or the result of an interview, even as an audio recording.

The applications of AI models include supporting an asset manager's investment decisions in the direct management of credit funds or in the management of the proprietary books of banks, insurance companies and pension funds. The assets under management are often very broad and consist of both the financial instruments typical of private markets, like corporate loans, SME loans, acquisition finance, private placements and club deals, and securities on liquid markets, like high-yield bonds, floating-rate notes and syndicated loans.

Focusing on semi-liquid or illiquid sub-investment grade products, the analysis process using IA models entails a series of similar steps to those described in the previous paragraphs, at least in the initial phase for the structuring of unstructured data.

In particular, the process begins with the acquisition of unstructured data (business plans, financial statements, commercial due diligence, financial due diligence, voice calls, etc.) and continues with the structuring of these data, feeding the calculation engines and then obtaining the output of the model. The output may be the quantification of the relevant risk parameters for an investment decision (such as probability of default, expected loss and the risk-adjusted spread) or for a non-binding opinion on an investment decision, which by its very nature must be binary: either invest or do not invest. In this context, AI systems may also be considered critical support decision systems.

The process described up to this point may then be completed with modules to present the results in the form of (i) verbal commentary on the system output, (ii) a report arguing for or against the valuation/decision or (iii) dialogue with a physical or digital android.[4] These methods of presenting the output may be used separately or together (see Chart 3.11).[5]

In the experience of one of the authors, deep neural networks are a particularly versatile tool to structure models of this type for a number of different reasons. First and foremost, neural networks have an extremely strong ability to learn and adapt to the data, in the sense that they can grasp the relationships between the data and the target variable even when such relationships are extremely non-linear. Second, neural networks can be organised to recognise qualitative elements within unstructured data and transform them into structured information. Additionally, they lend well to combining typically quantitative elements (i.e., signal intelligence [SigInt]) with qualitative elements (i.e., human intelligence [HumInt]). Lastly, they can be structured so as to request additional information when the system does not consider the feeding process sufficiently robust.

[4] In this respect, for example, the following physical and digital androids may be used: Robot Sophia by Hanson Robotics and Samsung's Artificial Humans by Neon.

[5] A one-minute video of an early non-commercial application incorporated in an Avatar Robot (Edgard, Singapore Nayang Technological University) is available at the following link: https://drive.google.com/file/d/1dJjtNEBA96w0vDo0lZykld1381rAaeXB/view?usp=drivesdk.

The name of the investee has been hidden for compliance reasons and the fund manager simply calls the Avatar Robot Mr. Neural Network.

Module 1	Module 2	Module 3	Module 4	Module 5
Unstructured Data: Business plan Financials Behavioural signals Mgmt team voice call Business DD Commercial DD etc.	Structuring Process: Transformation of the signals into organised data Check that the data are complete and consistent Any requests to complete the information	Calculation Engines: One or more deep neural networks fed by data organised when the data come in	Raw Output of the System: PD (and related further parameters) Binary investment opinion	Relational Output of the System: Verbal and Written (Limited) Argumentative capacity of decisions

Chart 3.11 The process (*Source* own processing)

The contribution of a model of this type is measured by comparing the expected losses of a portfolio with a similar composition and risk as with those of a portfolio selected and managed with the deep learning system and whose composition and risk is similar to the former,[6] as indicated in the three formulas below:

$$\begin{aligned}\text{Human Experience Differential} = {} & (\text{Expected Loss} - \text{Actual Loss}) \\ & \times \text{High Yield Portfolio Value} \\ & \times \text{Average Duration of the Portfolio}\end{aligned} \tag{3.1}$$

where:

Expected Loss is that measured on the basis of reliable external or internal rating systems

Actual Loss' is the value of credit losses that effectively arise in the portfolio.

A differential with a positive value reflects the greater value created by fund managers' experience over that of reliable internal or external rating systems.

$$\begin{array}{c}\text{Human Augmentation} \\ \text{vs} \\ \text{Human Experience Differential}\end{array} = \begin{array}{c}\left(\text{Actual Loss} - \text{Actual Loss}'\right) \times \\ \text{High Yield Portfolio Value} \times \\ \text{Average Duration of the Portfolio}\end{array} \tag{3.2}$$

[6] The most common measurement is made each year on a comparative basis and is the point-in-time EL of a portfolio with the same composition, rating classes and actual credit losses of the portfolio being valued (which is point-in-time by definition). There are other calculation methods that use through-the-cycle data, point-in-time moving averages, etc., and it is clear that the fund managers' human experience cannot be objectively assessed until the managed fund has reached the end of its life. Additionally, there are various indexes that can be used for benchmarking, where appropriate. The most well known are Proskauer Credit Default Index, the Cliffwater Direct Lending Index and the ELLI Index (European Leveraged Loan Index) for LBOs. In practice, in illiquid markets, data extracted from the actual loss tables broken down by rating classes and published by Moody's or Standard&Poor's are used.

where:

Actual Loss is the value of credit losses that effectively arise in the portfolio
Actual Loss' is the value of actual credit losses associated with the AI/DL selection criterion (Augmented Human)
A positive differential reflects the greater value created by the AI/DL system than the fund managers' abilities.

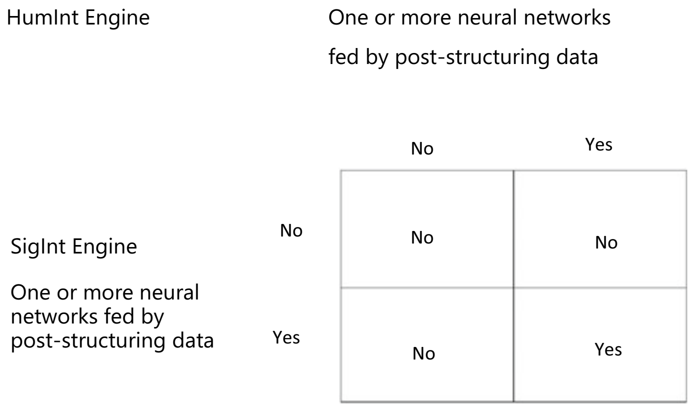
$$\begin{aligned}\text{AI/Deep Learning Differential} = & \left(\text{Expected Loss} - \text{Actual Loss}'\right) \\ & \times \text{High Yield Portfolio Value} \\ & \times \text{Average Duration of the Portfolio}\end{aligned} \tag{3.3}$$

where:

the Expected Loss is that measured on the basis of reliable external or internal rating systems
Actual Loss' is the value of credit losses that effectively arise in the portfolio selected with the AI/DL system.

It is easy to see that the result of equation [3] is the sum of the results of equations [1]+[2].

In the experience of one of the authors, a deep learning system was applied to asset management in the context of a specific business model named parallel lending. In this model, the asset manager collaborated with the bank and shared a portion of its medium/long-term loans under the same contractual terms and conditions as the bank, without competing against the latter for the related revenue arising from the medium/long-term loans. In this way, the bank split up the risk on the medium/long-term loans by involving the asset manager, while increasing the risk-adjusted commercial profitability on the relationship with the corporate customer. From the asset manager's standpoint, parallel lending makes it possible to build portfolios that are highly diversified in terms of sectors and regions and that are, especially, extremely granular (with targets of over 100 investees).



HumInt Engine

One or more neural networks fed by post-structuring data

SigInt Engine

One or more neural networks fed by post-structuring data

	No	Yes
No	No	No
Yes	No	Yes

Chart 3.12 The decision-making process (*Source* own processing)

In this specific experience, multiple deep neural networks work in parallel (e.g., two networks to exploit quantitative elements, SigInt, and one for qualitative signals, HumInt) and the last mile output is achieved by overcoming both levels. As shown in Chart 3.12, the result is positive, i.e. the output is "Yes", only if both neural networks (qualitative and quantitative) generate a positive output.

Deep neural networks are trained to minimise/eliminate actual losses (i.e., type 1 errors).[7]

With respect to the SigInt component, the model was developed on a training set of more 20,000 companies observed over 10 years and with a default definition that is consistent with that for supervisory capital regulations. The out-of-time back-testing analyses on a set of more than 10,000 companies over four years led to around 92% values in the area under the curve,[8] an excellent performance.

[7] Technically, both SigInt networks and HumInt networks are multi-layered, the activation function is a rectified linear unit (ReLU), the learning rate is set at 0.002, the loss function is binary-crossentropy, the beta is between 0.9 and 0.999 and there are about ten epochs.

[8] The relationship between the area under the curve (AUC) and the better known accuracy ratio (AR) is AUC = AR /2 + 50% (see Deutsche Bundesbank, Measuring the Discriminative Power of Rating Systems, no. 1, 2003).

Performance increases considerably when HumInt is added to the equation. The HumInt component of the deep neural network was developed on a smaller training set than that used for the SigInt but has the advantage of incorporating the assessment with all the relevant factors that arose in discussions with the Investment Committee. These factors relate to both the typical aspects of fundamental analyses of the target company (e.g., its market position, solidity and resilience of cash flows, etc.) and the quality of the co-sourcing relationship with the bank in the scope of the parallel investment, in addition to more sensitive aspects like the reputation and track record of the business owner and/or management team.

Given the small number of observations that in this business case constitute the HumInt component (just under 400 observations), although the accuracy ratio is particularly high, it is deemed still too soon to have statistics on the joint effectiveness of the SigInt and HumInt systems that are, overall, solid. However, it is also true that certain portfolios managed using the parallel lending business model in the past three years did not present any payment default events or, more simply, any moratoria on the repayment of principal or interest despite the occurrence of an atypical event like the COVID-19 pandemic.

Based on the in vitro tests (through back-testing) and the hands-on experience gained thus far, minimising risks in a portfolio of high-yield unsecured SME/corporate loans can lead to a reduction in the loss rate from 110 to 120bps[9] to 20–30bps, with a consequent initial guideline estimate of the creation of value in the ballpark of 90bps[10] per annum.

[9] In this case, a product sub-investment grade portfolio is used, 70% of which consists of unsecured assets invested in companies with an external rating or reliable internal rating of BB and 30% of which consists of unsecured assets invested in companies with an external rating or reliable internal rating of B. Based on the Annual Default Studies released by Moody's, average TTC data on actual credit losses by companies rated BB and B are, respectively, 60bps and 240bps (see Annual Default Study: Corporate Defaults and Recovery Rates, 1920–2017, Moody's Investor Services, Exbit 23, Annual Credit Loss Rates by Letter Rating 1983–2017). Conversely, the actual credit loss rate for the strategies currently managed by Muzinich & Co. SGR is zero, despite the impact of the pandemic. However, it is deemed reasonable to indicate a forward-looking range of 20bps–30bps.

[10] A more robust measurement should contemplate either broader historical series of the credit loss rates associated with the strategy or, alternatively, temporary consistency of the point-in-time credit loss rates when comparing the benchmark portfolios with the actual

It is also important to clarify that an investment strategy based on minimising type 1 error may be effectively applied where high type 2 error is tolerable (or does not present material costs), as in the case of the parallel lending business model.

Referring to the equation [3], the formulas would be:

$$\begin{aligned}\text{AI/Deep Learning Differential} = {} & \left(\text{Expected Loss} - \text{Actual Loss}'\right) \\ & \times \text{High Yield Portfolio Value} \\ & \times \text{Average Duration of the Portfolio}\end{aligned} \tag{3.3}$$

where:

the Expected Loss is that measured on the basis of reliable external or internal rating systems = 110bps–120bps

Actual Loss' is the value of credit losses that effectively arise in the portfolio selected based on the application of the AI/DL system. This value has been zero until now. However, a conservative estimate in the range of 20bps–30bps is given.

$$\begin{aligned}\text{AI/Deep Learning Differential} = {} & 0.9\% \times \text{High Yield Portfolio Value} \\ & \times \text{Average Duration of the Portfolio}\end{aligned} \tag{3.3}$$

3.4 Use of AI Models for the Validation/Benchmarking of Traditional Models

In recent years, ML techniques have progressively become the established practice for the development of supervisory and management reporting models. However, these models must be subject to life-cycle management similar to that of "traditional" models and the application of a model

portfolios. If the comparison horizon is limited to 2018–2021, the differential remains around 60–80bps. Finally, the back-testing confirms a differential of 70bps–110bps.

risk management framework (in terms of risk identification, mapping, assessment and consequent mitigation).

In addition to the development of stand-alone models, ML models are also challenging existing models—ML vs traditional and/or ML vs ML—both when used for supervisory reporting and for purely management reporting purposes, as clarified below.

In this respect, the applicable regulations refer to the validation role and the use of challengers (e.g., SR 11-7[11]). Particularly, SR 11-7 requires the validation of all the model's components (inputs, processing and reporting).

At present, there is no "standard" framework for the assessment of the aforesaid models, although practices for the assessment of the model design, performance and interpretation of the results of the challenger are spreading.

In this context, much importance is given to the data, which are taking on ever more significance in the scope of the broader model risk governance framework. Although this is a particular sensitive topic for traditional models as well, in the specific context of ML, data play an even more delicate role, as the quality of the models depends even more heavily on the quality of the data.

The purpose of this paragraph is to describe how AI techniques may be used as challengers in the validation of credit risk models.

A few business cases entailing the use of AI techniques for validation purposes are described below. The first two describe cases in which AI is used for supervisory reporting purposes and the last three for management purposes.

3.4.1 ML Techniques for Benchmarking Capital Requirements Models

In the measurement and monitoring of model risk, the application of ML techniques has been applied as a benchmark for models developed using traditional techniques.

In the first case, tests are performed for various groups of exposures to identify any obvious bias in the corporate rating model. The analysis was conducted by comparing the estimates resulting from traditional

[11] Supervisory Guidance on Model Risk Management established by the Federal Reserve Bank of the United States of America (4 April 2011).

methods (logistic regression) and alternative methods (decision trees, random forests, neural networks), leaving unchanged both the training set data (including the time horizon) and the explanatory variables (calibration and sensitivity analyses of the model in the two approaches). The results were assessed using established materiality thresholds and no significant differences arose, so the model risk was considered negligible.

In the second case study analysed (PD Retail model), the internal validation unit developed an alternative model using ML techniques to compare the performance obtained using traditional models.

In terms of the model design, the alternative model uses a different method of analysing the information to calculate the indicators, following, for instance, different guidelines:

- to estimate the link with the economic cycle;
- to treat the missing values and the extreme values;
- to include new or different hyperparameters, in particularly changing the number of estimators (trees) and the minimum weight of each of them;
- to change the number of variables to include in the alternative modules, with the preparation of more objective criteria;
- to calculate automatic alternative training algorithms for comparison with the performance of the original gradient boosting model.

The alternative model proposed resulting from the integration of the challenger modules was compared with the main model on both the in-sample set and the out-of-sample/out-of-time set. The performance, which was only slightly superior in all fields, demonstrated the accuracy of the alternative model, which was in any case validated. The intention was to use the alternative model as a "shadow model" for the periodic benchmarking of the main model's results, transforming it into an instrument that makes it empirically possible to perform new series of tests on the application of the model and, concurrently, more stringent safeguards on the application of the new technology.

3.4.2 Initial Applications for Management Purposes

A first example of the application of scoring models developed with ML techniques is the use of these models to challenge expert credit scoring.

One example in this respect is the development of a scoring model that expresses the customer's riskiness in the short-term (e.g., six months). The model was developed to provide a risk assessment of borrowers during COVID-19 and is distinctly point-in-time. This entailed the use of a long list of variables fed by multiple sources, such as financial statements, transactional data, credit bureau data, industry data, etc. The model was used to challenge the specific assessments provided by the network to evaluate borrowers' resilience following the impact of the pandemic.

The comparison with the network's assessments took the shape of a traffic light (green, yellow, orange and red) to represent the level of credit risk in the short term. The ML model's "traffic lights" were generated by the qualitative and quantitative aggregation of various risk buckets reflecting the growing probability that creditworthiness would worsen in the following six months. In general, the comparison did not give rise to any significant discrepancies with respect to the assessments made by the network. However, the need arose to conduct deeper analyses (i.e., single name controls) to investigate the reasons for the discrepancies that were found.

Another example of an alternative model used for management purposes is the development of one used to benchmark the internal rating model of the retail segment—estimated using a traditional logistic regression algorithm—for the purpose of recognising credit losses on a collective basis in accordance with IFRS 9.

The regulations applicable to these activities consists of the 26th update to Bank of Italy Circular no. 285 of 6 March 2019 and the "Guidelines on credit risk management practices and accounting for expected credit losses", published by the European Banking Authority (EBA) on 20 September 2017.

The fifth principle of the EBA's guidelines is devoted to "ECL model validation". This principle states that "Credit institutions should have robust policies and procedures in place to appropriately validate the accuracy and consistency of the models used to assess the credit risk and measure ECL, including their model-based credit risk rating systems and processes and the estimation of all relevant risk components, at the outset of model usage and on an ongoing basis".

Specifically, the benchmark is an artificial neural network model (of the deep learning variety), i.e., a statistical model consisting of elementary computational units (neurons) linked by weighted connections. These units are layered, so that each neuron in a layer is solely connected to the neurons in the previous layer and the subsequent layer. Statistical neural network models are based on natural neural networks, which are comprised of a vast number of nerve cells (about ten billion in a human being), called neurons, which are interconnected in a complex network. Intelligence is the result of many interactions between interconnected units.[12]

The input of a neuron depends on signals of different signs and intensities coming from the connected neurons. When a certain threshold is reached, through the appropriate transfer function, the neuron generates a signal that propagates, through the weights, to other neurons.

Before they are used to train the neural network, the model's inputs are pre-processed in a phase called one-hot encoding, which normalises the categorical variables.

To estimate the neural network benchmark, the following were used:

- a feedforward architecture (developed using Python algorithms) whereby the connections between the layers may only feed forward; therefore, the signal is transmitted only to the neurons in the next layer;
- two activation functions: "ReLU" for the neurons in the hidden layers and "sigmoid" (or logistic function) for the neurons in the output layers, which is particularly useful for classification problems and takes values between 0 and 1;
- supervised training given the presence of the dichotomic target variable (with the value of "1" if it will be classified as non-performing exposures within one year of the observation or "0" if not) which entails the use of a training dataset.

The neural network learns by examining individual records in the training set, generating a prediction for each record and making adjustments to the weights whenever it makes an incorrect prediction compared to the target variable. This process is repeated n times (called epochs) and the network continues to improve its predictions until one or more of the stopping criteria have been met. Initially, all weights are random. Once trained, the neural network is tested, checking how it behaves on a validation set made up of observations not used in the training. Therefore,

[12] See Sect. 2.2.

the purpose of validation is to evaluate the neural network's ability to generalise, i.e., to check whether it is capable of providing answers even for inputs that it has never seen before during training.

A multilayer perceptron (MLP) neural network was used to estimate the benchmark model, configured as follows: first an input layer, then two hidden layers consisting of 30 neurons each, and finally an output layer.

A training algorithm based on the gradient method calculated using back propagation was used to estimate the benchmark model. This algorithm initialises the weights randomly and updates them making small, gradual and progressive changes based on the estimated error gradient between the result produced by the network and the desired result. The training process continues for a finite number of iterations, and this is how the algorithm converges towards the desired solution. To reduce the risk of overfitting, the training set is split randomly into two sets of parameterisable percentages: the first set is used as the actual training set to optimise the network's parameters and the second is the validation set to evaluate the convergence and strength of the network during training.

Chart 3.13 shows the neural network training process by number of iterations/epochs.

Below is a summary of the main features of the estimated benchmark neural network model:

- feedforward and MLP architecture;
- input layer: 84 variables;
- "ReLU" activation functions for the hidden layers and "sigmoid" functions for the output layer;
- hidden layers: two layers with 30 neurons each;
- output layer: one neuron to estimate the binary target variable.

The performance of the benchmark model, which was developed using more recent data than the internal model, was calculated based on an out-of-sample set and subsequently compared with the internal model at the same date. For the purposes of the comparison, standard performance indicators (Cap Curve and Accuracy Ratio, Roc Curve and AUROC (AUC), Confusion matrix), as shown in Table 3.2 were considered.

The internal retail model presents, for all the calculated indicators, a statistical performance just below the benchmark model (neural network).

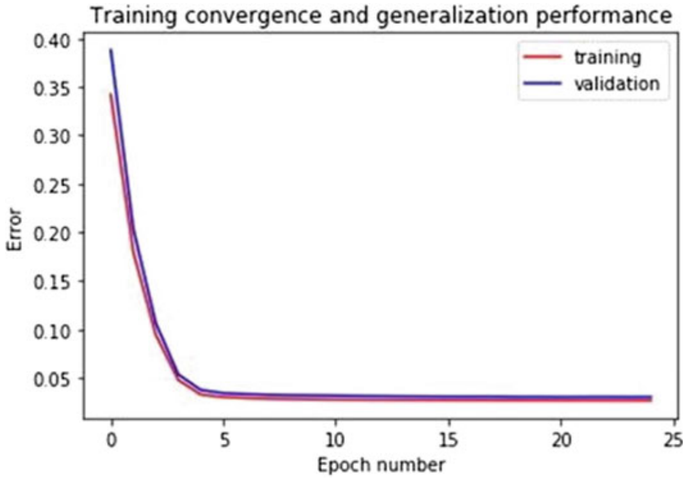


Chart 3.13 Error convergence as the number of iterations increases (*Source* own processing)

Table 3.2 Performance of the internal model compared to the neural network

	Internal model (%)	*Neural network (%)*
Accuracy ratio	83.51	86.55
AUROC	91.75	93.27
Correct classification rate	84.50	87.03
Sensitivity	83.98	86.83
Specificity	85.02	87.24
False positive rate	14.98	12.76
False negative rate	16.02	13.17

Source own processing

The negligible difference between the statistics of the two models confirms the internal model's high predictive power and accuracy and, therefore, that it is suitable for use to assess credit risk and measure expected credit losses in accordance with IFRS 9.

Furthermore, since this rating model is used for management purposes (approving credit products to existing customers or for ad hoc commercial campaigns), the positive results of the benchmark analysed demonstrate the rating's high discriminative power with limited model risk, thereby enabling effective origination policies.

The third business case for management purposes entails the use of an autoencoder neural network to assess the performance of the internal performance model for the retail segment and the benchmark neural network model for outliers.

Reference should be made to Sect. 2.2 ("Stand-alone AI Models") for a technical description of the autoencoder neural network algorithm.

A multilayer autoencoder was used to analyse the outliers, which is an implementation based on three hidden layers with a specular structure for better generalisation. In particular, an undercomplete model was developed in which the hidden layers, measuring 15 with an inner layer measuring 2 (bottleneck structure), were smaller than the input and output layers (84), obtained by normalising the real variables and one-hot encoding the categorical variables of the starting space.

This limitation forces the neural network to learn how to represent compressed data, i.e., to obtain as much information as possible from the input variables with a space of only two hidden variables.

In order to identify the model's outliers, decoding was used to reconstruct the input starting from the space of the two hidden variables. The distribution of errors between the reconstructions made by the autoencoder and the original inputs is cut according to a threshold past which the data are considered outliers.

Below is a summary of the characteristics of the autoencoder developed to analyse the *outliers:*

- multilayer autoencoder;
- input layer: 84 variables;
- hidden layer: three layers, the first with 15 neurons, the second with two neurons and the third with 15 neurons;
- output layer: 84 output variables;
- number of epochs: 20.

The autoencoder model was trained using the same training set that was used to train the benchmark neural network. Next, starting with a set of out-of-sample data, a number of outlier observations equal to roughly 0.1% of the total portfolio were identified. Considering only the perimeter

Table 3.3 Performance of the internal model compared to the neural network

	Internal model (%)	*Neural network*
AR	54	56
AUROC	77	78

Source own processing

of outliers, the internal model and the benchmark neural network were compared with certain performance indicators reported in Table 3.3.

Considering the perimeter of outliers, the internal model's performance was just under that of the benchmark neural network but was still sufficient to confirm the robustness of the model and its ability to identify default even when there are extreme observations.

The use of challenger models changes the way the validation function works, which is no longer limited to checking that the production models function correctly but checks that they are robust by developing alternative models that can be used to compare the result obtained with that of "official" models. Despite the substantial effort needed for the estimate of challenger models, experience has shown that they are particularly useful in model validation.

AI techniques used for validation, testing and benchmarking purposes have now become general operating practice to the same degree as ordinary validation techniques with respect to management models and boast excellent performance even though they are still continuously evolving.

They are, however, less frequently used for supervisory reporting purposes due to the greater regulatory constraints, although in recent years the regulator has asked for them to be used more.

References

Deutsche Bundesbank, Measuring the Discriminative Power of Rating Systems, no. 1, 2003.

Han, L., Han, L., Zhao, H. (2013). "Credit Scoring Model Hybridizing Artificial Intelligence With Logistic Regression". Journal of Networks, 8, 253–261.

Moody's Investor Services, Annual Default Study. Corporate Defaults and Recovery Rates, 1920–2017.

CHAPTER 4

The Validation of AI Techniques

Abstract This chapter describes the implementation of validation techniques aimed at monitoring and mitigate risks related to the development of AI models. The key trustworthy indicators are identified and detailed in coherence with the main trustworthy principles, namely accuracy, robustness, fairness, efficiency and explainability. Also, a focus on the interpretability of the AI models' outcomes, summarising the main regulatory requirements, and describing the methodological approaches aimed at assessing the stability of the models is detailed. In order to evaluate and interpret the results of the AI models, the contribution of each risk divers is assessed by means of specific methodologies.

Keywords Traditional models · Traditional techniques · Machine Learning—ML · Artificial Intelligence—AI · Validation · Trustworthy indicators · Trustworthy principles · Accuracy · Robustness · Fairness · Efficiency · Explainability · Credit risk · Credit risk assessment · Regulatory requirements · Interpretability · Stability · Model outcome · Results

R. Locatelli et al., *Artificial Intelligence and Credit Risk*,
https://doi.org/10.1007/978-3-031-10236-3_4

4.1 Possible Comparison Criteria Between Traditional Models and AI Models[1]

One of the main objectives specified in the regulatory framework on AI proposed on 21 April 2021 by the European Commission focuses on resolving and lessening problems related to the governance of risks deriving from AI applications. To monitor and mitigate such risks, it is necessary to intervene by formalising a series of key trustworthy indicators (KTIs) that meet the measurability and independence requirements with reference to the underlying ML model and that make it possible to comply with specific trustworthy principles like accuracy, robustness, fairness, efficiency and explainability.

It is precisely through the creation of the appropriate KTIs that we can ensure the adequate advancement of the "intelligence" of AI systems, resolving an ever-broader series of complex problems, making it easier to adapt these systems to situations and making AI methodologies more trustworthy in the sense that they will be more precise, stable, inclusive, efficient and explainable.

Based on the new requirements highlighted in the European Commission's proposed regulatory framework on AI, current research will focus increasingly on the operating effectiveness of the KTIs in the measurement of the established key principles.

A summary of the main approaches and algorithms being developed may be useful in providing a complete view of the recent contributions offered in the field of AI. Specifically, the tendency to take an iterative approach is increasingly taking hold which makes it possible, on the one hand, to include additional methods in the analysis and, on the other, to remove methods that prove to be unsuitable despite initially showing promise.

The five key principles for trustworthy artificial intelligence are presented and discussed below.

4.1.1 Principle 1: Accuracy

To evaluate the predictive accuracy of AI methods, new KTIs need to be defined on the basis of a direct comparison with the predictions obtained

[1] This contribution is based on research by the authors Paolo Giudici and Emanuela Raffinetti (Department of Economics and Business Management, University of Pavia), funded by the European Commission as part of Horizon 2020.

using a specific, selected ML model and the observed actual values. Typically, this may be accomplished by using the mean squared error (MSE) of the predictions, calculated in correspondence with a test dataset and generated by an ML model in turn "trained" on a training dataset, or in the area under the ROC curve (AUROC), based on the binary predictions obtained following a similar cross-validation approach.

However, the comparison of the predictive capacity of classes of ML models that differ in terms of their construction and the nature of the outcome variable gives rise to the need to extend the formalisation of "universal", general KTIs with predictive capacity, which can be used irrespective of the outcome variable (binary, continuous or ordinal) and the underlying ML model.

With respect to predictive accuracy, for example, neural network models adequately adapt to the statistical regularity of the training data but do not necessarily learn the generalisations of the application context, leading to low predictive accuracy. In this respect, the literature includes references whereby the neural network models perform well in terms of predictive accuracy for standard data but do drastically worse for data outside the reference distribution. This relates to the fact that the neural network models excel in the generalisation of "superficial" phenomena but not necessarily in the generalisation of causal structures within the data.

As noted earlier, the measures typically used to verify the predictive accuracy of ML models are: AUROC (for binary outcome variables) and MSE (for continuous outcome variables). Another criterion that can be used to meet the "universality" requirement is the rank graduation accuracy (RGA) measure, which is used to measure the predictive capacity of continuous, ordinal and binary target variable models (Giudici and Raffinetti, 2021b), based on a direct comparison of the natural arrangement of the observed values with the arrangement of the same as a result of the ranks associated with the respective predictions.

4.1.2 Principle 2: Robustness

Another aspect that must be adequately considered relates to the introduction of KTIs that can measure the robustness of the AI methods when the data vary (i.e., when there are anomalous or extreme data) and/or the scenarios considered change (i.e., in a scenario analysis). It follows that the concept of robustness is closely tied to the concept of limited change in the predictions generated (stability).

When it comes to robustness, the neural network models perform well in the processing of inputs that are similar to the training data, but their ability to detect any "similarities" in the data may be substantially weaker than that of an expert in the specific industry. For instance, in an artificial neural network, visual recognition of human movement and gestures may be unsuccessful due to poses, obstacles and types of movement that vary significantly from the "training" poses and base scenarios. This means that even the smallest change in the image could lead to obvious errors in visual recognition. One problem in this area relates to the trustworthiness of AI methods in situations that could be rare but dangerous. An example could be a self-driving car that must respond appropriately to unexpected events like a child crossing the street. Obtaining sufficient "training" data for such rare events can be costly, unethical or even impossible. Therefore, the lack of robustness could compromise the performance of the models even if specific data are included in the model, leading to distorted decisions.

An adequate operational measure to test robustness must, therefore, consist of retaining the predictive accuracy of the model if there are disruptions in the data or changes in the scenarios considered. One possible solution could be extending the RGA measure based, in this case, on the direct comparison of the natural arrangement of the predictions from undisrupted data with the arrangement of the same as a result of the ranks of the predictions obtained from disrupted data.

4.1.3 Principle 3: Fairness

To reduce inequities and improve sustainability at the same time, KTIs must be formalised that measure the potential distortion of AI applications in correspondence with specific population groups.

For fairness, the neural network models easily adapt to individual data schemes and are, therefore, also capable of rapidly adapting to distorted data and, in particular, to subsets/supersets of the specific reference subpopulations. In this context, the distortion may be measured ex-post according to the predictions obtained, by comparing the average values expected in the various groups in the reference population, or by deriving the respective distributions of conditioned probabilities.

Using the main properties that distinguish the RGA measure, a possible KTI to propose in order to evaluate fairness would be a new version based on the direct comparison of the rank of a specific statistical unit within the reference population with the marginal rank that it has within its category of the population.

4.1.4 Principle 4: Efficiency

One topic of particular interest following that set forth in the European Commission's proposed regulatory framework on AI relates to the issue of sustainability in terms of energy efficiency. In establishing the KTIs, there arises the need to measure the efficiency of the AI applications, both in statistical terms (i.e., how much data are necessary to achieve adequate predictive capacity) and in computational terms (i.e., how much time and resources are necessary to process a model that ensures the achievement of adequate predictive capacity). Satisfying the efficiency requirement, therefore, entails a concurrent improvement in energy efficiency.

As is commonly known, neural network models typically require abundant computational resources and, at the same time, the availability of vast quantities of high-quality data. This is due to the fact that they learn solely based on data.

Despite their advantages, many AI techniques (i.e., "good old-fashioned Artificial Intelligence") are known to be inefficient in execution (mainly because of the computational complexity) and vulnerable to specific dynamics, at times incomplete and uncertain. Moreover, the acquisition of knowledge of specific contexts and their representation in explicit theoretical models (ontologies) by human experts is extremely laborious and often lacking in practical applicability. The applicability of these models could be improved if they were based on mathematical formulas developed by experts in the relevant context as hybrid AI models.

Finally, the concept of efficiency goes hand-in-hand with that of adequacy. The word "efficiency" takes on a technical connotation that mostly reflects the concept of "data efficiency". In establishing specific KTIs, adequacy might be preferable to efficiency. Indeed, AI methods may be more or less adequate, but the use of more limited quantities of data (especially those relating to the personal sphere and therefore sensitive) could entail more adequate AI methods in addition to more sustainability in terms of costs.

4.1.5 Principle 5: Explainability

AI methods are often described as black boxes in which inputs are processed, producing outputs that are used for specific decision-making processes. Not being able to establish the internal mechanisms for the processing of the inputs can lead to incorrect decision-making processes with extremely adverse consequences. This gives rise to the need to ensure transparency in the application of AI methods by establishing KTIs that can measure the explainability/interpretability of the predictions that they generate.

Unlike the previous principles, which are assessed with respect to the application of the ML model to all the available data (at a global level), explainability is a local characteristic to be assessed specifically for each prediction.

In this respect, local methods based on the implementation of game theory concepts may be considered (like Shapley values) or based on simulations or local approximations (such as the LIME methods) (see Babaei et al. [2021]), and, alternatively, global methods may be considered based on the principle of statistical parsimony (like Bayesian model averaging, the ANOVA decomposition and the Shapley-Lorenz decomposition) [see, e.g., Giudici and Raffinetti 2020, 2021a, c].

As for the transparency requirement (interpretability/explainability), neural network models, as parametric estimators of statistical functions, are characterised by their ability to predict the desired outputs very accurately. However, it is general difficult to establish how much the neural network has learned and the dynamics underlying the generated output. Explainable AI ("XAI"—eXplainable Artificial Intelligence) includes certain methods like, for example, LIME (local interpretable model-agnostic explanations which do not depend on the model) and SHAP (Shapley Additive Explanations), which make neural networks interpretable and explainable. Although the automated learning process has undeniably seen progress in terms of the extraction of post-hoc explanations, these results cannot be generalised to cover architectures or arbitrary problems. Furthermore, it remains to be seen how reliable or useful the resulting explanations effectively are. One possible solution to the problem could be explainability by design, in which knowledge of the context, often in terms of knowledge graph embedding, is integrated ex-ante in the neural networks.

The operational measures used to test explainability are defined locally, i.e., for each individual observation. One promising alternative could be the use of the new RGA measure, which acts locally in terms of explainability and globally in terms of predictive accuracy. Specifically, the new operational measures to evaluate interpretability could be established by extending the RGA measure globally based on the setting for the Shapley values or LIME methods.

4.2 Interpretability and Stability of the Models' Outcomes

The use of AI models generally produces performance results expressed, for example, in terms of accuracy, that surpass those observed for classic models based on parametric estimates (e.g., logit/probit). However, the accuracy to be gained comes with greater methodological complexity, which focuses attention on interpretability compared to traditional models. The AI models are more complex than traditional models with respect to methodology, estimation algorithms and computational logic.

Referring to the principle of explainability[2] described in Sect. 4.1, the complexity of the AI models, therefore, paves the way to questions about the interpretability and usability of these models by the various stakeholders that interact with them in different ways, the various perspectives that they adopt in this context. The interactions of the different stakeholders with the AI models are generally both for strictly "management" purposes entailing the use of models to support decision-making and for supervisory reporting purposes as well.

The importance of interpretability in the context of decision-making processes highlights an important connection with the concept of the robustness, or stability, of the models themselves. Referring to Sect. 4.1 the stability of a model is its ability to retain predictive accuracy when unexpected disruptions in the data or in the model's application scenarios occur, which could lead to distorted decisions based on its use. Interpretability and stability are considered important features associated with a model in the context of its use in decision-making processes, because

[2] Explainability and interpretability are used as synonyms in this context.

while interpretability relates to the use of models in predictable conditions, stability provides information on how the model will behave in unexpected situations.

Below we explore both these concepts after introducing the main legislation on interpretability and stability.

4.2.1 Main Legislation

Prudential supervisory regulation allow banks to develop internal models to calculate their capital requirements. These internal models are subject to case-by-case approval by the supervisory authority.

The main regulations on algorithms contain general principles, and there are no explicit instructions not to use innovative algorithms like ML. The main articles that must be taken into account when validating the models are article 174 and subsequent articles of CRR 575/2013 and, for machine learning in particular, CRR 575/2013—art. 179 a) "*an institution's own estimates of the risk parameters PD, LGD, conversion factor and EL shall incorporate all relevant data, information and methods. [...] The estimates shall be plausible and intuitive and shall be based on the material drivers of the respective risk parameters. [...]*".

Although at first this regulatory structure made it difficult to implement AI methods, with its focus on the interpretability of the results, it is now converging towards the more and more frequent use of these techniques, even in the context of IRB models, thanks to the development of sophisticated techniques to explain the results.

Techniques to interpret the results are fundamental where the supervisory authority must fully understand not only the outputs but also the implications and functioning of the internal model.

Another important aspect of interpretability relates to legislation protecting the use of customer information. In this context, the interpretability of results and the understandability of the AI algorithms can help support the assessment of ethical and moral aspects (refer to Sect. 5.1 "The role of AI models in the credit risk of tomorrow") inherent to the outputs of these models, on which a bank could, for example, base its decision to approve a loan. Indeed, credit risk models have a determinant influence on credit approval and other important bank decisions and this is why the manager, valuers, the customer and the regulator must be able to understand and interpret the results and the logic that led to them.

Moreover, it is important to consider how the regulator assesses the inclusion of expert judgement[3] in statistical models and, consequently, decision-making processes, which in any case should be included in models based on AI techniques.

Turning to a more general context of technological development and innovation and the related legislation, we see how regulators are addressing how to encourage the integration of technological innovation and, in particular, artificial intelligent in the estimation processes subject to validation. In this respect, in January 2020, the EBA published a deep dive review on the use of big data and advanced analytics in the banking industry.[4] This report stressed the understandability of the model and the interpretability of results as fundamental elements of trust. The Financial Stability Board also referred to interpretability as crucially important from various angles, including for the understanding of potential effects on banks' balance sheets in a systemic shock.[5]

In 2018, the European Commission published a document[6] containing ethics guidelines for AI, prepared by a group of AI experts. In this document, the European Commission described the explainability and interpretability of AI algorithms as fundamental requirements for trustworthy AI. Finally, in a very recent publication, dated November 2021,[7] the EBA reiterated the importance of the interpretability of the results of AI models, but did not rule out the possibility of using them in internal rating-based (IRB) approaches, and even went so far as to open the door to the use of these systems: "*Depending on the context of their use, the complexity and interpretability of some ML models might pose additional challenges for the institutions to develop compliant IRB models. Nevertheless, this might not necessarily rule out the use of these models*".

[3] See article 180(d) CRR 575/2013 and updates.

[4] "Report on Big Data and Advanced Analytics"—EBA January 2020.

[5] "[…]Yet the lack of interpretability will make it even more difficult to determine potential effects beyond the firms' balance sheet, for example during a systemic shock. Notably, many AI and machine learning developed models are being 'trained' in a period of low volatility. As such, the models may not suggest optimal actions in a significant economic downturn or in a financial crisis, or the models may not suggest appropriate management of long-term risks.[…]"—Artificial Intelligence and machine learning in financial services—FSB 2017.

[6] "Ethics Guidelines for Trustworthy AI"—European Commission June 2018.

[7] EBA Discussion Paper on Machine Learning for IRB models.

As for the stability requirement applicable to credit risk measures, the legislation has generally devoted particular attention to the considerations relating to how the models behave when faced with expected and/or unexpected changes in the scenarios in which they operate.

In particular, stability has been an important aspect for PD models since the publication of the Working Paper 14 BIS, which in May 2005[8] addressed the problem of rating philosophy in the context of supervisory models, distinguishing between point-in-time (PIT) ratings and through-the-cycle (TTC) ratings.

In particular, with respect to obligor-specific PDs, Erik A Heitfield (at the time a member of the Board of Governors of the Federal Reserve System, Washington, DC) noted that "*Obligor-specific PDs that incorporate current credit-quality information and do not impose stress-scenario assumptions are likely to change rapidly as prevailing economic conditions change. They will tend to fall during upturns of the business cycle and rise during economic downturns* (PD PIT). *Obligor-specific PDs that do not incorporate dynamic information on credit quality or that impose stress-scenario assumptions will tend to remain relatively stable over the business cycle* (PD TTC). *Assuming that pooled PDs accurately reflect the average level of obligor-specific PDs, the characteristics of obligor-specific PDs will have an influence on the extent to which overall credit-risk capital requirements vary over the business cycle*".[9] The BIS' paper then pointed out that a crucial part of the decision also depended on "*supervisors' assessments of the tradeoffs between the benefits of credit-risk capital requirements that are sensitive to changing economic conditions versus the benefits of capital requirements that are relatively stable over the business cycle*".[10]

The stability of risk assessments to be performed by calibrating PD through-the-cycle was also emphasised by the EBA[11] in its guidelines detailing the new requirements for AIRB models. In this respect:

- "Where the rating assignment process is highly sensitive to economic conditions, grade assignment will change significantly, while default

[8] Basel Committee on Banking Supervision, Working Paper no. 14, "Studies on the Validation of Internal Rating Systems", Revised version, May 2005.

[9] Ibid., p. 11.

[10] Ibid., p. 11.

[11] EBA/GL/2017/16, "Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures", 20 November 2017.

rates of each grade will remain relatively stable. In contrast, when the assignment is less sensitive to economic conditions, the yearly default rates per grade component will capture the cyclicality of the global default rate"[12];

- "Institutions that have a ranking method in place which is very sensitive to macroeconomic conditions, reflecting point-in-time (PIT) rating philosophy but also aiming to obtain more stable through-the-cycle (TTC) PD estimates in their calibration and recalibration process (to ensure a certain degree of stability of capital requirements over the credit cycle), might show the same cyclicality of capital requirements as institutions that incorporate risk drivers less sensitive to the economic environment into their ranking method"[13];
- "institutions are asked to define a cycle for a fundamental review of models depending on the materiality of the models considered. In accordance with the GL, the scope of annual review includes the analysis of representativeness, the performance of the model, its stability over time and its predictive power. Full reviews including model design may be performed on a less frequent basis".[14]

This means that the use of AI techniques in credit risk assessment models must take into account this impact on the stability of the final risk metrics, adjusting, where necessary, for excessive pro-cyclicality.

4.2.2 Interpretability Methodological Notes

Different stakeholders approach assessing the validity of a predictive model from different perspectives. Consequently, the methodological approaches to the interpretability of AI models may be classified according to the perspective of who is interpreting them.

Methods that produce sensitivity analyses, based on a principle of manipulating inputs and observing the change in predictions, have found excellent reception. In this context, one family of methods focuses on the interpretability of local predictions, using methods like Shapley values, individual conditional expectations, counterfactual explanations

[12] Ibid., p. 20.

[13] Ibid., p. 24.

[14] Ibid., p. 44 and par. 217, p. 100.

and saliency maps. Similarly, sensitivity analyses may also be carried out on the behaviour of models on average or using global methods, e.g., partial dependence plots, individual conditional expectation curves, functional ANOVA, permutation feature importance, etc.

Finally, it is important to cite the interpretability methodologies based on surrogate models, i.e., interpretable models designed to replicate the behaviour of AI models. Surrogate models may be applied either at global or local level. LIME (local interpretable model-agnostic explanation) methods are used for the latter.

Based on the foregoing introduction, below we would like to focus on the application of interpretability methods for non-parametric AI models that combine the use of global surrogate models with the use of local sensitivity models. Indeed, the combination of these methodologies creates a valid methodological framework for the overall interpretability of an AI model.

4.2.3 Key Methodologies

To ensure effective interpretability, in certain cases, a surrogate model may be identified (e.g., a logit/probit model) to be used as a benchmark, adopting a reverse engineering approach, i.e., reconciling a more complex approach with one that is less sophisticated to then analyse the differences between the two. This approach makes it possible to compare the classic overall accuracy measures (e.g., ROC, AUROC, Brier score, etc.) to evaluation the better performance achieved with more complex methods.

The next step is to identify the contributions provided by the individual risk drivers, which is more complex for AI models than would be possible taking traditional logit/probit approaches. Recent academic research and the experience of the participants in this working group show the usefulness of adopting agnostic Shapley-based methods, where for each value in the statistical series associated with a specific point, it is possible to determine the marginal contribution to the overall estimate of the effect of interaction with the other indicators. The Shapley value expresses the distance between one point in the series and a middle measurement to be used a reference. The middle measurement is expressed in accordance with the methodology adopted while the difference is expressed as a different measurement depending on the type of point in the series (information expressed in different measurement units may coexist in a multi-varied distribution, which estimates a given target value). As previously noted in Sect. 4.1, starting with this basic assumption, recent studies have proposed normalisation methodologies (Lorenz Shapley measures*)* to resolve differences between measurement units.[15] The average of the marginal contributions for the points in the series relating to the risk drivers expresses a summary value that expresses the specific indicator's contribution:

$$E[C(w_{x1})] = E[F(x_1, x_2, \ldots, x_n) - F(x_2, x_3, \ldots, x_n)].$$

4.2.4 Focus Points

The working group noted the following focus points with respect to the interpretability methodologies introduced above. These critical points give a general view of the aspects to consider with regard to the interpretability of the results of the AI models.

It is important to bear in mind that while the aforementioned interpretation methodologies provide explanations about the functional aspects of the models and the sensitivity of the respective predictions, they do not associate a measure of confidence or uncertainty to them. This is an important aspect for example in the context of model risk management frameworks, which typically require the assessment and management of the model risk based on an appreciation of the uncertainty of the related estimates. In this regard, the uncertainty surrounding the estimates produced by AI models is generally addressed by implementing structural or distributional estimate hypotheses, which then make it possible to develop instruments that diagnose the associated uncertainty.

From a purely technical standpoint, the interpretability techniques proposed generally reflect features dependence in data, which is to say they suffer an empirical limit in the assignment of importance and effects to correlated features. While there are interpretability methods that approach this problem using techniques that alter the structures of data correlation—e.g., permutation techniques—it is important to emphasise how these methods produce results outside the domain of expected distribution and consequently limit the applicability of the interpretation.

[15] See N. Bussmann, 2021, R. Enzmann, P. Giudici, E. Raffinetti, "Shapley Lorenz values for Artificial Intelligence risk management".

In conclusion, the literature does not currently provide unambiguous guidance on how to measure interpretability capacity. In other words, studies are still ongoing about how to evaluate whether one interpretability method is better than another, based on both objective assessments—e.g., mathematically quantifiable metrics—and subjective assessments. In this respect, one proposal is particularly interesting, that of a "universal" measure, which could be the RGA metric introduced in Sect. 4.1. As described, it could contribute to local interpretability as well as global predictive accuracy.

4.2.5 Stability Methodological Notes

By their very nature, AI models are highly sensitive to changes in the application data, because they are geared towards capturing the best-fitting solutions based on the available data. Therefore, the risk measures that these models generate are normally characterised by a high degree of fitting and present more intertemporal variability than "traditional" statistical models.

Certain helpful expedients may be used when the AI models are trained to increase their through-the-cycleness and ensure that their application generates less variability of the resulting risk measures. An initial, simple solution could be to expand the size of the training set in terms of both time and the number of observations. In any case, the results of the AI models must be evaluated to identify how point-in-time (PIT) their results are. The more PIT the results, the more important it becomes to perform a solid estimate calibration that leads to PDs that present greater relative stability.

References

Babaei G., Giudici P., Raffinetti E., "Explainable fintech lending", 2021.

Basel Committee on Banking Supervision, Working Paper no. 14, "Studies on the Validation of Internal Rating Systems", Revised version, May 2005.

Bussmann, N., Enzmann, R., Giudici, P., Raffinetti, E., "Shapley Lorenz values for Artificial Intelligence risk management", 2021.

EBA, "Report on Big Data and Advanced Analytics", January 2020.

EBA/GL/2017/16, "Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures", 20 November 2017.

European Commission, "Ethics Guidelines for Trustworthy AI", June 2018.

Giudici, P., Raffinetti, E., Lorenz Model Selection, Journal of Classification, Volume 37, Issue 3, pp 754–768, 2020. https://doi.org/10.1007/s00357-019-09358-w.

Giudici P., Raffinetti E., "Shapley-Lorenz eXplainable Artificial Intelligence, Expert Systems With Applications", Expert Systems With Applications, Volume 167, Issue 14, p 104, 2021a. https://doi.org/10.1016/j.eswa.2020.114104.

Giudici P., Raffinetti E., "Explainable AI methods in cyber risk management, Quality and Reliability Engineering International", SSRN Electronic Journal, pp 1–9, 2021b. https://papers.ssrn.com/sol3/papers.cfm?abstractid=3,883,422.

Giudici P., Raffinetti E., "A generalised ROC curve", SSRN Electronic Journal, 2021c. https://ssrn.com/abstract=3892652.

CHAPTER 5

Possible Evolutions in AI Models

Abstract This chapter describes the possible evolution of AI models in the credit risk of tomorrow, evaluating the Regulatory position with respect to the implementation of such techniques with reference to the credit risk assessment, the position of the players in the market, the evaluative economic and macroeconomic environment. Also, an analysis of the outcomes of the AI models is reported, detailing the main aspects concerning ethics, transparency discrimination and inclusion.

Keywords Machine learning—ML · Artificial intelligence—AI · Evolution · AI models · Credit risk · Credit risk assessment · Credit risk evaluation · Credit risk estimation · Result · Model outcome · Ethics · Transparency · Discrimination · Inclusion

5.1 The Role of AI Models in the Credit Risk of Tomorrow

When the COVID-19 pandemic began spreading on a global scale in early 2020, public officials in the world's major economies began preparing dashboards of "high-frequency" data, such as daily airport passengers and credit-card-spending over the course of a day. These data were analysed to guide the policy that nearly all governments around the world were forced to abruptly adopt in response to a completely unforeseen threat.

R. Locatelli et al., *Artificial Intelligence and Credit Risk*,
https://doi.org/10.1007/978-3-031-10236-3_5

Gradually, in subsequent waves of COVID-19, new data were added to the data initially used, expanding the analyses.

The effectiveness of assessments based on high-frequency data, combined with the fact that high-frequency information covers countless areas in the lives of people and companies, has garnered the attention of economists, triggering the production of an extraordinary number of papers during the COVID-19 pandemic and giving life to a new approach to economic analysis which *The Economist* has defined "instant economics" (Chart 5.1).[1]

There is an obvious phenomenon at the foundation of this evolution: today, nearly all economic transactions and an enormous number of daily activities, including non-economic activities, involve computers, which automatically cause the accumulation of vast amounts of information referring to just as vast a group of people and companies.

The change in the economic context driven by the electronification of our lives is so significant that it has inevitably influenced the way in which the creditworthiness of individuals and companies must be analysed by banks and others competing with them to provide financial services.

The authorities that regulate and supervise banks have also addressed this change of context, as noted in various sections of this document.

In terms of rules, the EBA recently devoted a consultation paper to the challenges and opportunities of using AI techniques in IRB models. The paper, which will be followed by guidelines to help define the playing field on which AI is implemented in banking rating systems, clearly acknowledges the usability of ML techniques to quantify capital requirements for credit risk in the context of IRB models.

[1] The Economist, 23 October 2021, "The real-time revolution". According to The Economist, instant economics is such a revolutionary development that it has marked the third structural change in economic analysis since it began. In particular, according to the British weekly, the first wave of economics emerged with reflections on the big theoretical questions, starting with Adam Smith's "Wealth of Nations" in 1776. The second wave coincides with modern economics, in which theoretical analyses are supported by robust empirical evidence made possible by the availability of public statistics sources and advancements in computational capacity. The limit of these analyses is the delay between when the information is gathered and when it is made available to the researchers, largely due to the need to verify public source data. This limit is relaxed in third wave economics, when teams of researchers use large datasets containing granular and timely information to rapidly respond to practical economic problems, such as the real-time measurement of the performance of an economic sector or the impact of a policy choice just made.

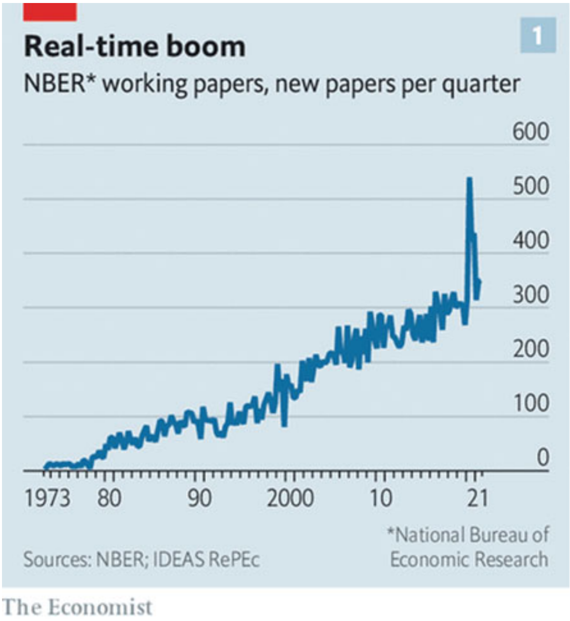


Chart 5.1 Trend in the number of working papers published according to an analysis by The Economist

In terms of supervision, since the onset of the COVID-19-triggered crisis, the European Central Bank (ECB) has demonstrated that it is clearly aware of the usefulness of credit management approaches geared towards incorporating AI, having encouraged banks to use all the available data—especially unstructured data—to rapidly gain an understanding of the effect of the pandemic on creditworthiness and include forward-looking elements in these analyses.

Furthermore, the ECB has also approved requests to change previously authorised IRB models to exploit unstructured data using AI techniques and therefore paving the way for the EBA's regulatory initiative noted earlier.

While it is, therefore, clear that the use of AI techniques and more or less instant data is, in many ways, a required course of action for credit risk analysis, one of the aspects that must be understood relates to how the new techniques will be used to assess risk. In other words, it remains to be seen whether the AI-based models that use unstructured data will replace traditional approaches and information or if they will supplement them, making them more predictive and responsive to context changes.

Since, most likely, the solution will vary depending on the purpose of the credit analyses, one clue might be the role that instant economics analyses are filling with respect to economic predictions.

In this regard, a perception is emerging that while the methodological approaches based on high-frequency data are extremely useful in rapidly capturing turnarounds in the economic cycle, they meet with more difficulty when they are used to precisely quantify the level of activity. In this field, traditional techniques maintain a competitive edge because of the higher quality of official statistics.

Similarly, it is plausible that traditional models will retain a role in credit risk analysis, as AI-based models progressively support them in fields where they are superior in connection with the type of information to be processed and the speed with which they make use of this information.

A key element to bear in mind in this structural credit risk modelling change relates to all cases in which AI techniques are applied to aspects of human life and concerns the need to preserve ethics and transparency

5.2 Ethics and Transparency of Results

This paragraph deals with the main aspects relating to the ethics and transparency of results of AI models, namely privacy, transparency, discrimination and inclusion.

Rapid technological progress in recent years is speeding up the adoption of AI by a growing number of financial institutions and this trend will continue in the years to come (EBA, 2020). Given its many benefits, the boundaries must be traced within which operations using these methodologies do not pose potential risks to the security of users.

Privacy and transparency are two key elements used to evaluate whether to deploy AI technologies. One of the key aspects relating to this problem can be summed up as machine unlearning: how to have an AI machine or model unlearn information.

5.2.1 Privacy

Within the methodology, privacy mainly refers to protecting the information in the training set (Shokri et al., 2019). It is in this step that the models might include the processing of personal data, which sometimes falls into the category of "sensitive data", like tax identification numbers, images, fingerprints, telephone traffic, medical information, biometric data or criminal sentences.

In this context, policymakers and regulators are careful to promote the principles of transparency and privacy for self-learning models (EC, 2020).

Legislation protecting personal data refers, in particular, to the principles and provisions in Regulation EU 2016/679 (General Data Protection Regulation, "GDPR") in force in the European Union since 25 May 2018 and Legislative decree no. 101/2018, which harmonised the Italian Personal Data Protection Code with the European Regulation.

In particular, according to the principle of accountability under the GDPR (Art. 5) institutions must be able to guarantee and demonstrate compliance with the principles established for personal data processing. In this respect, the GDPR—both for companies in general but especially for financial intermediaries—is useful in strengthening customers' trust with respect to how institutions process their data (EBA, 2020). The operators of this processing are defined as controllers and processors and "processing" is defined as (Art. 4) any operation performed on personal data, such as storage, organisation, alteration, retrieval, consultation, dissemination and erasure. As for this last point, Article 17 of the GDPR specifically protects the data subject by establishing the right to erase the data when they are no longer in the public's interest ("right to be forgotten").

With respect to this specific matter, Article 22 of the GDPR establishes a general principle whereby each data subject (natural person) has the right to not be subject to a decision based solely on the automated processing of his or her data if it produces legal effects concerning him or her or similarly significantly affects him or her. Although the rest of the article establishes exceptions to this principle, it is important to always bear in mind that, even in privacy by design, the European Regulator has taken this "protective" approach.

According to the same GDPR (Art. 32), institutions are required to take specific data protection measures to protect the data subject, with the appropriate technical and organisational structures. These may be pseudonymisation and encryption of personal data to prevent customers from being directly identify. Furthermore, Article 35 governs the introduction of the Privacy Impact Assessment to identify and correct risks to the rights and freedoms of the natural persons to which the processed data refer (both material—financial losses—and immaterial—discrimination, identity theft and reputational damage). These risks arise when there is a data breach, which must be suitably covered in the impact assessment procedures.

Intermediaries must, therefore, provide for all the aforesaid organisational and technological procedures to prevent the risk of sanctions by the authorities and avoid adverse events to the detriment of their customers, in order to preserve and strengthen the trust-based relationship at the core of operations.

5.2.2 Transparency

The transparency of a rating system is closely connected to its interpretability and the communication of how it functions and its outcome. This was already discussed in Sect. 4.2 “Interpretability and stability of the models’ outcomes”.

Communication obviously relates to all stakeholders in the system and, therefore, not only the parties subject to the credit scoring and decision, but also management and governance bodies, which are responsible for evaluating and approving the functioning in the context of business processes, the internal and external bodies and units responsible for validation and review, etc.

In general, the transparency requirement may be defined in terms of the process by which the system is constructed (the model design and development procedures that lead to the system’s output) and in terms of its outcome (content and justification of the result).

The second aspect is essentially important from the viewpoint of ethical repercussions. So the ethical aspect of transparency can be considered connected to the expectations of the loan applicant to know the reasons for the decision and specifically the reasons why the loan was denied.

More broadly, transparency may relate to the entire credit rating process, not just the final decision. This is especially true in financing for production activities, where the identification of strengths and weaknesses is a crucial part of the bank/company relationship.

Based on these premises, it is clear that transparency relates to any credit-related decision-making process, from a system based on experiential rules to ratings based on statistical algorithms or AI models.

Moreover, it is just as clear that as the system becomes more complex, it also becomes more difficult to interpret it and communicate its outcome. In a system based on rules, like the instalment/income ratio in retail loans, a breach that triggers denial is immediate. In a rating based on a plurality of aggregate data using multi-varied statistical techniques, the identification of the indicator or the combination of indicators that determines a certain assessment is certainly more complex, but there are long-standing statistical procedures and instruments that make it possible to interpret the outcome of the model.

Ratings based on AI constitute another qualitative step further in the direction of complexity, due to the expansion of sources and the processing algorithms. The pursuit of an adequate level of transparency means taking the appropriate precautions with respect to both aspects: in terms of the data, it being understood that expanding the dataset makes it possible to achieve, as described earlier, benefits not only in terms of predictions but also in terms of financial inclusion, with positive repercussions on ethical aspects, the mitigation of the risk of opacity entails analysing the individual variables introduced with respect to their materiality compared to the dependent variable, potentially excluding those whose relationship with credit risk is difficult to explain; in terms of the processing algorithms, reference must be made to that described in the section on the interpretability and stability of results, which are therefore of primary importance from the ethical standpoint of the transparency of information shared with the loan applicant.

5.2.3 Discrimination

To address the matter of discrimination, we must begin with the concept of bias, whose definition in statistics refers to the weakness of a model's training set in representing the entire population of application. When this occurs, the results could favour certain groups over others. This is because AI models learn from the past to predict the future and closely depend on the data used for their training. The social definition of bias is instead tied to human judgement based on preconceptions or prejudices, rather than an impartial assessment of the facts. In both cases, bias can lead to discrimination.

It should be emphasised that bias and the discriminatory impacts are not exclusive to the use of new analytical techniques like ML. Indeed, the human decision-making process is inherently distorted, either intentionally because of prejudice, or unintentionally because of limited knowledge or experience. In this context, the use of more automated credit scoring systems could have the effect of reducing discrimination by increasing the coherence and consistency of treatment and minimising individual judgement and discretion. Nevertheless, ML has led to the establishment of algorithms that learn from data that reflects human distortions and often reproduces these distortions, even amplifying them. In particular, spurious and potentially misleading correlations might have significant implications given the automated nature of ML algorithms and the underlying data could even lead to representations that increase the distortions and discriminations in society. Financial institutions already have a governance model, a model risk framework, processes to check the integrity of lending processes that frame and govern the statistical models used for credit decisions. Now they may adapt the current governance and risk management framework to ensure an ethical use of new technologies like ML. In the design and development of a framework that is not discriminatory but is inclusive, the most critical step is identifying where the distortions lie and how they can arise.

First and foremost, the ML training data, or the statistical data, merely reflect human history and previous decision-making processes. Consequently, they carry the prejudice and discriminatory potential of past experience. The training data of the models may introduce distortions and, therefore, discrimination when they do not adequately reflect the real diversity that distinguishes the entire population. They are, therefore, used to make generalisations about a group of subpopulations that differ from one another significantly, but that are not all well represented in the dataset.

Associative bias occurs when certain data in the training set are closely correlated with certain characteristics that may be sensitive and protected, like race. In this case, not even excluding demographic information from the algorithm to avoid discrimination would resolve the problem of what is called redundant encoding because, if the training dataset is packed with granular, diversified data, the algorithm, given specific pieces of highly correlated information, may deduce, even implicitly, certain sensitive characteristics.

For example, an ML algorithm built to predict the repayment probability of a loan excluding sensitive characteristics might learn from encoded data like where the applicant lives in order to discriminate based on race. It is, therefore, fundamental to carefully analyse the data's properties to ensure there are no direct distortions or indirect distortions hidden in their characteristics.

Certain distortions may be incorporated through the data cleansing and transformation process. For example, the creation of derivative variables can increase and aggregate certain attributes, while minimising others. This is why the process requires very careful analysis.

This clearly shows how, to avoid the risk of discriminating even unintentionally against a group of the population, it is necessary to design, verify and construct ML models with inclusion as a primary objective from the outset.

5.2.4 Inclusion

The current methodology that banks adopt for customer risk profiling is based on the clustering of socio-demographic and geographical information and statistical analyses of behaviour derived from the historical data of "regular" customers.

These models were also created in order to comply with regulatory requirements that limit historical observations to a minimum period of five years, meaning customers' behaviour cannot be readily and correctly captured, nor can their specific needs.

One major consequence of using models based on regulatory methodologies is that they are not sufficiently dynamic to predict customers' potential financial hardship and, above all, would qualify a very large group of customers like millennials (2.3 billion people in the world, including 11 million in Italy) as unrated or it would rate precarious workers and, in general, anyone without a credit history (e.g., those who previously worked abroad) as high risk.

The various papers that the EBA has published over the years (e.g., "*GLs on NPL and Forbearance*", 2018, or "*GLs on Loan Origination and Monitoring*", 2020) have led banks to study innovative models to supplement their mandatory models, enabling them to capture, thanks to early warning signals, their customers' smallest financial difficulties, while also correctly assessing them during loan origination.

AI is the tool used to develop and manage these new models and the available information sources may be internal, to determine transactional scores based on analyses of individual transactions over time, or external, using open-source data.

Transactional scoring is used to profile a customer's habits, analysing information on their utilities, taxes, incoming and outgoing bank transfers, credit card use and loan repayments, to define whether the customer is a reliable payer. Not only does this improve the credit scoring process but it also enables the bank to apply better interest rates.

The regular use of these scoring techniques for individual customers is usually combined with traditional models to equip the bank with a matrix that covers both their supervisory and management reporting needs, while also making the scoring/rating more dynamic and up to date.

PSD2 has significantly boosted the construction of transactional models, as it gives banks access to customers' accounts with other banks, for bigger datasets and, consequently, enabling them to define more accurate and predictive parameters, obviously with the customers' authorisation and based on agreements between the banks themselves.

To cover non-customers, operators may gather the information they need to build the model from open sources, which is, however, a very delicate method to use in terms of the sensitivity of the personal data and the vast amount of non-uniform information to be analysed.

The data usually used derive from telephone use, text messages, social media and online purchases, painting a picture of the customer's habits in order to score risk and meet business needs in the definition of products and prices.

It is implicit that the processing of retail customers' data must always be carried out in accordance with the GDPR and local regulations, and on this basis some of the information may not be processed for regulatory reasons.

An evolution of this type makes it possible to cover both risk management needs and a business gap, in the sense that younger generations are less willing to have physical contact with an advisor at the bank branch and prefer flexible, quick and easy tech support without the constraints of time and place.

It is precisely through AI that banks will be able to develop advanced platforms offering flexible products that meet lending and saving needs for the entire range of banking products.

Today, young consumers go online to find services with competitive prices or returns and that are, in certain cases, high risk, and they might even rely on online forum reviews to evaluate the services.

The greatest challenge for the future will be played out in IT and AI developments as the fiercest competition comes from fintechs and crowdfunding/crowdlending, which offer instantaneous, flexible and dynamic solutions that benefit from the lack of specific regulations applicable to those types of companies, whereas such regulations are a significant burden for traditional banks.

References

EBA, “Report on Big Data and Advanced Analytics”, 2020.

ECB, “Identification and Measurement of Credit Risk in the Context of the Coronavirus (COVID-19) Pandemic”, 2020.

European Commission, “White Paper on *Artificial Intelligence*—A European Approach to Excellence and Trust”, 2020.

Shokri, R., Strobel, M., Zick, Y., “Privacy Risks of Explaining Machine Learning Models”, 2019.

Appendix

The present manuscript is the result of a research and discussion process organised by Aifirm. All the members of the Working Group (Commission) are Aifirm Members. The organisational structure of the Aifirm Commission is described below. The editors of the present manuscript have coordinated with different roles the Commission's work and have supervised the final report. The Steering Committee is formed by the persons that played a coordinating role in one or more of the working groups that are responsible for single paragraphs of the manuscript, has specified below.

THE COMMISSION's SCIENTIFIC COORDINATOR

- **Rossella Locatelli** | Università degli Studi dell'Insubria

THE AIFIRM COORDINATOR

- **Fabio Salis** | Credito Valtellinese

THE STEERING COMMITTEE

- **Stefano Biondi** | Banca Mediolanum
- **Dario Cavarero** | Intesa Sanpaolo
- **Giovanna Compagnoni** | Banca ICCREA
- **Pasquale Costa** | Banco Popolare di Bari
- **Matteo Crippa** | Standard Chartered Bank

R. Locatelli et al., *Artificial Intelligence and Credit Risk*,
https://doi.org/10.1007/978-3-031-10236-3

- **Silvio Cuneo** | Università dell'Insubria
- **Giovanni Della Lunga** | Banca Monte dei Paschi di Siena
- **Lorenzo Ducci** | Credito Valtellinese
- **Emanuele Giovannini** | UniCredit
- **Paolo Giudici** | Università di Pavia
- **Rita Gnutti** | Intesa Sanpaolo
- **Andrea Minuti** | Credit Agricole
- **Gianluca Oricchio** | Muzinich & Co
- **Carlo Palego** | Banca Di Asti
- **Massimo Pezzini** | Intesa Sanpaolo
- **Emanuela Raffinetti** | Università di Pavia
- **Fabio Salis** | Credito Valtellinese
- **Francesca Scaglia** | Banco Desio
- **Marco Tarenghi** | Mediobanca

THE WORKING GROUP—Chapter 1

- **Pasquale Costa** and **Massimo Pezzini** (coordinators)
- **Alberto Cabrini** | Intesa Sanpaolo
- **Antonio Ciccaglione** | Università di Bologna
- **Giansimone Ghiottone** | Banco Popolare di Puglia e Basilicata
- **Francesco Iacono** | Credem
- **Fulvio Radeschi** | BPER Banca

THE WORKING GROUP—Section 2.1

- **Giovanni Della Lunga** (coordinator)
- **Emanuele De Meo** | Unipolsai Assicurazioni
- **Giuseppe Missaglia** | IBL Banca
- **Nicola Picchiotti** | Banco BPM
- **Alessia Zalunardo** | Credit-Agricole Italia

THE WORKING GROUP—Section 2.2

- **Lorenzo Ducci** and **Gianluca Oricchio** (coordinators)
- **Rino Colorio** | Banca IFIS
- **Sepehr Jafari** | KPMG

THE WORKING GROUP—Section 3.1

- **Dario Cavarero** and **Emanuele Giovannini** (coordinators)
- **Marco Bagnato** | Soft Jam
- **Antonio Ciccaglione** | Università di Bologna
- **Eugenio Maglio** | Avantage Reply
- **Anna Pieretti** | Banca Nazionale del Lavoro
- **Luca Seia** | Intesa Sanpaolo

THE WORKING GROUP—Section 3.2

- **Francesca Scaglia** and **Marco Tarenghi** (coordinators)
- **Riccardo Fiori** | Banco Desio
- **Luca Molinari** | BPER Banca
- **Marco Proverbio** | Mediobanca

THE WORKING GROUP—Section 3.3

- **Gianluca Oricchio** | Muzinich & Co

THE WORKING GROUP—Section 3.4

- **Rita Gnutti** and **Stefano Biondi** (coordinators)
- **Fiorella Bernabei** | Intesa Sanpaolo
- **Riccardo Fiori** | Banco Desio
- **Dario Girardi** | Experian
- **Victor Jalba** | Banca Mediolanum
- **Fabrizio Manstretta** | Banca Mediolanum
- **Roberta Ranaldi** | Intesa Sanpaolo
- **Marco Vignolo** | Intesa Sanpaolo

THE WORKING GROUP—Section 4.1

- **Paolo Giudici** and **Emanuela Raffinetti** (coordinators)
- **Marco Bagnato** | Soft Jam
- **Emanuela Biasotti** | Credit-Agricole Italia
- **Marco Carpineta** | BPER Banca
- **Valeria Lazzaroli** | ARISK

- **Francesco Martoni** | Consultant
- **Marco Riani** | Università di Parma
- **Luciano Tarantino** | Studio Tarantino & Partners

THE WORKING GROUP—Section 4.2

- **Matteo Crippa**, **Andrea Minuti** and **Carlo Palego** (coordinators)
- **Filippo Azzarelli** | Banco BPM
- **Marco Bellone** | Banca Sella
- **M. Simone Fontani** | Avantage Reply
- **Francesco Montorio** | Banca Di Asti
- **Giacomo Petrini** | Banco BPM
- **Boris Fausto Rovani** | Avantage Reply

THE WORKING GROUP—Section 5.2

- **Silvio Cuneo** and **Giovanna Compagnoni** (coordinators)
- **Elisa Corsi** | Volksbank
- **Valentina Lagasio** | Università degli Studi di Roma la Sapienza

SUPERVISED BY:

- **Corrado Meglio** | Vice President AIFIRM

PMO COORDINATION

- **Giovanni Pepe** | KPMG
- **Simona Schiappa** | KPMG
- **Nicole Morselli** | KPMG
- **Riccardo Rossetto** | KPMG

Glossary

ANOVA: Analysis of variance, i.e., all statistical techniques that make it possible to test the difference between two or more sets comparing the internal variability of the sets with the variability between the sets.

Application portfolio: The portfolio of credit exposures included in the scope of application of the PD or LGD model when a risk parameter is estimated.

AUROC: Area Under the Receiver Operating Characteristics, a metric used to measure the performance of the classification of an ML algorithm. In particular, it refers to the area under the ROC curve, which is created in a line chart by tracing the rate of correctly classified observations by the model as belonging to a certain class compared to the rate of observations improperly classified by the model as belonging to the same class.

Calibration segment: A uniquely identified subset of exposures under the scope of application of the PD or LGD model which is separately calibrated.

Calibration: The purpose of this phase is to adjust the output of the integration, ensuring that the resulting PD estimates match the medium/long-term default rate.

Central tendency: Medium/long-term default rate or observed on the overall time series.

R. Locatelli et al., *Artificial Intelligence and Credit Risk*,
https://doi.org/10.1007/978-3-031-10236-3

Development of the model: Part of the risk parameter estimation process that leads to an appropriate differentiation of risks by identifying the relevant risk drivers, applying statistical methods to assign exposures to the rating classes and estimating the parameters of the model.

Development of the modules: the modules are defined to identify the various modules used to analyse the specific characteristics of the exposure in the portfolio in order to quantify its risk level. The development phase entails estimating the model (in traditional PD models this is logistic regression) for each module, where the dependent variable is the dichotomic variable 1/0 (default/non-default) and the independent variables are the specific indicators of each module.

Estimate of risk parameters: The entire modelling process for the risk parameters, including the selection and preparation of the data, the development of the model and calibration.

External performance module: The purpose of this module is to analyse the customers' ability to repay their debt by analysing their behaviour with banks. The module is based on data from outside the bank, i.e., Bank of Italy's central credit register data.

Financial module: This module is used to analyse the customers' ability to repay their debt based on financial statements analysis. It uses external data gathered from the Bank of Italy's financial statements database.

Integration: Process in which the outputs of the various modules are combined in order to estimate the model's overall output. In traditional credit risk models, it is usually performed using logistic regression, where the dependent variable is the dichotomic variable 1/0 (default/non-default) and the independent variables are the outputs of the models of each module.

Internal performance module: The purpose of this module is to analyse customers' ability to repay their debt by analysing their behaviour in terms of payments, income and current accounts. The module is based on the bank's internal data (i.e., data gathered on existing customers).

Latent variable: This is a variable that is not directly observed in the training set but inferred from other variables that are observed.

Machine learning (ML): The study of computer algorithms that can improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so.

Natural language processing (NLP): All the techniques of natural language processing.

One-hot encoding: Encoding technique in which the various categories belonging to a categorical variable are represented by additional numerical variables to make them easier to interpret for an ML algorithm.

Outliers: Observations that are clearly an abnormal distance from the other available observations for a variable.

Overfitting: This is when a statistical model contains an excessive number of parameters compared to the number of observations and therefore achieves an excellent performance on the training set, but a weak performance on the validation sets.

PD model: All the data and methods used as part of a rating system and relating to the differentiation and quantification of PD estimates, used to assess default risk for each borrower or exposure covered by this model.

Point-in-time PD: A measurement of the risk of default in the subsequent year, whose model typically considers information about the borrower and macroeconomic performance indicators.

Rating system: A set of methods, procedures, controls, data and information systems that support credit risk assessment, the assignment of internal creditworthiness scores and the quantitative estimate of default.

Reference dataset (RDS): All the data used to estimate the risk parameters, including the dataset used to develop the model, and the dataset used to calibrate a risk parameter.

Risk drivers: Specific indicators for each module, which include the variables used to capture the risk of borrowers and the specific variables of the products.

Shapley value: The average of all the marginal contributions of the individual input variables of a model compared to the output generated.

Socio-demographic module: The purpose of this module is to analyse the customers' ability to pay their debt based on socio-demographic information (e.g., economic sector, business owner's age, legal status and geographic area).

Structured data: Typically used in "traditional" credit risk models, structured data meet a set of predetermined rules or can be defined in terms of type (date, name, number, characters and address) and mutual relationships.

Target variable: The dependent variable. In PD models, it is the dichotomic variable 1/0 which identifies the default event.

Through-the-cycle PD: A measurement of the risk of default in the subsequent year considering specific information about the borrower based on economic conditions in the long run.

Unstructured data: Data without a predefined model that cannot be organised in rows and columns. Examples of unstructured data are images, audio recordings, videos, e-mails, spreadsheets and objects stored as files.

Validation of the model: All the analyses to verify the adequacy of a statistical model.

XAI: eXplainable Artificial Intelligence, it refers to all the methods and processes that enable the understanding and assessment of the reliability of the outputs generated by ML algorithms.

Bibliography

Babaei, G., Giudici, P., Raffinetti, E., "Explainable Fintech Lending", 2021.

Bank of Italy, 26th update to Circular no. 285, 2019.

Basel Committee on Banking Supervision, Working Paper no. 14, "Studies on the Validation of Internal Rating Systems", Revised version, May 2005.

BCBS, "The IRB Use Test: Background and Implementation", 2006.

Bussmann, N., Enzmann, R., Giudici, P., Raffinetti, E., "Shapley Lorenz values for Artificial Intelligence risk management", 2021.

Capital Requirements Regulation (EU) No 575/2013, 2013.

Carroll, P., Rehmani, S., "Alternative Data and the Unbanked", Oliver Wyman, 2017.

Deutsche Bundesbank, "Measuring the Discriminative Power of Rating Systems", no. 1, 2003.

EBA, "GLs on Loan Origination and Monitoring", 2020.

EBA, "Guidelines on Credit Institutions' Credit Risk Management Practices and Accounting for Expected Credit Losses", 2017.

EBA, "Guidelines on Management of Non-Performing and Forborne Exposures", 2018.

EBA, "Report on Big Data and Advanced Analytics", January 2020.

EBA/GL/2017/16, "Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures", 20 November 2017.

ECB, "Identification and Measurement of Credit Risk in the Context of the Coronavirus (COVID-19) Pandemic", 2020.

European Commission, "Ethics Guidelines for Trustworthy AI", June 2018.

European Commission, "White Paper on *Artificial Intelligence*—A European Approach to Excellence and Trust", 2020.

R. Locatelli et al., *Artificial Intelligence and Credit Risk*,
https://doi.org/10.1007/978-3-031-10236-3

Giudici, P., Raffinetti, E., Lorenz Model Selection, Journal of Classification, Volume 37, Issue 3, pp 754–768, 2020. https://doi.org/10.1007/s00357-019-09358-w.

Giudici, P., Raffinetti, E., "A Generalised ROC Curve", 2021a.

Giudici, P., Raffinetti, E., "Explainable AI Methods in Cyber Risk Management," Quality and Reliability Engineering International, pp. 1–9, 2021b.

Giudici, P., Raffinetti, E., "Shapley-Lorenz Explainable *Artificial Intelligence*, Expert Systems With Applications", Volume 167, 2021c.

IIF, "Bias and Ethical Implications in Machine Learning", 2019.

Kamber et al., "Generalization and Decision Tree Induction: Efficient Classification in Data Mining", 1997.

Kandal, E.R., Schwartz, J.H., Jessel, T.M., Siegelbaum, S.A., Hudspeth, A.J. "Principles of Neural Sciences" V Edition, pp. 1583 ss., 2015.

Laza, S. "The Disruptive Path of Neuroeconomics", Duke University, Las Vegas, 2021.

Leslie, D., "Understanding *Artificial Intelligence* Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector", The Alan Turing Institute, 2019.

Liberati, C. et al. "Personal Values and Credit Scoring: New Insights in the Financial Prediction", *Journal of the Operational Research Society*, February 2018.

Lu Han, Liyan Han, Hongwei Zhao, "Credit Scoring Model Hybridizing *Artificial Intelligence* With Logistic Regression", *Journal of Networks*, 8, 253–261, 2013.

Oricchio, G., Capone, F., Capone, G., Di Pino, G., Florio, L., Di Lazzaro, V. "Linking Cognitive Abilities with the Propensity for Risk-Taking: The Balloon Analogue Risk Task", *Neurological Sciences*, 37, 12, 2016.

Oricchio, G., Capone, F., Capone, G., Di Pino, G., Ranieri, F., Di Lazzaro, V. "The Effect of Practice on Random Number Generation Task: A Transcranial Direct Current Stimulation Study", in *Neurobiology of Learning and Memory*, 2014.

Regulation EU 2016/679 of the European Parliament and the Council, "General Data Protection Regulation (GDPR)", 2016.

Shokri, R., Strobel, M., Zick, Y., "Privacy Risks of Explaining Machine Learning Models", 2019.

SSM-2020–0744, "Identification and measurement of credit risk in the context of the coronavirus (COVID-19) pandemic", Frankfurt am Main, 4 December 2020.

Index

A
Accuracy, 6, 22, 33, 38, 39, 41, 46, 56, 57, 60, 69, 70, 74, 76
Advanced analytics, 71
Alternative credit data, 13
Artificial intelligence (AI), 2, 5–7, 19, 31, 42, 43, 45, 47, 48, 55, 62, 65, 69–72, 74–76, 80–82, 85, 88, 89
Autoencoder, 26, 61

B
Business case, 26, 29, 32, 42, 44, 46, 53, 55, 61

C
Counterparty risk, 36
Covid-19, 3, 6, 36, 53, 57, 79, 80
Credit risk, 4, 5, 14, 29, 57, 60, 72, 80–82, 85
Credit risk assessment, 74
Credit risk model, 4, 29, 55, 71, 82
Current account transactional data, 30

D
Data cleansing, 11, 87
Decision tree, 23, 31, 56
Discrimination, 82, 84–86

E
Early warning, 6, 19, 36, 37, 87
Ethics, 72, 82
Explainability, 69, 72

G
Gradient boosting, 31, 56

I
Inclusion, 29, 71, 82, 85, 87
Interpretability, 32, 38, 39, 41, 43–46, 69–72, 74–76, 84, 85
Interpretability of results, 32

L
Linear model, 42, 44–46
Linear regression, vi

R. Locatelli et al., *Artificial Intelligence and Credit Risk*,
https://doi.org/10.1007/978-3-031-10236-3

Loan approval, 37
Logistic regression, 30, 31, 42, 56, 57

M
Machine learning (ML), 4–6, 28, 29, 31–34, 38, 39, 41, 42, 44, 45, 54–57, 70, 71, 80, 86, 87
Model outcome, 5, 32, 85
Model results, 19

N
Natural language processing (NLP), 30, 31, 34
Neural network, 31, 48, 52, 53, 56–59, 61, 62

P
Part-of-speech tagging, 11
Payment cards, 30
POS transactional data, 30
Probability of default (PD), 28–32, 34, 38, 39, 43, 44, 48, 56, 72, 73, 77

R
Random forests, 22, 23, 31, 38, 39, 41, 45, 56
Regulatory capital, 28
Risk discrimination, 34
Robustness, 5, 62, 70

S
Seasonal business, 34
Sentiment analysis, 11
SHAP summary plot, 32
Social network data, 30, 31
Structured data, 10
Structured information, 30, 31, 34, 48

T
Topic modelling, 11
Traditional model, 4, 6, 34, 38, 42–44, 55, 56, 69, 82, 88
Traditional modelling technique, 42
Traditional technique, 55, 82
Transactional data, 5, 6, 13, 19, 29–31, 34, 47, 57
Transparency, 5, 82–85
Two-step approach, 44–46

U
Unstructured data, 48, 81, 82
Unstructured information, 30, 31, 34
Use case, 43

V
Validation, 21, 39, 55, 56, 58, 59, 62, 71, 84