University of Pavia

Doctoral Thesis

Explainable Artificial Intelligence Methods in FinTech Applications

Author:
Niklas Bussmann

Supervisor:
Prof. Paolo Stefano Giudici

Co-Supervisor:
Alexander Koch

A thesis submitted in fulfilment of the requirements for the

degree of Doctor of Philosophy

in the

XXXVI CYCLE
Electronics, Computer Science and Electrical Engineering

September 2023

Acknowledgements

Dedicated to Larissa, the love of my life and my biggest supporter, and my family, to whom I owe everything.

Abstract

The increasing amount of available data and access to high-performance computing allows companies to use complex Machine Learning (ML) models for their decision-making process, so-called ”black-box” models. These ”black-box” models typically show higher predictive accuracy than linear models on complex data sets. However, this improved predictive accuracy can only be achieved by using more complex and confusing methodologies which leads to a deterioration of the model's explanatory power. This will be further analysed in chapter 1 ”Open the black box” and make the model predictions explainable is summarised under the research area of Explainable Artificial Intelligence (XAI). Using black-box models also raises practical and ethical issues, especially in critical industries such as finance. For this reason, the explainability of models is increasingly becoming a focus for regulators. Applying XAI methods to ML models makes their predictions explainable and hence, enables the application of ML models in the financial industries. The application of ML models increases predictive accuracy and supports the different stakeholders in the financial industries in their decision-making processes.

This thesis consists of five chapters: a general introduction, a chapter on conclusions and future research, and three separate chapters covering the underlying papers. Chapter 1 proposes an XAI method that can be used in credit risk management, in particular, in measuring the risks associated with borrowing through peer-to-peer lending platforms. The model applies correlation networks to Shapley values and thus the model predictions are grouped according to the similarity of the underlying explanations. Chapter 2 develops an alternative XAI method based on the Lorenz Zonoid approach. The new method is statistically normalised and can therefore be used as a standard for the application of Artificial Intelligence (AI) in credit risk management. The novel ”Shapley-Lorenz”-approach can facilitate the validation of model results and support the decision of whether a model is sufficiently explained. In Chapter 3, an XAI method is applied to assess the impact of financial and non-financial factors on a firm's ex-ante cost of capital, a measure that reflects investors' perceptions of a firm's risk appetite. A combination of two explanatory tools: the Shapley values and the Lorenz model selection approach, enabled the identification of the most important features and the reduction of the independent features. This allowed a substantial simplification of the model without a statistically significant decrease in predictive accuracy.

Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence

AUROC Area Under the Receiver Operating Characteristic curve

BaFin Bundesanstalt für Finanzdienstleistungsaufsicht

CAPM Capital Asset Pricing Model

CRD Capital Requirements Directive

CRR Capital Requirements Regulation

EAD Exposure at Default

EBA European Banking Authority

EBIT Earnings before interest and taxes

EBITDA Earnings before interest, taxes, depreciation and amortization

ECAI European External Credit Assessment Institution

ECB European Central Bank

EIS Environmental Innovation Score

EU European Union

ESG Environmental, Social, and Governance

FN False Negatives

FP False Positives

FPR False Positive Rate

GDPR General Data Protection Regulation

GDP Gross domestic product

GPU Graphics Processing Unit

LGD Loss given Default

LIME Local Interpretable Model-agnostic Explanations

MaRisk Mindestanforderungen an das Risikomanagement

ML Machine Learning

MST Minimum Spanning Tree

P2P Peer-to-peer

PDP Partial Dependence Plots

PD Probability of default

ROC Receiver Operating Characteristic curve

ROE Return on Equity

SHAP SHapley Additive exPlanations

SME Small and medium-sized Enterprises

TN True Negatives

TP True Positives

TPR True Positive Rate

WACC Weighted Average Cost of Capital

XAI Explainable Artificial Intelligence

XGBoost Extreme Gradient Boosting

General Introduction

In recent years the amount of produced data increased drastically (Reinsel, Gantz, & Rydning, 2017). Together with simplified access to high-performance computers, many companies are able to use more models to support their decision-making process (Jordan & Mitchell, 2015). In some areas the amount of available data is too big and too complex to be analyzed by simple linear models, hence many companies use complex ML models for their decision-making process (Chen & Guestrin, 2016).

AI and ML models can analyze huge and complex data sets and show typically a higher predictive accuracy than linear models on these kinds of data sets. However, this improved predictive accuracy can only be achieved through a deterioration in explainability. The increasing complexity of ML models, in combination with the large amount of data processed by the models, reduces the capability to explain a model's decision. Hence, the predictions of a model are no longer comprehensible (Linardatos, Papastefanopoulos, & Kotsiantis, 2020).

In other words, there is a trade-off between the model performance and its explainability. Simple models, such as linear regression and logistic regression models, can satisfy the explainability condition. However, the predictive accuracy of these models is diminished when they are applied to large and complex data sets, especially with non-linear relationships. The use of sophisticated ML models, such as neural networks and random forests, provides a high predictive accuracy but it leads to limited explainability. For this reason, the literature also calls these complex ML models "black box" models (Molnar, 2020).

Although complex models show better model performance, it is difficult to trust them. Using black box models also includes practical and ethical issues, especially when they are used in crucial industries like healthcare or finance. These industries require trustworthy models and a basic requirement for trustworthiness is explainability. This requirement shows the importance of XAI methods (Gunning & Aha, 2019).

This doctoral thesis focuses on the financial industries and the potential applications of AI or ML models, enabled by the use of XAI methods. Credit risk management is one of the main application areas of ML models in the financial industries. It is a key banking area and addresses one of the material risks, financial institutions are facing. The risk arises primarily from potential credit defaults when a financial institution lends money to borrowers. Credit Risk Management includes the identification, measurement, and monitoring of these risks, an appropriate treatment, and the implementation of adequate risk models (BCBS, 2000).

To measure and manage the credit risk of a financial asset, banks use three key measures: 1) Probability of default (PD); 2) Loss given Default (LGD); and 3) Exposure at Default (EAD) (Altman & Heine, 2006). Credit risk, measured by the above-mentioned models is crucial to a bank and the stability of the banking system. Hence, the three risk models are strictly supervised by the European regulator and as such financial institutions have to fulfil strict requirements. For the European Union (EU) these requirements are defined inter alia in the Capital Requirements Regulation (CRR) (EC, 2013b) and the Capital Requirements Directive (CRD) IV (EC, 2013a). Similar requirements are defined by other regulators as well. Even though these requirements for credit risk management models are defined as technology-neutral, the regulator requires the financial institutions to approve the respective models internally. To be able to approve the models, the senior management needs to comprehensively understand the respective models. Hence, a lot of banks and financial institutions use simple models to measure credit risk, for example, logistic regression models.

The regulators identified that the risk associated with the increased use of models does not only apply in credit risk management but in general and responded with various regulations. Two of the most rele-

vant regulations are the US SR 11-7 (FED, 2011) and the European Central Bank (ECB) Guide to Internal Models (Consolidated version from 2019) (ECB, 2019).
Additionally, the regulators focus increasingly on the application of AI and ML models in financial institutions. This is reflected in various guidelines and recommendations, which have been published during the last years. In the following paragraph, some of the most important guidelines, reports, and regulations are presented that illustrate the importance of explainability for the application of AI and ML models in financial institutions.

In 2016 the European Parliament published the General Data Protection Regulation (GDPR), which states that data processing via automated decision-making should include information about the underlying logic (EC, 2016).

The Financial Stability Board highlights in their report from 2017 (FSB, 2017) the following areas regulators and supervision need to focus on:

- Governance and accountability
- Data quality and bias
- interpretability and transparency
- Risk management and validation
- Ethical considerations

The high-level expert group on AI from the European Commission published in 2019 the Ethics guidelines for trustworthy AI (EC, 2019a). They defined a framework for trustworthy AI including three pillars that require AI to be 1) compliant with existing laws and regulations, 2) it should be ethical, and 3) robust against technical and social influences. To realise a trustworthy AI system, they defined seven concrete requirements. One of these requirements is transparency. The component transparency includes the requirements that AI should be traceable, explainable, and it should communicate with the user. They further describe that the decision-making process of AI should be explainable whenever it affects people's lives and that this explanation should be adjusted according to the expertise of the respective stakeholder. The seven requirements to realise a trustworthy AI system are:

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental well-being
- Accountability

The European Banking Authority (EBA) Report on Big Data and Advanced Analytics from 2020 defined seven elements of trust in big data and advanced analytics (EBA, 2020). Again, explainability is one of the key elements in the context of the trustworthiness of AI models. The EBA defines in their report that explainable models have to enable humans to understand how the model generated the output and that the models need to be able to justify what the output is based on. These are the defined seven elements of trust in big data and advanced analytics:

- Ethics
- Explainability and interpretability
- Fairness and avoidance of bias
- Traceability and auditability (including versioning)
- Data protection and quality
- Security
- Consumer protection

The most recent regulation is the proposal of the EU Artificial Intelligence Act which categorizes some applications of AI and ML models in banks as high-risk applications that require extensive regulation. They explicitly define systems as high risk, which are used to estimate the creditworthiness of a person or estimate a credit score. The regulation states that the European Union aims for a solid European regulatory Framework for trustworthy AI. Hence, they defined that high-risk AI systems need to meet certain requirements, such as traceability and transparency (EC, 2021).

In 2023 the German Regulator Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin) published the updated Minimum Standards for Risk Management Mindestanforderungen an das Risikomanagement (MaRisk). The minimum standards explicitly state that in addition to the predictive accuracy, sufficient explainability must be ensured, particularly for the application of AI models (BaFin, 2023).

All these guidelines and regulations show the important role of explainability in the context of transparency and trustworthiness concerning the application of AI or ML models in the financial industries. The fact that expert groups, regulators, and supervisors have focused on the topic of explainability for many years emphasises that financial institutions must include the requirement of explainability in their model development process. The possibility of explaining the model needs to be as much prioritized as the performance of the model itself.

The important role of explainability leads to some fundamental questions, such as ”What is the correct definition of explainability” and ”When is a model actually explainable”?
Next to explainability, the term interpretability is often used synonymously. However, there is an ongoing discussion in the literature about the correct terminology. (Sokol & Flach, 2021) extensively discussed among others these two terminologies and their meaning within their paper. They show that the research in XAI and interpretable ML can be traced back to 1990 and if expert systems are considered as well, even to the 1970s (Rudin, 2019; Gregor & Benbasat, 1999; Leondes, 2001). They further elaborate that interpretability is a passive characteristic, whereas explainability is an active one. In other words, the degree of interpretability of a model describes the passive structure and the extent to which it enables an explainee

to understand the underlying structure of a model and the decisions based on it. The explainability of a model describes the active functions and methods implemented in a model in illustrating its decisions and predictions to an explainee.
Finally, Sokol and Flach decide to use the term explainability instead of interpretability and define it as "[...] insights that lead to understanding (the role of an explanation) [...]" (Sokol & Flach, 2021). They argue that this is a commonly used and frequently applied definition in the social sciences. Based on this definition, they defined the following equation:

$$Explainability = \underbrace{Reasoning\,(Transparency\,|\,Background\,Knowledge)}_{understanding} \tag{1}$$

This equation illustrates that explainability is a process of understanding through reasoning. This requires taking the transparency of a model and the background knowledge of the explainee into account.

Taking the explainee into the focus of the explanation of a model is a key element for the application of AI models in financial services. The explainee can vary between different stakeholders. (Bracke, Datta, Jung, & Shayak, 2019) defined at least six different stakeholders in the financial institutions:

- Model developers/implementers
  - The persons who develop and implement the ML model.
- First-line model checkers
  - The persons who ensure that the ML models have been developed sufficiently. This could be achieved via the implementation of Peer-to-peer (P2P) reviews between different model development teams.
- Responsible management
  - Many banks have the concept of a model owner. This person is ultimately responsible for the model.
- Second-line model checkers
  - The persons who independently validate a model to ensure the quality of development and employment.
- Conduct regulators
  - The regulators ensure that a model meets the respective conduct rules.
- Prudential regulators
  - The regulators ensure that a model meets the respective prudential requirements.

In addition to the listed explainees, one could add members of the board, corporate audit services, and customers. Some of the above-mentioned explainees have similar requirements regarding XAI methods and can be combined accordingly.

To address the requirements of different explainees it is important to highlight the already existing XAI methods. The methods can be categorized into different types, according to their respective characteristics. The following list provides a brief overview. A more complete overview and a detailed description of the respective example methods can be obtained via the following documents: (e.g. Molnar, 2020; Guidotti et al., 2018).

- White-box models vs. black-box models in combination with posthoc XAI methods
  - White-box models: This category includes all models that are interpretable due to their simple underlying structure, such as linear regression models or decision trees. These models do not need to be explained.
  - Black box models in combination with post-hoc XAI methods: This category includes all methods that are applied to already trained black box models (post-hoc). Post-hoc methods are for example Local Interpretable Model-agnostic Explanations (LIME) or the Shapley values approach.
- Model agnostic vs. model-specific
  - Model agnostic methods: These methods can be applied to all models independent of the models' structure. Examples are Partial Dependence Plots (PDP) or Shapley values approach.
  - Model-specific methods: These methods are only applicable to a specific category of models, like the analysis of the weights of regression coefficients in a linear regression model.
- Local vs. Global explanations
  - Local explanations: These methods explain individual predictions of a model. A famous example from credit risk management is the question of why a specific credit application has been rejected. Example methods are LIME or Counterfactual Explanations.
  - Global explanations: These methods explain the effect one feature has on the overall model predictions, e.g. PDP.
- Statistical methods vs. Visualisation methods
  - Statistical methods: These methods provide summary statistics regarding the variables, like feature importance values.
  - Visualisation methods: Many of the above-mentioned methods use this method to explain a model. Visualisations can be PDP, heatmaps, or diagrams.

The above-described model agnostic post-hoc XAI methods do not satisfy the requirements of all explainees in the financial industries. Explainees, like the members of the board, need to get a sufficient understanding of a model quickly to decide without being able to familiarise themselves intensively with the respective topic. Other explainees like corporate audit services need to understand every detail of a model and compare it with other models. Hence, the results of a model need to be clearly and comprehensively explained and the explainability of different models needs to be comparable with each other. Therefore, this doctoral thesis aims to create methods that can address many different explainees, including the above-mentioned, by extending existing XAI methods and developing new methods. The contributions to the existing literature are presented in three self-contained chapters.

Chapter 1 investigates how ML models in combination with XAI methods can be applied in credit risk management to estimate the PD of a potential customer. The paper on which this chapter is based was published in the Springer Journal of Computational Economics (2020). In this chapter, an existing XAI method is extended by applying a Correlation Network to Shapley Values. The output of the model is made explainable by the Shapley Values and is further simplified by the visual representation as a network and by the clustering of the underlying Shapley Values. This improves the degree of explainability of the model even more.

For this purpose, a data set was analysed that included an evaluation of various financial characteristics of 15,000 small and medium-sized enterprises that requested a loan on a P2P-platform. The analysis was based on the PD of the customers. A logistic regression model and an XGBoost model were trained on a training dataset. The model performance of the two models was compared using the area under the curve Area Under the Receiver Operating Characteristic curve (AUROC). The comparison shows that the ML model has a better model performance and was able to predict the default probability of the customers more accurately.
To explain the XGBoost model, we first calculated the Shapley values of the predictions. Since calculating the Shapley values for a large number of variables requires high computational power, we used the TreeSHAP method by (Lundberg et al., 2020). To further increase the explanatory power of the Shapley Value approach and to visualize the structure within the Shapley Values, we used a MST (a single linkage cluster). Within the graph, we coloured individual observations to distinguish between the defaulted loans and the non-defaulted ones.

In chapter 2 the Shapley value approach described in chapter 1, is further developed by combining it with the Lorenz Zonoid approach to obtain a new XAI method, the Shapley Lorenz value approach. The paper on which this chapter is based on has been published in the Classification and Data Analysis Group 2021 post-proceedings Springer collection. Shapley values are not normalised and therefore difficult to understand. This problem has made it difficult to use Shapley values in economics so far.
The Shapley-Lorenz values provide a normalised value between 0 and 1 for every feature of a model. This enables stakeholders to compare models and the respective drivers of the models' predictions with each other based on the explainability. The novel method is used to analyse the same dataset as in chapter 1. We applied a logistic regression model to the data, as it is easy to interpret and allows us to better evaluate the new explainability method.
To analyse the results we compared the Shapley-Lorenz values with the Shapley values and with the contribution of each variable to the deviance G^2. The Shapley-Lorenz values were the easiest to interpret. Additionally, this approach is more robust to outlier observations.

Chapter 3 addresses the topic of sustainability, which is currently a highly relevant topic in the financial industry. The paper on which this chapter is based on has been published on SSRN. Environmental, Social, and Governance (ESG) topics have become increasingly important in the last years. In this chapter, the important ESG variables to estimate the cost of capital were identified. We trained a ML model on financial and non-financial factors to predict the cost of capital, which in this context, reflects the investor's perceptions of a firm's risk profile.
Our data set contains more than 1400 companies, listed on global stock markets. We applied two XAI Methods to the results of the ML model, the Shapley Values approach and the Lorenz Zonoid approach. Our model is the well-known eXtreme Gradient Boosting ML model, XGBoost. Based on the Shapley Values, we are able to prioritize the variables regarding their importance. We used the Lorenz Zonoid

approach, which we used in chapter 2 as well to further develop the Shapley values approach, but this time we use it to identify pivotal factors influencing the implied cost of capital.

The results showed that financial and country characteristics are most important for the model's predictions, but it also showed that non-financial factors contribute to the prediction of the cost of capital, especially environmental and governance-related factors. The results indicate that investors penalize the most polluting companies, highlighting the necessity of fostering the shift to a more sustainable economic structure.

1 Explainable Machine Learning in Credit Risk Management

> *What we learn from history is that people don't learn from history.*
>
> – Warren Buffett

1.1 Introduction

Black box AI is not suitable for regulated financial services. To overcome this problem, Explainable AI models, which provide details or reasons to make the functioning of AI clear or easy to understand, are necessary.

To develop such models, we first need to understand what ”Explainable” means. Recently, some important institutional definitions have been provided. For example, (Bracke et al., 2019) states that ”Explanations can answer different kinds of questions about a model's operation depending on the stakeholder they are addressed to” and (Croxson, Bracke, & Jung, 2019) ”interpretability will be the focus of explainability, generally taken to mean that an interested stakeholder can comprehend the main drivers of a model-driven decision”.

Explainability means that an interested stakeholder can comprehend the main drivers of a model-driven decision; (FSB, 2017) suggests that "lack of interpretability and auditability of AI and ML methods could become a macro-level risk"; (Croxson et al., 2019) establishes that "in some cases, the law itself may dictate a degree of explainability."

The European GDPR (EC, 2016) regulation states that "the existence of automated decision-making should carry meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject." Under the GDPR regulation, the data subject is therefore, under certain circumstances, entitled to receive meaningful information about the logic of automated decision-making.

Finally, the European Commission High-Level Expert Group on AI presented the Ethics Guidelines for Trustworthy Artificial Intelligence in April 2019. Such guidelines put forward a set of seven key requirements that AI systems should meet in order to be deemed trustworthy. Among them three relate to the concept of ”eXplainable Artificial Intelligence (XAI)” , and are the following.

- Human agency and oversight: decisions must be informed, and there must be a human-in-the-loop oversight.

- Transparency: AI systems and their decisions should be explained in a manner adapted to the concerned stakeholder. Humans need to be aware that they are interacting with an AI system.

- Accountability: AI systems should develop mechanisms for responsibility and accountability, auditability, assessment of algorithms, data and design processes.

Following the need to explain AI models, stated by legislators and regulators of different countries, many established and startup companies have started to embrace Explainable AI models. In addition, more and more people are searching for information about what ”Explainable Artificial Intelligence” means.

In this respect, (**Figure 1.1**) represents the evolution of Google searches for explainable AI related terms.

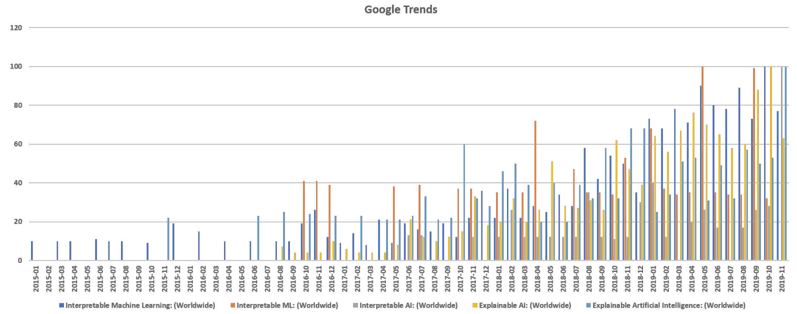

Figure 1.1: Google trend searches for explainable AI related terms.

From a mathematical viewpoint, it is well known that “simple” statistical learning models, such as linear and logistic regression models, provide a high interpretability but, possibly, limited predictive accuracy. On the other hand, ”complex” ML models, such as neural networks and tree models, provide high predictive accuracy at the expense of limited interpretability.

To solve this trade-off, we propose to boost ML models, that are highly accurate, with a novel methodology, that can explain their predictive output. Our proposed methodology acts in the post-processing phase of the analysis, rather than in the preprocessing part. It is agnostic (technologically neutral) as it is applied to the predictive output, regardless of which model generated it: a linear regression, a classification tree or a neural network model.

The ML procedure proposed in the chapter processes the outcomes of any other arbitrary ML model. It provides more insight, control, and transparency to a trained, potentially black box ML model. It utilises a model-agnostic method aiming at identifying the decision-making criteria of an AI system in the form of variable importance (individual input variable contributions).

A key concept of our model is the Shapley value decomposition of a model, a pay-off concept from cooperative game theory. To the best of our knowledge, this is the only explainable AI approach rooted in an economic foundation. It offers a breakdown of variable contributions so that every data point (e.g. a credit or loan customer in a portfolio) is not only represented by input features (the input of the ML model) but also by variable contributions to the prediction of the trained ML model.

More precisely, our proposed methodology is based on the combination of network analysis with Shapley values (see (Lundberg & Lee, 2017), (Joseph, 2019b), and references therein). Shapley values were originally introduced by (L. Shapley, 1953) as a solution concept in cooperative game theory. They correspond to the average of the marginal contributions of the players associated with all their possible orders. The advantage of Shapley values, over alternative XAI models, is that they can be exploited to measure the contribution of each explanatory variable for each point prediction of a ML model, regardless of the underlying model itself (see, e.g.(Lundberg & Lee, 2017)). In other words, Shapley-based XAI models combine the generality of application (they are model agnostic) with the personalisation of their results (they can explain any single-point prediction).

Our original contribution is to improve Shapley values, improving the interpretation of the predictive output of a ML model using correlation network models. To exemplify our proposal, we consider one area of the financial industry in which Artificial Intelligence methods are increasingly being applied: credit risk management (see for instance the review by (Giudici, 2018)).

Correlation networks, also known as similarity networks, have been introduced by (Mantegna & Stanley, 1999) to show how time series of asset prices can be clustered in groups based on their correlation matrix. Correlation patterns between companies can similarly be extracted from cross-sectional features, based on balance sheet data, and they can be used in credit risk modeling. To account for such similarities we can rely on centrality measures, following (Giudici, Hadji-Misheva, & Spelta, 2019a) and (Giudici, Hadji-Misheva, & Spelta, 2019b), who have shown that the inclusion of centrality measures in credit scoring models does improve their predictive utility. Here we propose a different use of similarity networks. Instead of applying network models in a pre-processing phase, as in (Giudici et al., 2019a) and (Giudici et al., 2019b), which extract from them additional features to be included in a statistical learning model, we use them in a post-processing phase, to interpret the predictive output from a highly performing ML model. In this way, we achieve both predictive accuracy and explainability.

We apply our proposed method to predict the credit risk of a large sample of small and medium enterprises. The obtained empirical evidence shows that, while improving the predictive accuracy concerning a standard logistic regression model, we improve, the interpretability (explainability) of the results.

The rest of the chapter is organized as follows: Section 1.2 introduces the proposed methodology. Section 1.3 shows the results of the analysis in the credit risk context. Section 1.4 concludes and presents possible future research developments.

1.2 Methodology

1.2.1 Statistical Learning of Credit Risk

Credit risk models are usually employed to estimate the expected financial loss that a credit institution (such as a bank or a P2P lender) suffers if a borrower defaults to pay back a loan. The most important component of a credit risk model is the PD, which is usually estimated statistically by employing credit scoring models.

Borrowers could be individuals, companies, or other credit institutions. Here we focus, without loss of generality, on small and medium enterprises, whose financial data are publicly available in the form of yearly balance sheets.

For each company, n, define a response variable Y_n to indicate whether it has defaulted on its loans or not, i.e. $Y_n = 1$ if the company defaults, $Y_n = 0$ otherwise. And let X_n indicate a vector of explanatory variables. Credit scoring models assume that the response variable Y_n may be affected ("caused") by the explanatory variables X_n.

The most commonly employed model of credit scoring is the logistic regression model. It assumes that

$$\ln(\frac{P_n}{1 - P_n}) = \alpha + \sum_{j=1}^{J} \beta_j x_{nj}$$

where p_n is the PD for company n; $\mathbf{x}_n = (x_{i,1}, ..., x_{i,J})$ is a J-dimensional vector containing the values that the J explanatory variables assume for company n; the parameter α represents an intercept; β_j is the j-th regression coefficient.

Once the parameters α and β_j are estimated using the available data, the PD can be estimated, inverting the logistic regression model, from:

$$p_n = (1 + exp(\alpha + \sum_{j=1}^{J} \beta_j x_{nj}))^{-1} \tag{1.1}$$

1.2.2 Machine learning of credit risk

Alternatively, credit risk can be measured with ML models, able to extract non-linear relations among the financial information contained in the balance sheets. In a standard data science life cycle, models are chosen to optimise predictive accuracy. In highly regulated sectors, like finance or medicine, models should be chosen to balance accuracy with explainability (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019). We improve the choice-selecting models based on their predictive accuracy and employ a posteriori algorithm that achieves explanability. This does not limit the choice of the best-performing models.

To exemplify our approach we consider, without loss of generality, the Extreme Gradient Boost model, one of the most popular and fast ML algorithms (e.g. Chen & Guestrin, 2016).

XGBoost is a supervised model based on the combination of tree models with Gradient Boosting. Gradient Boosting is an optimisation technique able to support different learning tasks, such as classification, ranking, and prediction. A tree model is a supervised classification model that searches for the partition of the explanatory variables that best classify a response (supervisor) variable. Extreme Gradient Boosting improves tree models strengthening their classification performance, as shown by (Chen & Guestrin, 2016). The same authors also show that XGBoost is faster than tree model algorithms.

In practice, a tree classification algorithm is applied successively to ”training” samples of the data set. In each iteration, a sample of observations is drawn from the available data, using sampling weights that change over time, weighting the observations with the worst fit. Once a sequence of trees is fit, and classifications made, a weighted majority vote is taken. For a more detailed description of the algorithm see, for instance, (Friedman, Hastie, & Tibshirani, 2000).

1.2.3 Learning model comparison

Once a default probability estimation model is chosen, it should be measured in terms of predictive accuracy, and compared with other models, so to select the best one. The most common approach to measure the predictive accuracy of credit scoring models is to randomly split the available data into two parts: a ”train” and a ”test” set; build the model using data in the train set, and compare the predictions the model obtains on the test set, $\hat{Y_n}$, with the actual values of Y_n.

To obtain $\hat{Y_n}$ the estimated default probability is rounded into a ”default” or ”non-default”, depending on whether a threshold is passed or not. For a given threshold T, one can then count the frequency of the four possible outputs, namely: False Positives (FP): companies predicted to default, that do not; True Positives (TP): companies predicted to default, which do; False Negatives (FN): companies predicted not to default, which do; True Negatives (TN): companies predicted not to default, which do not.

The misclassification rate of a model can be computed as:

$$\frac{FP + FN}{TP + TN + FP + FN} \tag{1.2}$$

and it characterizes the proportion of wrong predictions among the total number of cases.

The misclassification rate depends on the chosen threshold and it is not, therefore, a generally agreed measure of predictive accuracy. A common practice is to use the Receiver Operating Characteristics

(ROC) curve, which plots the False Positive Rate (FPR) on the Y axis against the True Positive Rate (TPR) on the X axis, for a range of threshold values (usually percentile values). FPR and TPR are then calculated as follows:

$$FPR = \frac{FP}{FP + TN} \tag{1.3}$$

$$TPR = \frac{TP}{TP + FN} \tag{1.4}$$

The ideal ROC curve coincides with the Y axis, a situation which cannot be realistically achieved. The best model will be the one closest to it. The ROC curve is usually summarised with the Area Under the ROC curve value AUROC, a number between 0 and 1. The higher the AUROC, the better the model.

1.2.4 Explaining model predictions

We now explain how to exploit the information contained in the explanatory variables to localise and cluster the position of each individual (company) in the sample. This information, coupled with the predicted default probabilities, allows a very insightful explanation of the determinant of each individual's creditworthiness. In our specific context, information on the explanatory variables is derived from the financial statements of borrowing companies, collected in a vector $\mathbf{x}_n$, representing the financial composition of the balance sheet of institution n.

We propose to calculate the Shapley value associated with each company. In this way, we provide an agnostic tool that can interpret in a technologically neutral way the output from a highly accurate ML model. As suggested in (Joseph, 2019b), the Shapley values of a model can be used as a tool to transfer predictive inferences into a linear space, opening a wide possibility of applying them to a variety of multivariate statistical methods.

We develop our Shapley approach using the SHAP (Lundberg & Lee, 2017) computational framework, which allows us to compute Shapley values expressing predictions as linear combinations of binary variables that describe whether each single variable is included or not in the model.

More formally, the explanation model $g(x')$ for the prediction $f(x)$ is constructed by an additive feature attribution method, which decomposes the prediction into a linear function of the binary variables $z' \in \{0,1\}^M$ and the quantities $\phi_i \in \mathbb{R}$:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i. \tag{1.5}$$

In other terms, $g'(z') \approx f(h_x(z'))$ is a local approximation of the predictions where the local function $h_x(x') = x$ maps the simplified variables x' into x, $z' \approx x$ and M is the number of the selected input variables.

Indeed, (Lundberg & Lee, 2017) prove that the only additive feature attribution method that satisfies the properties of *local accuracy*, *missingness* and *consistency* is obtained attributing to each feature x'_i an effect ϕ_i called Shapley value, defined as

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \setminus i) \right] \tag{1.6}$$

where f is the trained model, x the vector of inputs (features), x' the vector of the M selected input features. The quantity $f_x(z') - f_x(z' \setminus i)$ is the contribution of a variable i and expresses, for each single prediction, the deviation of Shapley values from their mean.

In other words, a Shapley value represents a unique quantity able to construct an explanatory model that locally linearly approximates the original model, for a specific input x,(*local accuracy*). With the property that, whenever a feature is locally zero, the Shapley value is zero (*missingness*) and if in a second model the contribution of a feature is higher, so will be its Shapley value (*consistency*).

Once Shapley values are calculated, we propose to employ similarity networks, defining a metric that provides the relative distance between companies by applying the Euclidean distance between each pair $(\mathbf{x}_i, \mathbf{x}_j)$ of company predicted vectors, as in (Giudici et al., 2019a) and (Giudici et al., 2019b).

We then derive the MST representation of the companies, employing the correlation network method suggested by (Mantegna & Stanley, 1999)). The MST is a tree without cycles of a complex network, that joins pairs of vertices with the minimum total ”distance”.

The choice is motivated by the consideration that, to represent all pairwise correlations between N companies in a graph, we need $N * (N-1)/2$ edges, a number that quickly grows, making the corresponding graph not understandable. The MST simplifies the graph into a tree of $N-1$ edges, which takes $N-1$ steps to be completed. At each step, it joins the two closest companies, in terms of the Euclidean distance between the corresponding explanatory variables.

In our Shapley value context, the similarity of variable contributions is expressed as a symmetric matrix of dimension n x n, where n Is the number of data points in the (train) data set. Each entry of the matrix measures how similar or distant a pair of data points is in terms of variable contributions. The MST representation associates to each point its closest neighbour. To generate the MST we have used the EMST Dual-Tree Boruvka algorithm, and its implementation in the R package “emstreeR”.

The same matrix can also be used, in a second step, for a further merging of the nodes, through cluster analysis. This extra step can reveal segmentations of data points with very similar variable contributions, corresponding to similar credit-scoring decision-making.

1.3 Application

1.3.1 Data

We test our proposed model to data supplied by the European External Credit Assessment Institution (ECAI) which specializes in credit scoring for P2P platforms focused on Small and medium-sized Enterprises (SME) commercial lending. The data is described by (Giudici et al., 2019a) to which we refer for further details. In summary, the analysis relies on a dataset composed of official financial information (balance-sheet variables) on 15,045 SMEs, mostly based in Southern Europe, for the year 2015. The information about the status (0 = active, 1 = defaulted) of each company one year later (2016) is also provided. The proportion of defaulted companies within this dataset is 10.9%.

Using this data, (Giudici et al., 2019a) and (Giudici et al., 2019b) have constructed logistic regression scoring models that aim at estimating the PD of each company, using the available financial data from the balance sheets and, in addition, network centrality measures that are obtained from similarity networks.

Here we aim to improve the predictive performance of the model and, for this purpose, we run an XGBoost tree algorithm (e.g. Chen & Guestrin, 2016). To explain the results from the model, typically highly predictive, we employ similarity network models, in a post-processing step. In particular, we employ the cluster dendrogram representation that corresponds to the application of the MST algorithm.

1.3.2 Results

We first split the data into a training set (80%) and a test set (20%), using random sampling without replacement.

We then estimate the XGBoost model on the training set, apply the obtained model to the test set, and compare it with the best logistic regression model. The ROC curves of the two models are contained in (**Figure 1.2**) below.

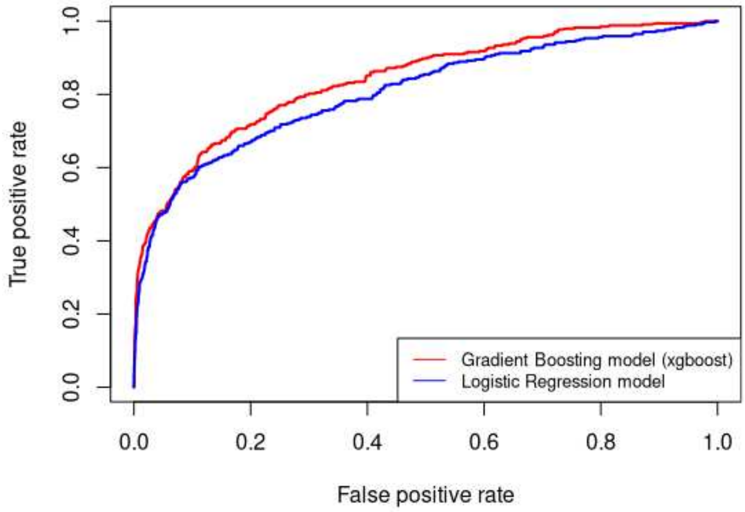

Figure 1.2: Receiver Operating Characteristic (ROC) curves for the logistic credit risk model and for the XGBoost model. In blue, we show the results related to the logistic models while in red we show the results related to the XGBoost model.

From (**Figure 1.2**) note that the XGBoost clearly improves predictive accuracy. Indeed the comparison of the Area Under the ROC curve AUROC for the two models indicates an increase from 0.81 (best logistic regression model) to 0.93 (best XGBoost model).

We then calculate the Shapley values of the companies in the test set, using the values of their explanatory variables. The MST (a single linkage cluster) is used to simplify and interpret the structure present among Shapley values. We can also ”colour” the MST graph in terms of the associated response variables values: default, not default.

(**Figure 1.3**) and (**Figure 1.4**) present the MST representation. While in (**Figure 1.3**) company nodes are coloured according to the cluster to which they belong, in (**Figure 1.4**) they are coloured according to their status: not defaulted (grey); defaulted (red).

In (**Figure 1.3**), nodes are coloured according to the cluster in which they are classified. The figure shows that clusters are quite scattered along the correlation network.

To construct the coloured communities in (**Figure 1.3**), we used the algorithm implemented in the R package “igraph” that directly optimizes a modularity score. The algorithm is very efficient and easily scales to very large networks.

In (**Figure 1.4**), nodes are coloured in a simpler binary way: whether the corresponding company has defaulted or not.

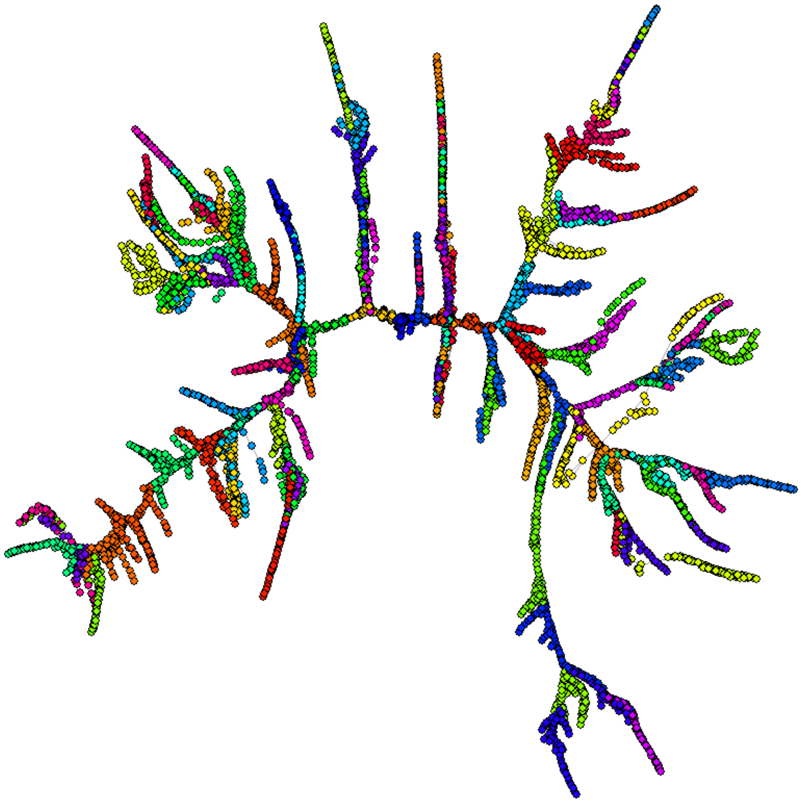

Figure 1.3: MST representation of the borrowing companies. Clustering has been performed using the standardized Euclidean distance between institutions. Companies are coloured according to their cluster of belonging.

From (**Figure 1.4**) note that default nodes appear grouped in the MST representation, particularly along the bottom left branch. In general, defaulted institutions occupy a precise portion of the network, usually to the leaves of the tree, and form clusters. This suggests that those companies form communities, characterised by similar predictor variables' importance. It also suggests that not defaulted companies that are close to default ones have a high risk of becoming defaulted as well, being the importance of their predictor variables very similar to those of the defaulted companies.

To better explain the explainability of our results, in (**Figure 1.5**) we provide the interpretation of the estimated credit scoring of four companies: two that defaulted and two that did not.

(**Figure 1.5**) clearly shows the advantage of our explainable model. It can indicate which variables contribute more to the prediction of default. Not only in general, as is typically done by statistical and ML models, but differently and specifically for each company in the test set. Indeed, (**Figure 1.5**) clearly shows how the explanations are different ("personalised") for each of the four considered companies.

The most important variables, for the two non-defaulted companies (left boxes) regard: profits before taxes plus interests paid, and Earnings before interest, taxes, depreciation and amortization (EBITDA), which are common to both; trade receivables, for company 1; total assets, for company 2.

Economically, a high proficiency decreases the PD, for both companies; whereas a high stock of outstanding invoices, not yet paid, or a large stock of assets, helps reduce the same probability.

On the other hand, (**Figure 1.5**) shows that the most important variables, for the two defaulted companies (right boxes) concern: total assets, for both companies; shareholder's funds plus non-current liabilities, for company 3; profits before taxes plus interests paid, for company 4.

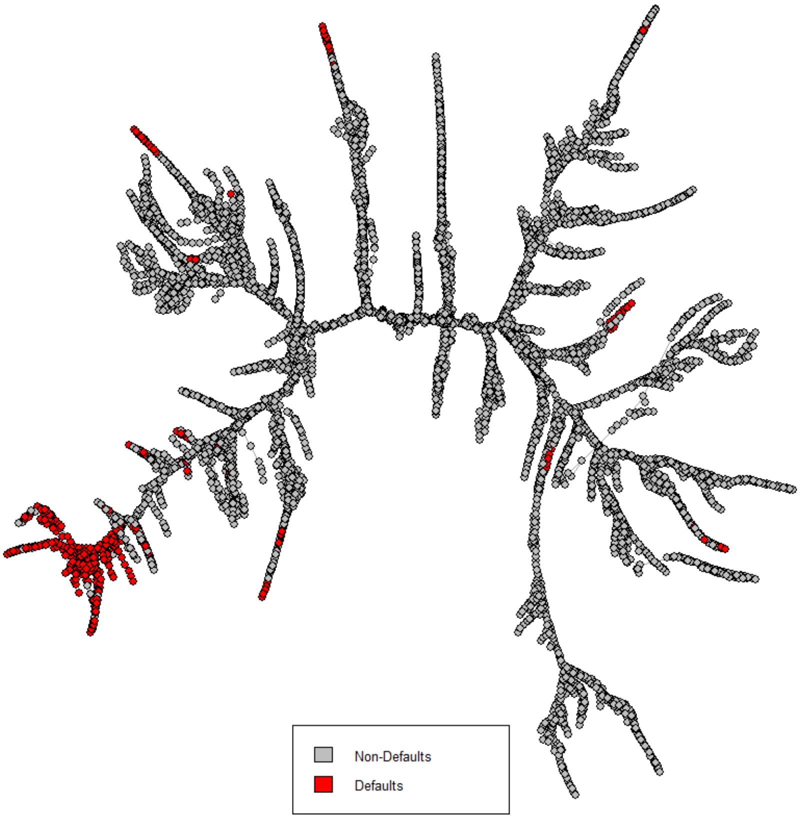

Figure 1.4: MST representation of the borrowing companies. Clustering has been performed using the standardized Euclidean distance between institutions. Companies are coloured according to their default status: red= defaulted; grey= not defaulted.

In other words, lower total assets coupled, in one case, with limited shareholder funds and, in the other, with low proficiency, increase the PD of these two companies.

The above results are consistent with previous analysis of the same data: both (Giudici et al., 2019a) and (Giudici et al., 2019b) select, as the most important variables in several models, the return on equity, related to both EBITDA and profit before taxes plus interests paid; the leverage, related to total assets and shareholders' funds; and the solvency ratio, related to trade payables.

We remark that (**Figure 1.5**) contains a "local" explanation of the predictive power of the explanatory variables, and it is the most important contribution of Shapley value theory. If we average Shapley values across all observations we get an "overall" or "global" explanation, similar to what is already available in the statistical and ML literature. (**Figure 1.6**) below is the global explanation in our context: the ten most important explanatory variables, over the whole sample.

From (**Figure 1.6**) note that total assets to total liabilities (the leverage) are the most important variable, followed by the EBITDA, along with profit before taxes plus interest paid, measures of operational efficiency; and by trade receivables, related to solvency, in line with the previous comments.

1.4 Conclusions and future research

The need to leverage the high predictive accuracy brought by sophisticated ML models, making them interpretable, has motivated us to introduce an agnostic, post-processing methodology, based on correla-

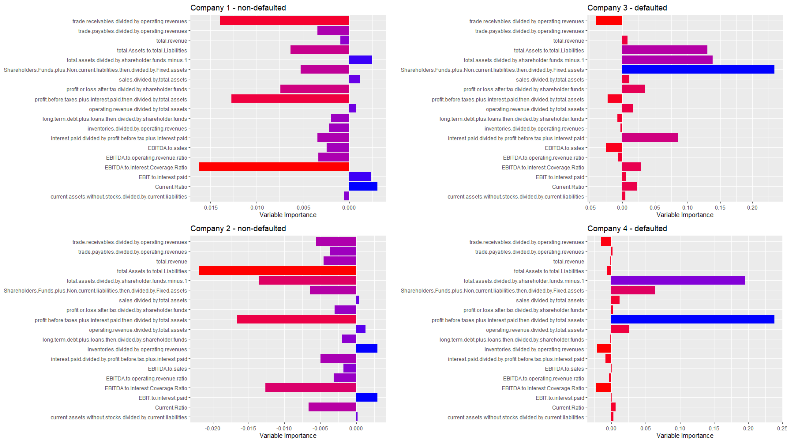

Figure 1.5: Contribution of each explanatory variable to the Shapley's decomposition of four predicted default probabilities, for two defaulted and two non-defaulted companies. The more red the color the higher the negative importance, and the more blue the color the higher the positive importance.

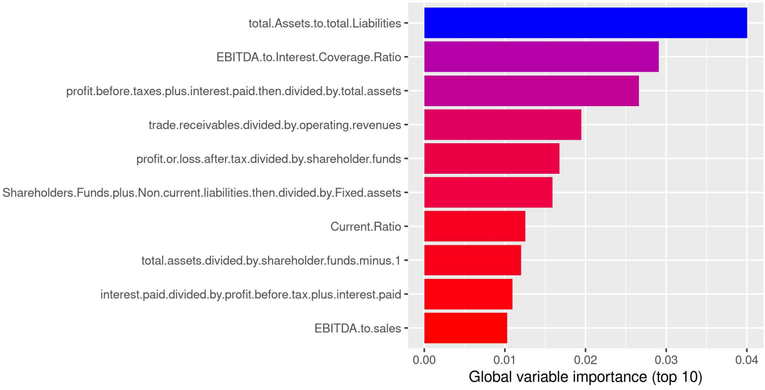

Figure 1.6: Mean contribution of each explanatory variable to Shapley's decomposition. The redder the colour the higher the negative importance, and the bluer the color the higher the positive importance.

tion network models. The model can explain, from a substantial viewpoint, any single prediction in terms of the Shapley value contribution of each explanatory variable.

For the implementation of our model, we have used TreeSHAP, a consistent and accurate method, available in open-source packages. TreeSHAP is a fast algorithm that can compute SHAP values for trees in polynomial time instead of the classical exponential runtime. For the XGBoost part of our model we have used NVIDIA Graphics Processing Unit (GPU)s to considerably speed up the computations. In this way, the TreeSHAP method can quickly extract the information from the XGBoost model.

Our research has important policy implications for policymakers and regulators who are in their attempt to protect the consumers of artificial intelligence services. While artificial intelligence effectively improves the convenience and accessibility of financial services, it also triggers new risks. Our research suggests that network-based explainable AI models can effectively advance the understanding of the determinants of financial risks and, specifically, of credit risks. The same models can be applied to forecast the PD, which is critical for risk monitoring and prevention.

Future research should extend the proposed methodology to other datasets and, in particular, to imbalanced ones, for which the occurrence of defaults tends to be rare, even more than what is observed for the analysed data. The presence of rare events may inflate the predictive accuracy of such events (as shown in (Bracke et al., 2019)). Indeed, (Thomas, Edelman, & Crook, 1997) suggests dealing with this problem via oversampling and it would be interesting to see what this implies in the proposed correlation network Shapley value context.

Finding comprehensive data sets which contain information about credit risk is difficult, especially those which include information about credit defaults. There are multiple reasons for this but the main ones are privacy and regulatory concerns. Credit risk data contains sensitive information that allows drawback conclusions about the respective customer. Hence, this data is subject to strict regulatory requirements, such as the European GDPR. Sharing this data and making it publicly available could lead to serious issues for the respective institution (EC, 2016). Accordingly, we were not able to test the approach on other strongly imbalanced data sets.

The credit default rate of 10.9 % of the applied data set shows that the amount of credit defaults which the model should estimate is scarce and the credit risk data is imbalanced. However, compared to the default rate of other data sets, it is still comparatively high. Other data sets can have imbalances in the form of 1,000:1 or even 10,000:1. Working with these imbalanced data sets describes a well-known issue of model development in the financial sector. Imbalanced data sets lead to difficulties in the training process of models such as biased learning, which means that the model is trained on the majority class (no credit default) and is less sensitive against the minority class (credit default). Since the model is mainly learning from cases of the majority class, it is overfitting on this class and hence captures noise and outliers instead of identifying the desired patterns in the data (H. He & Garcia, 2009). The above-described difficulties can also be transferred to the MST. The risk of biased learning could impact the MST in the form of a biased structure of the tree. This means the MST structure mainly reflects connections within the majority class and has fewer connections within the minority class. This could lead to a reduction of the explanatory power of the MST. Again noise and outliers could have a significant impact on the structure of the MST and would lead to longer or less meaningful branches within the tree (H. He & Garcia, 2009).

2 Shapley Lorenz values for credit risk management

Fear of the unknown is the greatest fear of all.

– Yvon Chouinard

2.1 Introduction

A key point in the application of Artificial Intelligence methods is risk measurement. When applied to regulated industries, such as energy, finance, and health, artificial intelligence methods lack explainability, and, therefore, authorities aimed at monitoring the risks arising from their application may not validate them. The interpretability requirement is strong, especially in regulated industries, such as banking, finance, and insurance, where data have to be exploited in order to draw conclusions from them and predict future trends (e.g. FSB, 2017; Joseph, 2019a). In these fields, comprehensible results need to be obtained to allow organizations to detect risks, especially in terms of the factors that can cause them. This objective is more evident when dealing with AI systems, which have a black-box nature resulting in automated decision-making and can classify a user into a class associated with the prediction of the individual behaviour, without explaining the underlying rationale. In order to avoid wrong actions being taken as a consequence of ”automatic” choice, AI methods have to be as much as possible explainable.

To develop explainable AI methods, the notion of ”explainability” has to be clarified. Some relevant institutional definitions have been recently provided. (Bracke et al., 2019), for instance, states that ”Explainability means that an interested stakeholder can comprehend the main drivers of a model-driven decision”, meaning that, to be explainable, AI methods have to provide details or reasons clarifying their functioning.

From a mathematical viewpoint, the requirement of high explainability can be fulfilled by resorting to simple ML models, (such as, e.g., logistic and linear regression models). Nevertheless, these models provide a reduced predictive accuracy. To improve predictive accuracy, the implementation of complex ML models (such as neural networks or random forests) seems necessary but this leads to a limited interpretability. This trade-off can be solved by boosting highly accurate ML models with innovative methodologies able to explain the corresponding predictive output. A recent attempt in this direction can be found in (Bussmann, Giudici, Marinelli, & Papenbrock, 2021), who proposed to apply correlation networks (e.g. Mantegna & Stanley, 1999) to Shapley values (e.g. L. Shapley, 1953) so that Artificial Intelligence predictions are grouped according to the similarity in the underlying explanations. The proposal was validated in the area of credit lending, in which the use of AI methods for credit risk measurement is developing fast, to detect the variables that mostly contribute to the prediction of default.

In this chapter, we propose an explainable ML model aimed at accurately measuring credit risks. To achieve this goal we develop a methodology based on the combination of the Shapley value approach and the Lorenz Zonoid tool, described in (Giudici & Raffinetti, 2021). Shapley values belong to the class of local explanation approaches since they can be exploited to interpret individual predictions (e.g. Molnar, 2020; Joseph, 2019a), at the single unit level. Lorenz Zonoids can be used to describe a model as a whole, in terms of which explanatory variables most determine its predictions, for all units. We propose to extend the Shapley value game theoretic approach to the Lorenz Zonoid framework, recently proposed by (Giudici & Raffinetti, 2020). This leads to a new class of global explanation approaches: the Shapley-Lorenz approach.

The chapter is organized as follows: the next section illustrates the methodology; Section 2.3 discusses the empirical findings obtained by applying our proposal to real credit lending data; finally Section 2.4 contains some concluding remarks.

2.2 Methodology

To meet the requirement that risk measurement is explainable, leading To develop a trustworthy application of AI in credit lending, in this Section, we propose an explainable ML method to measure credit risk. Our proposal derives from the combination of two research streams. The first one concerns the development of risk management models to analyse credit lending data. The second concerns the development of explainable methods to understand the results of advanced ML models. The result is a novel method for credit risk management that is, at the same time, predictively accurate, interpretable, and robust.

2.2.1 Modeling default

Let Y be a response binary variable, which expresses whether a company defaults ($Y = 1$) or not ($Y = 0$). Given K explanatory variables $X_1, \dots, X_K$, a logistic regression model for Y can be specified as follows:

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} = \eta_i, \tag{2.1}$$

where $\eta_i = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik}$; π_i represents the PD for the i-th company; $\mathbf{x}_i = (x_{i1}, \dots, x_{ik})$ points out the K-dimensional vector reporting the values taken by the K explanatory variables referred to the i-th company; β_0 and β_k are the parameters representing the intercept and the k-th regression coefficient, respectively.

By means of the Maximum Likelihood Estimation method, the parameters β_0 and β_k can be estimated leading to derive the predicted PD as:

$$\hat{\pi}_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}. \tag{2.2}$$

2.2.2 The Shapley-Lorenz decomposition for credit risk data

To meet the conditions of predictive accuracy and interpretability, (Giudici & Raffinetti, 2021) have proposed a global explainable AI model, named Shapley-Lorenz decomposition, which combines the interpretability power of the local Shapley value game theoretic approach (e.g. L. Shapley, 1953) with a more robust global approach based on the Lorenz Zonoid model accuracy tool (e.g. Giudici & Raffinetti, 2020). The Lorenz Zonoids, originally introduced by (G. A. Koshevoy & Mosler, 1996), were further developed by (Giudici & Raffinetti, 2020) as a generalisation of the ROC curve in a multidimensional setting and, therefore, the Shapley-Lorenz decomposition has the advantage of combining predictive accuracy and explainability performance into one single diagnostics. Furthermore, the Lorenz Zonoid is based on a measure of mutual variability that is more robust to the presence of outlying observations with respect to the standard variability around the mean. These theoretical properties can be exploited to develop partial dependence measures that allow the detection of the additional contribution of a new predictor into an existing model.

Shapley values were originally introduced by (L. Shapley, 1953) as a pay-off concept from cooperative game theory. When referring to ML models, the notion of pay-off corresponds to the model prediction. Thus, for any single statistical unit i ($1 = 1, \dots, n$), the pay-offs are computed as

$$p_{off}(X_i^k) = \hat{f}(X^{'} \cup X_k)_i - \hat{f}(X^{'})_i, \tag{2.3}$$

where $\hat{f}(X^{'})_i$ are the predicted values provided by a ML model, depending only on $X^{'}$ predictors; $\hat{f}(X^{'} \cup X_k)_i$ are the predicted values generated by the ML model, depending both on the $|X^{'}|$ predictors and the additional included X_k predictor.

The main advantage of Shapley values, over alternative Explainable AI models, is that they can be exploited to measure the contribution of each explanatory variable for each point prediction of a ML model, regardless of the underlying model itself (e.g. Lundberg & Lee, 2017; Strumbelj & Kononenko, 2010). In other words, Shapley-based XAI models combine the generality of application (they are model agnostic) with the personalisation of their results (they can explain any single point prediction).

The main drawback of Shapley values is that they provide explainability scores that are not normalised. They can be used to compare the relative contribution of one variable to that of another, but they cannot be used to assess the absolute importance of each variable.

The key benefit related to the employment of the Lorenz Zonoid tool is the possibility of evaluating the contribution associated with any additional explanatory variable to the whole model prediction with a normalised measure that can be used to assess the importance of each variable.

The Lorenz Zonoid measure was introduced by (Giudici & Raffinetti, 2020) to develop new partial dependence measures. Specifically, given a set of K explanatory variables, let $\hat{Y}_{X' \cup X_k}$ and $\hat{Y}_{X'}$ be the predicted values provided by a model, including also covariate X_k, and the predicted values provided by a reduced model, excluding covariate X_k, respectively. The additional contribution related to the inclusion of covariate X_k can be determined in terms of the Partial Gini Contribution measure as follows:

$$PGC_{Y,X_k|X'} = \frac{LZ(\hat{Y}_{X' \cup X_k}) - LZ(\hat{Y}_{X'})}{LZ(Y) - LZ(\hat{Y}_{X'})}, \tag{2.4}$$

where $LZ(\hat{Y}_{X' \cup X_k})$, $LZ(\hat{Y}_{X'})$ and $LZ(Y)$ define: the Lorenz Zonoids computed on the estimated values provided by the model including also covariate X_k; the Lorenz Zonoids computed on the estimated values provided by the model including the $X^{'}$ covariates but excluding covariate X_k; the Lorenz Zonoid computed on the Y target variable values.

The PGC measure allows us to assess the partial contribution provided by the additional explanatory variable X_k in explaining the response variable mutual variability which is not explained by the $X^{'}$ explanatory variables.

Our proposal can be applied to the game's theoretical context by translating the pay-off notion in terms of the numerator of the PGC measure in equation (2.4). For a set of statistical units ($i = 1, \dots, n$), it derives that the pay-off notion translated in terms of the Lorenz Zonoids ($LZ(\cdot)$) is given by

$$p_{off}(X^k) = LZ(\hat{Y}_{X' \cup X_k}) - LZ(\hat{Y}_{X'}), \tag{2.5}$$

where $LZ(\hat{Y}_{X' \cup X_k})$ and $LZ(\hat{Y}_{X'})$ describe the (mutual) variability of the response variable Y explained by the models including the $X^{'} \cup X_k$ predictors and the $X^{'}$ predictors, respectively.

As we are dealing with a binary response variable, denoting the active and defaulted status of the companies, the terms $\hat{Y}_{X' \cup X_k}$ and $\hat{Y}_{X'}$ can be re-written as the predicted PD $\hat{\pi}_{X' \cup X_k}$ and $\hat{\pi}_{X'}$, when resorting to the logistic regression model including also the explanatory variable X_k and to the logistic regression model not including the explanatory variable X_k, respectively. Thus, equation in (2.5) becomes

$$p_{off}(X^k) = LZ(\hat{\pi}_{X' \cup X_k}) - LZ(\hat{\pi}_{X'}), \tag{2.6}$$

The Lorenz Zonoids $LZ_{d=1}(\hat{Y}_{X' \cup X_k})$ and $LZ_{d=1}(\hat{Y}_{X'})$ in equation (2.6) can be computed by resorting to the covariance operators, i.e.,

$$LZ(\hat{\pi}_{X' \cup X_k}) = \frac{2}{\sum_{i=1}^{n} \hat{\pi}_{iX' \cup X_k}} Cov(\hat{\pi}_{X' \cup X_k}, r(\hat{\pi}_{X' \cup X_k})) \quad \text{and}$$
$$LZ(\hat{\pi}_{X'}) = \frac{2}{\sum_{i=1}^{n} \hat{\pi}_{iX'}} Cov(\hat{\pi}_{X'}, r(\hat{\pi}_{X'})).$$

The Shapley-Lorenz decomposition expression is the result of a combination of the Shapley value-based formula and the Lorenz Zonoid tools. Formally, the contribution of the additional variable X^k, expressed in terms of the differential contribution to the global predictive accuracy, equals to

$$LZ^{X_k}(\hat{\pi}) = \sum_{X' \subseteq C(X) \setminus X_K} \frac{|X^{'}|!(K - |X^{'}| - 1)!}{K!} [LZ(\hat{\pi}_{X' \cup X_k}) - LZ(\hat{\pi}_{X'})], \tag{2.7}$$

where $LZ(\hat{\pi}_{X' \cup X_k})$ and $LZ(\hat{\pi}_{X'})$ measures the marginal contribution provided by the inclusion of variable X_k; K is the number of available predictors; $C(X) \setminus X_k$ is the set of all the possible model configurations which can be obtained with $K - 1$ variables, excluding variable X_k; $|X^{'}|$ denotes the number of variables included in each possible model.

Finally, it is worth noting that the Shapley-Lorenz decomposition presents as an agnostic eXplainable Artificial Intelligence method that can be applied to the predictive output, regardless of which model and data generated it.

2.2.3 Algorithm

The code used to compute the Shapley Lorenz Zonoid values is available on Github.com.

Following the Shapley value attribution method, computing exact Shapley Lorenz Zonoid covariate contribution measures for K covariates, requires the computation of Lorenz Zonoid marginal contributions across 2^K different subsets, per covariate. Computationally this becomes intractable for non-conservative covariate sizes and therefore an approximate solution is implemented in the Shapley Lorenz Zonoid package. The method can be summarised as follows:

2.3 Application

2.3.1 Data

We apply our proposed model to data supplied by the ECAI which specializes in credit scoring for P2P platforms focused on SME commercial lending. The data is described by (Giudici et al., 2019b) to which we refer for further details. In summary, the analysis relies on a dataset composed of official financial information, extracted from the balance sheets of 15,045 SMEs, mostly based in Southern Europe, for the year 2015. The information about the status (0 = active, 1 = defaulted) of each company one year later (2016) is also provided. The observed proportion of defaulted companies is equal to 10.9%. .

Table 2.2 lists the nineteen financial variables included in our dataset. All variables are continuous financial ratios, calculated from the balance sheet variables.

We remark that, for the variables in Table 2.2, and particularly for those reflecting the operations of the companies, there is a noticeable presence of unusually large or small values when compared to the mean. These outliers should not be substituted or deleted as they can provide important explanations for the companies included in the sample. However, their presence is going to affect the robustness of ML models, and of Shapley values in particular. This further motivates the use of Shapley-Lorenz values, more robust to outliers.

2.3.2 Results

Using the data, (Giudici et al., 2019b) have constructed logistic regression scoring models that aim at estimating the PD of each company, using the available financial data from the balance sheets and, in addition, network centrality measures that are obtained from similarity networks.

To improve the predictive performance of the model, (Bussmann et al., 2021) have applied to the same database the Gradient Boosting Tree algorithm, and obtained a substantial increase in predictive performance: the AUROC increases from a value of 0.81 obtained with the application of the logistic regression, to a value of 0.93, obtained with the Gradient Boosting method.

The same authors identify Variable 3: Total Assets/Total Liabilities; Variable 7: EBITDA/Interest paid and Variable 8: (Profit or Loss before tax + Interest paid)/Total asset as the variables that mostly contribute to Shapley values decomposition. This is quite consistent with most credit scoring models, which typically include, among the explanatory variables of credit default: a measure of financial leverage (such as Variable 3) and a measure of profitability (such as Variables 7 and 8).

We use the same data and apply the Gradient Boosting Tree after the data is split into a training set (80%) and a test set (20%). We then calculate, on the same split, the contribution of each of the twenty-three explanatory variables to the estimate of the PD, using two explainable AI methods: the Shapley value approach and the Lorenz-Shapley approach that we propose.

Table 2.3 contains the result of the comparison. For each variable, we report: the value of the Shapley Lorenz Zonoid contribution and the Shapley Value contribution, calculated as the sum of the Shapley values over all observations. For comparison purposes, we also report the contribution of each variable to the deviance (G^2), calculated using the Shapley value formula.

From Table 2.3 note that the variables which most contribute to the prediction of default, according to the sum of the Shapley values, is Variable 8: (Profit or Loss before tax + Interest paid)/Total asset, followed at a considerable distance by Variable 13 and 14 (both related to EBITDA) and by variable 3 (Total Assets/Total liabilities). In terms of G^2, instead, the differences between variable 8 (the highest contributor) and variables 14,15 and 3 are lower. The role that Variable 13 has in terms of Shapley value is replaced by Variable 15. Note that the variables selected as the most important by both Shapley and G^2 have a similar structure: one variable that indicates leverage (Variable 3) and a few variables that express profitability (8,13,14 or 8,14,15). The latter are highly correlated, as they are based on similar information: this may indicate a weakness of the selection methods, possibly redundant and more sensible to outlier observations.

The first column of Table 2.3, giving the Shapley Lorenz values, indicates, instead, that Variable 8, with a value of 0.16, and Variable 3, with a value of 0.11, are one magnitude order higher than the others. This indicates a more clear-cut choice, with only two variables being selected: a measure of leverage, and a measure of profitability. In the latter case, only the most contributing one, among the several that measure profitability, is chosen.

Comparing the results obtained with the different methods, the Shapley Lorenz selection is evidently the easiest to interpret and, by definition, the more robust to outlier observations, as can be easily noticed by repeating the analysis for different training/validation splits.

2.4 Concluding remarks

The chapter proposes a new model agnostic XAI method, based on the combination of Shapley values and Lorenz Zonoids, that can be used to interpret the results of a highly performing ML risk management algorithm.

The proposed method, like other explainable AI models, can identify the variables that most affect the predictions.

A XAI method with normalised values has several desirable advantages. Firstly, the results are user-friendly and easier to interpret in comparison to other XAI methods. This means that users who do not have sufficient technical expertise, such as political decision-makers or members of the board of a company, can quickly and easily identify the features of the model with the most significant impact on the result. They can incorporate this information into their decision-making process. The advantage of normalised values should not be neglected, especially in the context of trustworthiness and in connection with ethical and regulatory requirements for models. As described in the general introduction of this thesis, regulators continuously focus on the transparency and trustworthiness of models in the financial sector. Normalised XAI values support regulators in deciding whether a model fulfils the requirements e.g. transparency, trustworthiness, and ethics.

There are further advantages concerning model debugging and feature selection. By simplifying the identification of features which have a significant impact on the model and features which are redundant, redundant features can be quickly removed from the analysis. Normalised values also support the identification of potential errors in the model, which manifest themselves in unexpected or undesirable model behaviour. These errors can be quickly identified and resolved.

The application of the proposal to a credit risk management use case shows its superior performance, in terms of selectivity, consistency with economic intuition, and robustness.

It identifies one measure of leverage and one measure of profitability as those that most matter.

We can thus conclude that the proposed method is satisfactory and can be proposed as a use case to standardise risk measurement and management in the application of artificial intelligence to credit lending.

Table 2.1: Shapley Lorenz Zonoid Algorithm

```
0.    A pre-defined (user supplied) upper bound on a total number of fully considered subset permutations is defined, given, by say, n_perm. This is similar to the nsamples parameter used in the kernel SHAP module by (Lundberg & Lee, 2017). Unlike the kernel SHAP module, however, only subsets are considered, for which full permutations can be considered, given n_perm. Subsets are considered sequentially, in order of highest to lowest Shapley kernel weights, defined by $\frac{X'(M-X'-1)!}{M!}$. Due to the symmetric property of this Shapley kernel weight, a given subset is always considered pairwise, with its complement, i.e. first all permutations of subset size 1 and those of size $K-1$ are considered. If all permutations of the next subset can be considered as well, given the upper bound in n_perm, this is added to the subset sizes considered in the next step.
1.    Do for $k \in K$: (i.e. for all covariates)
1a)       Do for $s \in \mathcal{C}(X)$ (i.e. for all subset permutations):
1b)           Do for $i \in N$:
                  Let $\tilde{X}$ contain a given permutation of a given subset size without $k$.
                  Compute $E[f(X) \mid \tilde{X} = \tilde{X}_i]$. Once for $\tilde{X}/k$ and once for $\tilde{X} \cup X_k$.
                  Assuming covariate independence, this can be approximated by $\frac{1}{N}\sum_{j=1}^{n} f(X_j/\tilde{X}, \tilde{X}_i)$, and analogously for $\tilde{X}_i \cup X_k$.
                  $X/\tilde{X}$ represents all covariates not included in the subset $X' \cup X_k$ and is obtained by replacing those covariates with either training data or a row-wise shuffled variant of the original covariate matrix.
                  The result of this step, thus is $E[f(X) \mid \tilde{X} = \tilde{X}_i]$, approximated by the sample mean, over the underlying distribution of $X$.
2.            Sort the obtained values for the current permutation of the current subset size iteration.
3.            Compute the Lorenz Zonoid share for the current permutation of the current subset size iteration. Once for the permutation not including $k$ and once for the permutation including $k$. Then compute the difference.
4.            Weight obtained Lorenz Zonoid difference, by the kernel Weight as defined above.
5.        Compute weighted sum of differences
```

ID	Formula or Description	Type
1	Total Assets/Equity	Continuous
2	(Long term debt + Loans)/Shareholders Funds	Continuous
3	Total Assets/Total Liabilities	Continuous
4	Current Assets/Current Liabilities	Continuous
5	(Current assets - Current assets: stocks)/Current liabilities	Continuous
6	Shareholders Funds + Non current liabilities)/Fixed assets	Continuous
7	EBIT/interest paid	Continuous
8	(Profit or Loss before tax + Interest paid)/Total assets	Continuous
9	Return on Equity	Continuous
10	Operating revenues/Total assets	Continuous
11	Sales/Total assets (Activity Ratio)	Continuous
12	Interest paid/(Profit before taxes + Interest paid)	Continuous
13	EBITDA/interest paid	Continuous
14	EBITDA/Operating revenues	Continuous
15	EBITDA/Sales	Continuous
16	Trade Payables/Operating Revenues	Continuous
17	Trade Receivables/Operating Revenues	Continuous
18	Inventories/Operating Revenues	Continuous
19	Turnover	Continuous

Table 2.2: List of financial ratios used as independent variables.

ID	Variable	Shapley-Lorenz	G^2	Shapley
1	(Total assets/Equity	0.00	0.16	2.53
2	(Long term debt + Loans)/Shareholders Funds	-0.00	0.54	-202.80
3	Total assets/Total liabilities	0.11	1088.12	-1273.97
4	Current assets/Current liabilities	0.05	553.68	-641.69
5	(Current assets - Current assets: stocks)/Current liabilities	-0.00	479.06	-93.51
6	(Shareholders Funds + Non current liabilities)/Fixed assets	-0.00	13.16	4180.56
7	EBIT/interest paid	-0.01	411.10	1504.44
8	(Profit (loss) before tax + Interest paid)/Total assets	0.16	1633.51	-13115.53
9	Return on Equity	0.05	826.96	-1993.98
10	Operating revenues/Total assets	0.06	17.36	-289.46
11	Sales/Total assets	-0.02	10.96	252.59
12	Interest paid/(Profit before taxes + Interest paid)	0.01	103.26	379.73
13	EBITDA/interest paid	0.02	418.00	-1697.31
14	EBITDA/Operating revenues	0.03	1254.63	-1419.43
15	EBITDA/Sales	0.02	1122.05	-785.95
16	Trade Payables/Operating revenues	0.00	14.73	-193.60
17	Trade Receivables/Operating revenues	0.05	475.40	-585.58
18	Inventories/Operating revenues	0.01	126.78	1190.47
19	Turnover	0.02	85.26	1072.37

Table 2.3: Marginal contribution of each explanatory variable in terms of Shapley-Lorenz zonoids, G^2 and total Shapley values

3 Explainable Machine Learning to Predict the Cost of Capital

The greatest threat to our planet is the belief that someone else will save it.

– Robert Swan

3.1 Introduction

The employment of AI tools in finance is becoming quite common: by leveraging multidimensional and high-frequency data, AI tools can help in the prediction of returns and risk of securities and risk management (Ortmann, 2016; J. Lin, 2018; Simonian, 2019; Liu, Chen, & Wang, 2022; Cao, 2022). However, traditional AI tools can also be very opaque, making the economic and financial interpretation of the results of AI applications very difficult for an investor. Indeed, regulators have warned investment firms and financial institutions about the use of AI tools, as their interpretability and accountability are key in the policymakers' agenda (Weber, Carl, & Hinz, 2023).

One way to address the interpretability issue is to use models and algorithms from the XAI set of instruments, and those that are able to "open" the black box, such as the Shapley Values or the SHAP Framework (Kumar, Venkatasubramanian, Scheidegger, & Friedler, 2020; Fryer, Strümke, & Nguyen, 2021). Within this framework, this chapter is the first to apply XAI tools to estimate the cost of capital for a sample of large listed companies. The cost of capital represents the remuneration investors require to provide funds to a firm and it is determined by a company's financial and non-financial characteristics and country-specific features. Previous studies choose between two main approaches to proxy the cost of capital: a historical approach (ex-post) or an implied (ex-ante) approach. The first approach is suitable for finding the determinants of the historical cost of capital (e.g., Weighted Average Cost of Capital (WACC) or Capital Asset Pricing Model (CAPM) (Wong et al., 2021; Desender, LópezPuertas-Lamy, Pattitoni, & Petracci, 2020; Shad, Lai, Shamim, & McShane, 2020). The second approach, based on the ex-ante or implied cost of capital, interprets the cost of capital as the risk associated with an investment in the company by an investor. Studies taking this second approach often employ Price Earning Growth models. These rely on 'analysts' forecasts for future earnings to predict the cost of capital (García-Sánchez, Hussain, Khan, & Martínez-Ferrero, 2021; Gupta, 2018; E. P.-y. Yu, Tanda, Luu, & Chai, 2021).

According to the literature, a company's cost of capital is generally determined by internal firm financial characteristics, market features, and, less often, country characteristics (Breuer, Mueller, Rosenbach, & Salzmann, 2018; Desender et al., 2020; Wang, Kartika, Wang, & Luo, 2021; E. P.-y. Yu et al., 2021). Recently also the non-financial behaviour of companies has been studied as a possible determinant of the cost of capital (El Ghoul, Guedhami, Kwok, & Mishra, 2011; Dorfleitner, Halbritter, & Nguyen, 2015). The non-financial performance of companies can determine their riskiness and value (D'Amato et al., 2017; GSIA, 2018; Widyawati, 2020; E. P.-y. Yu, Guo, & Luu, 2018; E. P.-y. Yu et al., 2021). Additionally, the institutional quality of countries where firms are located can affect the perceived riskiness of their business and, as a result, the cost of capital. The previous empirical literature has employed different measures. For instance, (Eldomiaty, Al Qassemi, Mabrouk, & Abdelghany, 2016) employ the Economic Freedom Indicator, while (Grira, Hassan, Labidi, & Soumaré, 2019) more recently employ the measures developed by the International Country Risk Guide on the quality of institutions, democratic tendencies, corruption, and government action. The studies find that institutional quality improves the cost of equity. In this chapter, we choose to employ the World Bank's non-financial features and the Human Development Index as we expect the country's non-financial characteristics to have predictive power on the cost of capital of listed companies.

Additionally, almost the entirety of previous studies employs linear models to investigate the relationship between financial and non-financial characteristics on the cost of capital, while this might not be the case. To overcome this limitation, this chapter applies the XGBoost model and two explainable AI methods, Shapley Values and Lorenz Zonoids, to detect which financial and non-financial factors are good candidates as predictors of the implied cost of capital of more than 1,400 multinational companies listed worldwide.

Thanks to our approach, we are able to provide an intuitive explanation of the contribution of each variable to the model prediction, thereby "opening" the black box of ML.

We contribute to the literature by determining the most relevant financial and non-financial features that predict the implied cost of capital, without making any a priori assumption on the relationships between them and investigating the role of financial and non-financial features both at firm and country levels. We find that besides the traditional drivers of cost of capital - i.e. size, profitability, and liquidity - non-financial features of companies and countries are able to drive the prediction of the cost of capital. Emission intensity is found to predict a higher cost of capital, suggesting that investors penalise companies with high emissions. On the other hand, companies with good governance practices or located in countries with good institutional quality benefit from a lower cost of capital.

We underline that our results have important managerial implications: on one hand, investors can use our results to choose the portfolio allocation that best aligns with their preferences and, on the other hand, companies can have a better understanding of how to improve their financial and non-financial indicators to access to more funding, and at a lower cost.

The remainder of the chapter is organized as follows. Section 2 introduces our proposal; Section 3 describes the data and the variables employed; Section 4 discusses the empirical findings; and, finally, Section 5 concludes.

3.2 Proposal

To analyse the data set and predict the cost of capital, we use the well-known extreme gradient boosting ML model XGBoost. XGBoost is an ensemble learning method that is particularly well suited to large structured data sets. It is a supervised ML model that combines decision tree models with gradient boosting. The model applies decision trees, which are weak classifiers, to a data set, where each subsequent decision tree is built to correct the errors of the previous tree model; (e.g., Chen & Guestrin, 2016). The XGBoost model is a black-box model: its predictions are not explained in terms of their drivers. However, as shown in several recent papers, different XAI methodologies can be applied to explain the predictions of ML models and hence 'open' the black box (Lundberg et al., 2020; Bussmann et al., 2021; Gramegna & Giudici, 2021).

The application of these methods is becoming more common in corporate finance (Ghoddusi, Creamer, & Rafizadeh, 2019). Recently (B. Lin & Bai, 2022) applied a ML approach to estimate the determinants of the cost of debt for 40 listed companies in the mining, steel, and power industries. (Tron, Dallocchio, Ferri, & Colantoni, 2023) investigate the ability of corporate governance features of non-listed companies to determine corporate defaults. Other contributions study the risk management in finance (Gan, Wang, & Yang, 2020) or the application of AI to corporate financial functions (e.g., Polak, Nelischer, Guo,

& Robertson, 2020). In this chapter, by applying different XAI methods to our XGBoost model we aim to identify which financial and non-financial market characteristics mostly affect the cost of capital.

To produce a ranking of the variables, the XGBoost Python package includes an integrated feature importance plot function. The algorithm measures how often each variable is used to split the data, across all decision trees. With this technique, variables that are often used for important splits are identified as the most important for the model predictions (Chen & Guestrin, 2016).

Another popular method to explain complex ML models is the SHAP framework. The SHAP framework defines an interpretation for each prediction in the form of an explanation model. It calculates the average marginal contribution of each feature to the predictions, across all possible feature combinations (Lundberg & Lee, 2017). The underlying Shapley values method (L. S. Shapley, 1953) belongs to the class of additive feature attribution methods and derives from cooperative game theory.

The SHAP algorithm calculates Shapley values, which characterize predictions as linear combinations of binary variables, indicating whether or not each variable is included in the model. As a result, a SHAP value is calculated for each variable, representing the relative contribution to the model predictions (Lundberg & Lee, 2017). The explanation model is a linear function of the binary variables and is defined as in Eq. 3.1.

$$g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \tag{3.1}$$

where:

- $x' \in \{0,1\}^M$,
- $\phi_i \in \mathbb{R}$,
- M is the number of independent variables.

The Shapley value approach, underlying the SHAP algorithm, belongs to the class of additive feature attribution methods. Indeed, (Lundberg & Lee, 2017) showed that the Shapley value method is the only explanation model that jointly satisfies the characteristics of local accuracy, missingness, and consistency. Local accuracy indicates that the sum of all variables of the explanation model approximates the output of the original model. Missingness denotes that missing variables do not receive any importance in the explanation model. Consistency states that a change in the model, which leads to an increase in the contribution of a variable, cannot decrease its importance (Lundberg et al., 2020).

The above characteristics are achieved by assigning to each feature vector, a feature attribution value, which is defined as follows (Eq. 3.2).

The i-th Shapley value of a variable $X_k, (k = 1, \dots, K)$ is:

$$\phi(\hat{f}^k(X_i)) = \sum_{X' \subseteq C(X) \setminus X_k} \frac{|X'|!(K - |X'| - 1)!}{K!} [\hat{f}(X' \cup X_k)_i - \hat{f}(X')_i], \tag{3.2}$$

where $C(X) \setminus X_k$ is the set of all the possible model configurations which can be obtained excluding variable X_k; $\hat{f}(X' \cup X_k)_i$ and $\hat{f}(X')_i)$ are the predictions obtained including and excluding variable X_k.

The Shapley contribution of X_k is the sum (or the mean) of all Shapley values (Lundberg et al., 2020). Although Shapley values are much used in the recent ML literature, they have a drawback: their values are not normalised and, therefore, cannot be easily interpreted and compared across different applications.

To overcome this issue, we propose to employ the Lorenz Model Selection approach introduced by (Giudici & Raffinetti, 2020) to perform variable selection and simplify the ML model. The underlying

Lorenz Zonoid approach is based on the research of (G. Koshevoy, 1995) for empirical distributions and of (Mosler, 1994) for general probability distributions.

Lorenz Model Selection offers a novel method to select variables not based on correlation, but based on a mutual notion of variability. This makes them more robust to outliers. In the univariate case, the Lorenz Zonoid values equate to the Gini coefficient, which can be used to measure the contribution of each explanatory variable to the predictive power of a linear model more accurately. As shown by (Lerman & Yitzhaki, 1984) in the univariate case the Lorenz Zonoid $LZ_{d=1}$ can be expressed by the formula:

$$LZ_{d=1}(Y) = \frac{2Cov(Y, r(Y))}{\mu} \tag{3.3}$$

where:

- Y is the dependent variable,
- μ is the mean value of Y, and
- $r(Y)$ is the rank score of Y variables.

(Giudici & Raffinetti, 2020) show that if we consider the dependent variable Y and the independent variables $X_1, ..., X_h, ..., X_k$ with $h = 1, ..., k$, and we apply a model on this data set, we receive the predictions $\hat{Y}_{X_1,...,X_k}$. The Lorenz Zonoid values are defined accordingly as:

$$LZ_{d=1}(Y) = \frac{2Cov(Y, r(Y))}{n\mu} \tag{3.4}$$

and

$$LZ_{d=1}(\hat{Y}_{X_1,...,X_k}) = \frac{2Cov(\hat{Y}_{X_1,...,X_k}, r(\hat{Y}_{X_1,...,X_k}))}{n\mu} \tag{3.5}$$

where:

- n is the number of all observations,
- $r(\hat{Y}_{X_1,...X_k})$ is the rank score of the predicted variables $\hat{Y}_{X_1,...,X_k}$.

The formulae described above can be rearranged in such a way that the underlying model predictions are generalised and rearranged in a non-decreasing manner, thus yielding a measure of marginal dependence, called the Marginal Gini Coefficient (MGC), which determines the explanatory power of each variable. The MGC can be calculated with the following formula, for any variable X_h, $(h = 1, ..., k)$:

$$MGC(Y|X_h) = \frac{LZ_{d=1}(\hat{Y}_{X_h})}{LZ_{d=1}(Y)} = \frac{Cov(\hat{Y}_{X_h}, r(\hat{Y}X_h)}{Cov(Y, r(Y))} \tag{3.6}$$

The previous formulae can also be rearranged to calculate the additional (partial) contribution of a new explanatory variable, $X_k + 1$, to an existing model, resulting in the partial Gini coefficient (PGC):

$$PGC(Y, X_k + 1|X_1, ..., X_k) = \frac{LZ_{d=1}(\hat{Y}_{X_1,...,X_k+1}) - LZ_{d=1}(\hat{Y}_{X_1,...,X_k})}{LZ_{d=1}(Y) - LZ_{d=1}(\hat{Y}_{X_1,...,X_k})} \tag{3.7}$$

We employ the PGC to measure the contribution of each additional variable to the predictive accuracy of our model, within a stepwise model selection procedure.

To compare any two models, we need to define the payoff. To do this, we calculate the following difference for any statistical unit i:

$$Poff(X_i^k) = \hat{f}(X \cup X_k)_i - \hat{f}(X)_i, \tag{3.8}$$

where:

- $\hat{f}(X)_i$ represents the predictions of a model and
- $\hat{f}(X \cup X_k)_i)$ represents the predictions of a model after including an additional independent variable.

If we replace the model predictions with the PGC, we receive for a given set of statistical units the following expression:

$$Poff(X^k) = LZ_{d=1}(\hat{Y}_{X_1,...,X_k}) - LZ_{d=1}(\hat{Y}_{X_1,...,X_{k-1}}), \tag{3.9}$$

where:

- $LZ_{d=1}(\hat{Y}_{X_1,...,X_{k-1}})$ represent the Lorenz Zonoid values of a model and
- $LZ_{d=1}(\hat{Y}_{X_1,...,X_k})$ represent the predictions of a model after including an additional independent variable.

Once calculated, the pay-off can be assessed in terms of statistical significance, by means of an appropriate test that compares the predictive accuracy of the two models being compared.

As the cost of capital is a continuous variable, we propose to employ the Diebold Mariano test (Diebold & Mariano, 2002), which compares the forecasting accuracy of a continuous response by two competing models.

To perform the test the model predictions need to be compared with the actual observations, and forecast errors calculated. The null hypothesis of the test states that the forecast errors of any two models do not show statistically significant differences and thus the models being compared could not be identified as statistically significantly different in terms of their predictive accuracy.

The null hypotheses of the null difference between the forecast errors are defined by $E[g(e_it)]$, or $E[d_t] = 0$, where $g(e_it)$ is a function of the forecast error and $d_t = [g(e_it) - g(e_jt)]$ is the loss difference. In other words, the null hypothesis that the predictive accuracy of both models is equal can also be expressed as a null hypothesis that the difference between the population mean of the losses is equal to zero.

To determine whether the difference is statistically significant or not, the test statistic can be compared to a critical value from an appropriate distribution, whose parametric form depends on the assumptions about the prediction errors (Diebold & Mariano, 2002).

3.3 Data

To understand which financial and non-financial features contribute to the prediction of the implied cost of capital, we collect data from 2013 to 2019 for more than 1,400 publicly listed companies in 43 countries. Data and information are retrieved from many sources, including Refinitiv Eikon, I/B/E/S, and Bloomberg. Our dependent variable ('implied' or 'ex-ante' cost of capital) is derived from the forward earnings price ratio (Pinto, 2020). As independent variables, we employ all the financial and non-financial variables that are indicated by the literature as relevant in the determination of the cost of capital, as well as country-specific features.

Table 3.1 includes a list of all the variables used in the chapter, with their description.

Table 3.1: The considered explanatory variables

Variable name	Variable description
Firm's financial features	
SIZE	Value of a company's asset size.
ROE	The ratio of net income to 'shareholders' equity (Return on Equity).
VOLATILITY	Volatility of a company's stock price.
Beta	Beta is a proxy of systematic risk and shows how share prices move according to the movements of the relative market index.
CURR	Current ratio: Indicator for a company's liquidity.
QUICK	Quick ratio: Indicator for a company's liquidity.
EPS-GROWTH(t-1)	Value of a company's growth rate on earnings per share (EPS) in t-1.
EPS-GROWTH	Value of a company's growth rate on earnings per share (EPS).
TRAD LIQ	Trading liquidity.
GROWTH SALES 1	A company's growth rate on sales, is based on a ratio of variations in sales over the previous year.
GROWTH SALES 3	Value of a company's average annual growth rate on sales over the previous three years.
RD EXPEND TO NET SALES	Amount of Research and Development expenses divided by net sales.
LEV	Leverage: A ratio of a company's total debt to total assets.
VOL	The trading volume is used for calculating the volume-weighted average price.
SHARES OUT	A company's outstanding shares are available on the market.
GROWTH-EPS	Value of a company's growth rate on earnings per share (EPS).
Firm's non-financial features	
E-DISC	Bloomberg environmental disclosure score measures the amount of environmental information a company reveals to the public. If companies provide all data points collected by Bloomberg, the maximum value of 100 is awarded. The minimum value starts from 0.1.
G-DISC	Bloomberg governance disclosure score measures the amount of governance information a company reveals to the public. If companies provide all data points collected by Bloomberg, the maximum value of 100 is awarded. The minimum value starts from 0.1.
ESG-DISC	Bloomberg ESG disclosure score measures the amount of environmental, social, and governance information a company reveals to the public. If companies provide all data points collected by Bloomberg, the maximum value of 100 is awarded. The minimum value starts from 0.1.

S-DISC	Bloomberg social disclosure score measures the amount of social information a company reveals to the public. If companies provide all data points collected by Bloomberg, the maximum value of 100 is awarded. The minimum value starts from 0.1.
EMIS-INT	Emission intensity of a company.
EIS	Environmental Innovation Score (EIS) is a company's environmental innovation degree, measured based on a company's green revenue and its research and development expenses.
INSI-OWN	Insider ownership: a percentage of equities held by insiders.
BD-SIZE	The number of board members.
BD-INDEP	A percentage of the independent directors on a company's board.
INST-OWN	A percentage of equities held by a company's institutional investors.
Country non-financial features	
WB-V	Indicator of voice for country i represents observations of how a country's inhabitants have the right to vote for their government and freedom to convey their opinions.
WB-RQ	Indicator of the regulatory quality for country i represents opinions on how a government implements its prudent policies to help the private sector grow.
WB-RL	Indicator of the rule of law for country i represents perceptions of how agents have confidence in the general public rules.
WB-GE	Indicator of the government effectiveness for country i represents the quality of a country's public service and a government's creditability to the public.
WB-C	Indicator of control of corruption for country i represents observations of the degree to which the elite and the public power pursue their private interests.
WB-S	Indicator of political stability for country i represents perceptions of the prospects of political uncertainty and terrorism.
HDI	The Human Development Index quantifies a country's development in these key dimensions: its education, health, and economic aspects.
Other control variables	
INF	We collect the inflation rate for our sample countries from the IMF World Economic Outlook Database.
GDPpc	Log (Gross domestic product (GDP) per capita) is measured based on the purchasing power parity exchange rates in 2011. Unit: the U.S. dollar.

From Table 3.1 note that we proxy financial information with the firm's key balance sheet and economic indicators. Non-financial information is proxied via different scores and variables. First, we measure the firm's relevant non-financial information, captured using ESG disclosure and performance scores, in line with the literature (e.g., Breuer et al., 2018; Desender et al., 2020; Mariani, Pizzutilo, Caragnano, & Zito, 2021; Wang et al., 2021)[1]. Second, we also investigate the role of the country features, namely the cultural, socioeconomic, and regulatory framework, that can influence a firm's ex-ante cost of capital. Finally, we include some well-known country and macroeconomic variables, namely inflation and GDP per capita.

3.4 Empirical findings

At first, we split the available data set into an eighty per cent train set and a twenty per cent test set. Before training the XGBoost model, we use the GridSearchCV function from the "sklearn" Python package to determine the optimal hyperparameter settings: it resulted in a learning rate, equal to 0.015; and a maximal depth, equal to 4. We then applied the XGBoost model to the training data set and applied the learned model to predict the response values (Cost of Capital values) in the test data set.

The XGBoost model performs rather well: the predicted mean cost of capital in the test set is 6.44% against an actual mean cost of 6.42%. Furthermore, the root mean Squared Error (RMSE) between the predicted and actual observations is equal to 3%, about half of the mean value, indicating a small variability of the errors.

To explain the obtained predictions, we applied several different XAI methods. First, we analysed the results using the Feature Importance plot, based on the Gini Index, which is included in the XGBoost Python package. The results of the application are shown in Figure 3.1.

From Figure 3.1 we note that the most explainable variables are, for each company, the systemic risk proxy Beta, the environmental innovation score, the stock price volatility, the Return on Equity (ROE), and the size.

However, it is well known that the feature importance plot is a component of tree models, whose results are not stable, as obtained on subsamples, and not globally (Altmann, Toloi, Sander, & Lengauer, 2010). To improve the robustness of the explanations, and overcome the weaknesses of the Feature importance plot, we analysed the same predictions using Shapley values. The calculated SHAP values can be visualised as a summary plot, as in Figure 3.2.

The SHAP summary plot in Figure 3.2 shows the importance of the variables according to their contributions to the model predictions of the cost of capital. The variables are ordered according to their importance, from the most important (top) to the last important (bottom). In the Figure, each dot represents one observation of the underlying data set. When the dots of the variable are located at the right of the 0.000 vertical line it means that the variable has a positive impact on the prediction of the cost of capital; the opposite occurs when the dot is on the left. Blue shades of the dots represent low values of the underlying independent variable and red represents high values of the independent variable.

From Figure 3.2 we note that the most explainable variables are, for each company: the size, the ROE,

[1] It is just the case to recall that the measures employed for ESG scores in empirical papers are not homogeneous, and often lead to different scores for the same company. Among them, the measures obtained by commercial databases (e.g. Bloomberg or Thomson Reuters) are commonly used, but other proxies are also employed: the inclusion in sustainable/ESG indexes (e.g., Eom & Nam, 2017) or initiatives (Fisher-Vanden & Thorburn, 2011); own developed measures, sometimes based on previous literature michaels2017relationship,lau2019economic; hybrid measures based on a mix of the above (García-Sánchez et al., 2021). In this chapter, we choose to rely on the Refinitiv Eikon and Bloomberg ESG information

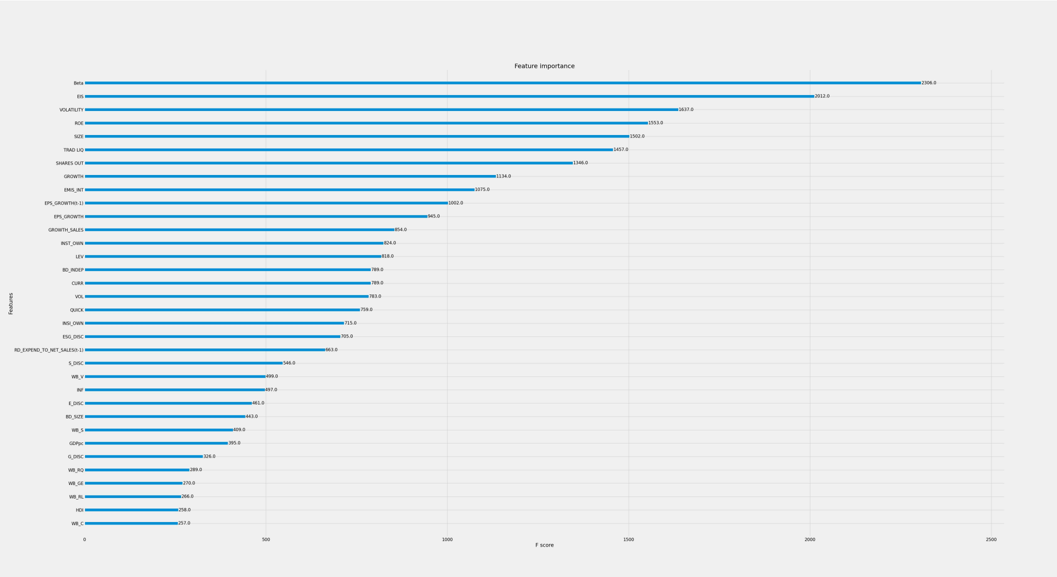

Figure 3.1: Feature importance plot. The Figure presents the feature importance plot computed using the XGBoost algorithm.

the stock price volatility, WB_V (i.e. a variable which describes citizens' right to vote, and freedom to convey opinions) and the liquidity of the company, described by the Trade liquidity. Furthermore, Beta (a measure of the systematic risk of the companies' stock) and emission intensity appear quite highly placed in the features ranking.

Comparing the five most important variables in Figure 3.1 with those in Figure 3.2 note that three of them are the same, namely, Size, ROE, and Stock price volatility. Whereas the systematic risk proxy Beta is first in Figure 3.1 and sixth in Figure 3.2. The Trade liquidity is sixth in Figure 3.1 and fifth in Figure 3.2.

Differences include, for instance, the placement of the variable EIS, which is ranked second in Figure 3.1 and only 25th in Figure 3.2. Conversely, the country's institutional quality, namely the variable 'Voice' (WB_V), is captured among the most important variables only by Shapley values.

The difference between the two XAI tools may be due to the inclusion of many variables in the ML model, some of which have only a very small impact. This suggests performing a preliminary feature selection, to improve the robustness of the model.

To this aim, we have created a series of sub-datasets based on the feature ranking in the SHAP approach. The first data set consists only of the most important variable (SIZE). The second data set consists of the most important and the second most important variables (SIZE and ROE). We continued this subdivision until we obtained 35 sub-datasets, corresponding to all considered variables. We calculated the Lorenz Zonoid values for each of the chosen sub-datasets, which corresponds to an increasing number of explanatory variables: from 1 to 35.

Figure 3.3 represents graphically the Lorenz Zonoid values calculated on each of the 35 subsets, ordered from the smallest (with only one variable included in the model) to the largest (all variables included in the model).

From Figure 3.3 note that the highest Lorenz Zonoid value (0.1865) is achieved with the inclusion in the model of 13 variables. In other words, Figure 3.3 indicates that, according to the parsimony principle, good predictions are likely to be obtained by drastically simplifying the model from 35 to 13 features: a much simpler model.

Before concluding with the choice of a model with 13 variables, note that Figure 3.3 shows lower increments of Lorenz Zonoid values after including four variables. When comparing the MSE of the model with 13 variables (0.0010) with the MSE of the model with 4 variables (0.0012), it can be seen that the model which includes 13 variables performs only slightly better. The Lorenz Zonoids of the corresponding sub-sets of four variables are plotted in Figure 3.4, extracted and magnified from Figure 3.3.

From Figure 3.4 note that the feature SIZE, which represents the asset size of a company, explains about 11% of the predictive accuracy of the model. When ROE is added to SIZE there is an increase of about 2% in accuracy. Adding VOLATILITY a further increase of about 2% and adding WB_V produces an increase of about 1%.

To gain a better insight on whether to further simplify the chosen ML model, from 13 to four variables, we further analysed our results with the help of the Diebold Mariano test (Diebold & Mariano, 2002).

More precisely, we compared the model which consists of only four variables with the model which consists of 13 variables, based on the results of the Lorenz Zonoid approach. The result of the test gives a p-value of 0.999. Since the p-value is higher than 0.05 the null hypothesis that the predictive power of the simpler model (with four variables) is as good as the predictive power of the more complex model (with 13 variables) cannot be rejected.

Thus, the result of the Diebold Mariano test shows that we can exclude all other variables from the data set and select a model that only contains four variables: SIZE, ROE, VOLATILITY, and WB_V.

From an economic viewpoint, the chosen four variables are found to be the most relevant in predicting the ex-ante cost of capital. Three of them refer to the firm financial characteristics. The fourth one is a non-financial country-related feature.

The variable SIZE is the most important variable for the XGBoost model and it represents the asset size of a company. Concerning the sign of importance, it can be seen from the SHAP summary plot in 3.2 that companies with a large asset size have a positive impact on the model's predictions of the cost of capital. Hence, the model predicts a higher cost of capital for companies with a large asset size and a lower cost of capital for companies with a small asset size. Asymmetries of information would call for the opposite effect, with larger companies being less exposed to asymmetries of information (Armstrong, Core, Taylor, & Verrecchia, 2011; Embong, Mohd-Saleh, & Sabri Hassan, 2012; W. P. He, Lepone, & Leung, 2013). Nevertheless, the peculiarity of our sample suggests that this pool of companies - which are all very large multinational listed companies - is characterised by having larger total assets. This might imply being 'too' large to be understood and, hence, growing after a certain threshold might induce investors to perceive the complexity of the company as a hurdle rather than an advantage for the future cash flows determining future profitability and, in the final stance, the cost of capital.

The same plot in 3.2 shows that the importance of ROE indicates that especially low values of return on equity have a strong impact on the models' predictions.

From the SHAP summary plot, it can also be seen that high values of the volatility of a company lead to increased predictions of the cost of capital. A possible explanation for this result is that investors associate the high volatility of a company's stock price with higher risk and uncertainty. This indicates that future values of the company stock price are uncertain and hence, lead to higher cost of capital.

As already mentioned, the fourth most important variable is 'WB_V', a country-specific feature. The variable indicates the political and regulatory framework of the country, describing how a country's citizens express their votes for the government and how their opinions are conveyed and heard. We can see from the SHAP summary plot in 3.2 that especially high values of this variable have a strong impact on the models' predictions, leading to an increase in the predicted cost of capital.

We finally remark that our empirical findings indicate that the company's emissions are a significant predictor of the cost of capital, although the variable is not included in the selected parsimonious model with four variables. This result may be due to corporate emissions being related to important variables such as the size and ROE of a company, as well as the institutional quality of a country, described by variable 'WB_V'. We can thus conclude that our findings indicate that Environmental, Social, and Governance factors, such as a company's emission and country regulatory characteristics have an important role in determining the cost of capital for a company, either directly or indirectly.

3.5 Conclusions

This chapter investigates for the first time the determinants of the cost of capital through a ML model, in combination with the SHAP framework and the Lorenz Zonoid approaches to make it explainable. We are able to overcome the a priori hypothesis on the linearity of the relationship among variables and are able to individuate and rank the features that contribute more to the prediction of the cost of capital.

Overall, our results show that a firm's size, ROE, portfolio volatility risk, ESG behaviour and country's institutional quality are the most valuable variables in predicting a firm's ex-ante cost of capital. Concerning non-financial features, the Shapley values approach shows that some of the non-financial indicators, proxied by ESG factors, such as Emission intensity or corporate governance settings, can be adopted as good predictors of the cost of equity besides the traditional financial features of companies. These results corroborate the proposals made by policymakers and indicate that the market penalises companies with high emission intensity with more expensive capital funding. On the other hand, the market awards companies with good corporate governance practices by charging a lower cost of capital.

Additionally, our empirical results employing Lorenz Zonoid show that a firm's cost of capital is well predicted by the level of the country's voice, which we use to proxy the institutional quality of the country where the firm is incorporated.

Our study provides supporting evidence that some key non-financial features both at firm and country levels can contribute to shaping investors' risk perception.

Our findings also suggest that investors perceive the most polluted firms as riskier and more costly in the future, they consequently require these firms with higher emission intensity with a more expensive cost of capital. In other words, this chapter provides evidence indicating that investors punish the most polluted firms.

According to this finding, we suggest policymakers call for more transparency in disclosing the ESG data at the firm level, which will help investors make better decisions on their long-term investment strategy and asset allocations. Future research can be devoted to understanding if and how these results change depending on the industries considered or over time, as regulation is modified and sustainability becomes integrated into the institutional setting of the different countries.

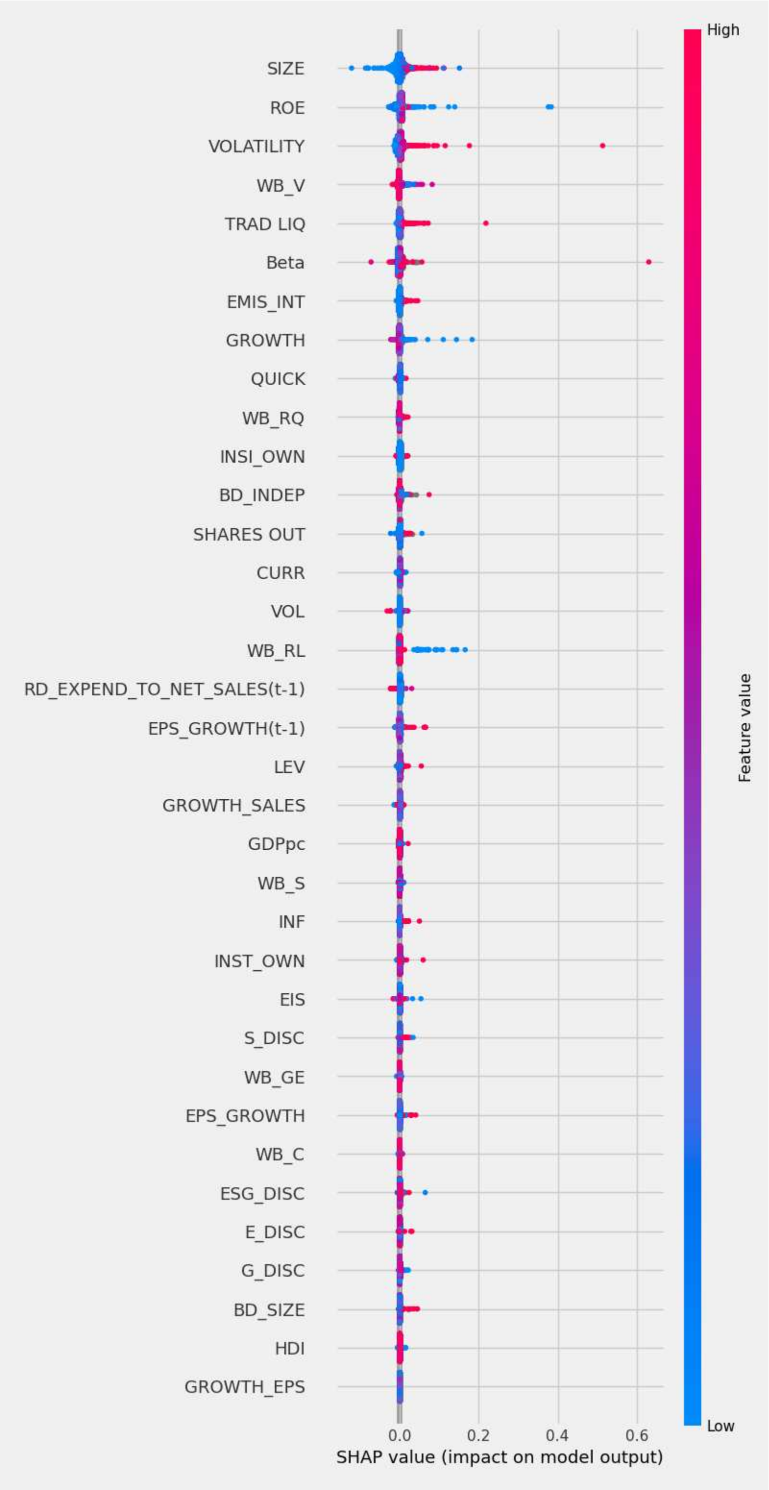

Figure 3.2: SHAP summary plot.

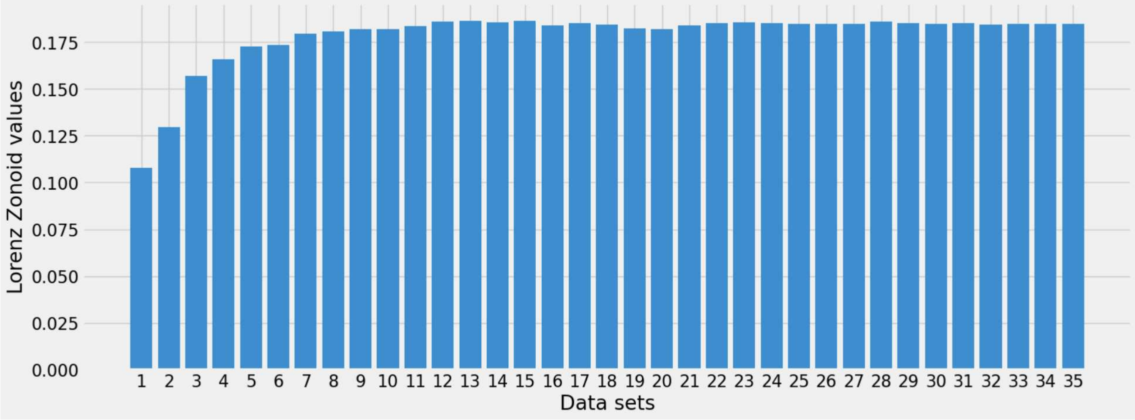

Figure 3.3: Lorenz Zonoid values plot.

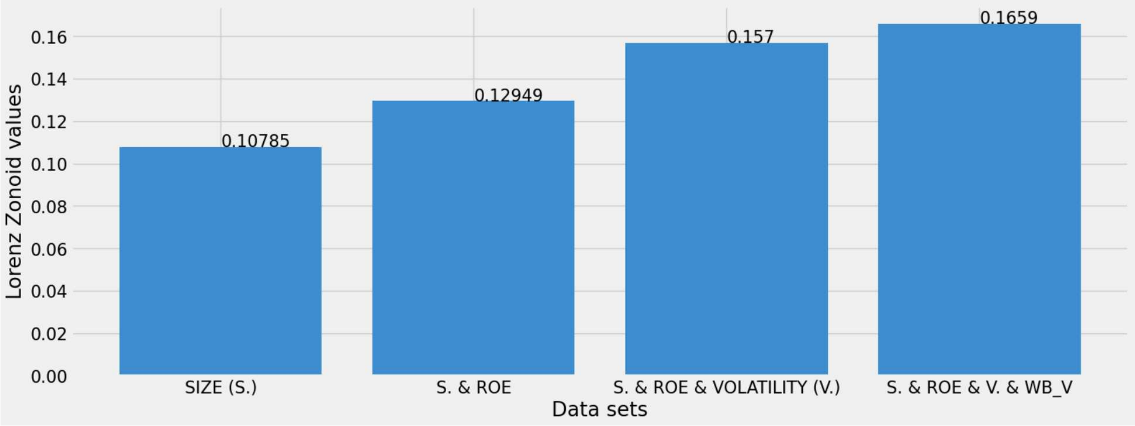

Figure 3.4: Lorenz Zonoid values plot for the first four subsets.

4 Concluding remarks and future research

The increasing amount of available data and the simplified access to high-performing computers enables the use of ML models. Companies that operate in highly regulated markets such as the financial market are also taking advantage of this opportunity. However, the use of ML models is accompanied by various risks, such as the decreasing explainability of these algorithms. The motivation of this doctoral thesis was on the one hand to take a closer look at these risks and on the other hand to further analyse the opportunities that this technology represents for the financial industries.

As described in the general introduction of this doctoral thesis, there are different types of XAI methods. One already widespread method in the financial industries is the application of black box models in combination with the use of model agnostic post-hoc XAI methods. Hence, this thesis focuses on the above-mentioned methods and does not discuss the topic of intrinsically (a priori) explainable ML models.

Many of the currently available methods are well suited to explain the results of models to model developers or data scientists. However, they are too complex and confusing for non-specialist stakeholders such as members of the board or regulators. Hence, the principle of the explainee was introduced in the general introduction. It describes that XAI methods must be adapted to the corresponding person to whom the model should be explained. This explainee-oriented application of XAI methods is also described in (Bracke et al., 2019). In other words, to make a model explainable, different XAI methods should be applied and the method that is the most suitable for the level of experience of the particular explainee should be used.

Therefore, in this doctoral thesis, existing XAI methods were analysed and further developed to make the results of AI models more explainable and easier to understand for non-specialist stakeholders. This provides added value, especially in highly regulated markets, such as the financial market, where the use of models has to be approved by senior management, regulators or policymakers.

In chapter 1, we used a XAI method to explain the predictions of a complex ML model which predicts the PD of loan applicants. We contributed to the ongoing debate about the application of ML models in credit risk management. We demonstrated that XAI methods show promising results regarding the explainability of the ML model's decisions. The comparison of the models proved the higher model performance of the XGBoost model. We analysed the results using the SHAP framework to identify the most important variables of the predictions. We visualised the resulting SHAP values using a MST and standardized Euclidean distance to represent 1) general clusters and 2) the defaulted companies as individual clusters.

The results of our analysis show that every single prediction of the model could be explained by the applied XAI methods. This demonstrates that the application of ML models enables the analysis of credit risk parameters more accurately. The additional implementation of network-based XAI methods provides stakeholders like e.g. policymakers and regulators deep insights into the model processes but with a lower complexity by using a visualisation. Especially, the second network model enables stakeholders to identify that most of the defaulted observations are clustered in similar areas of the network. The network is based on Shapley values. Hence, showing stakeholders the network in combination with the SHAP bar plot of the most important variables for the models' predictions streamlines the traceability of the Shapley values and as such, increases the overall explainability of the model.

Future research should investigate how the Shapley Correlation Network deals with highly imbalanced data and how the results change, especially the network model that highlights the defaulted companies.

In chapter 2, we developed the Shapley-Lorenz approach and contributed to the ongoing development of new XAI methods to further improve the explainability of AI models. The method is more accessible and hence, can be used by several different stakeholder groups to make the output of AI models explainable. The applied combination of Shapley Values and the Lorenz Zonoids approach provides normalised values as a result and enables comparable interpretations of different models' predictions. The comparison of the Shapley-Lorenz approach, the Shapley values approach, and the calculated contribution of each variable to the deviance G^2, described in chapter 2 shows that the Shapley-Lorenz approach is the easiest method to interpret. Additionally, the Shapley-Lorenz approach is more robust to changes in the database than the other XAI methods. The possibility to compare variables according to their explainability due to the normalised Shapley-Lorenz values is an advantage that supports the applicability of AI models, especially in the financial industry. Showing the explainability of a model with normalised values supports model risk managers, model validators, regulators, auditors, but also the model developers themselves in their decision-making process and it also improves the model output quality. The Shapley-Lorenz approach can facilitate the validation of model results and supports the decision of whether a model is sufficiently explained.

The methodology of the Shapley-Lorenz values needs further improvements since the current version requires significant amounts of computational power. To avoid long run times we included a warning message to the algorithm, which recommends users to use a sample data set of not more than 50 observations. Future research should improve the existing code and perform analyses including enlarged data sets.

Chapter 3 addresses the currently very relevant topic regarding the application of ESG variables in the financial industries. European guidelines like the "Guide on Climate-Related and Environmental Risks" ECB (ECB, 2020) or the EU Sustainable Finance Disclosure Regulation" (European Commission) (EC, 2019b) expect financial institutions to incorporate climate-related and environmental risks into their risk management framework and to disclose this sustainability information to their stakeholders.

These guidelines and regulations do not specify exactly which ESG factors should be incorporated into the risk framework. Chapter 3 contributes to the current debate on how ESG risks should be integrated into the risk framework of financial institutions due to their importance and grade of information. Using a ML model, we were able to identify which financial and non-financial factors are the most important drivers of a company's overall risk from an investor's perspective. This research can help policymakers, regulators, and financial institutions to identify the most important ESG variables for their risk analysis. It also provides insights into the impact of different financial, economic, and ESG risk-related variables in predicting a company's cost of capital, which we used as a proxy for a company's risk. The results show that in addition to the traditional financial variables, many ESG variables were also important in predicting the outcome. Some of the ESG variables were more important than traditional economic variables such as growth in GDP per capita.

The data set was compiled manually from various sources in a time-consuming process. There are no comparable, publicly available data sets that combine financial and non-financial risk areas. Some of the analysed variables showed a rather poor database. If the database improves, future research should repeat the performed analysis, including the improved database and see how the results will change. It would further be interesting to repeat the experiment with an enlarged database on a yearly basis over several years to see how the importance of certain variables changes over time.

This thesis analysed the existing methodologies to explain the output of black box models and fur-

ther developed them to increase the usability of ML models in the financial industries. Since this thesis is the result of an executive PhD program the studies are focused on real-world scenarios in the financial industries. Future research should extend the different methodologies by applying them to different data sets as mentioned above.
Even though this thesis contains many positive aspects of the use of XAI methods to explain black box ML models, there is also criticism. For example, (Rudin, 2019) points out that in recent years there has been an increasing focus on the development of XAI methods to explain black box ML models instead of focusing on the development of inherently explainable ML models. XAI methods are often only downstream models that cannot provide a complete explanation of the black box ML models and therefore might be misunderstood. This is particularly problematic in high-risk markets, such as the financial market, since a full understanding of the models with explanations at a detailed level is often required. As an alternative, (Rudin, 2019) suggests focusing on the development of ML models that inherently provide an explanation for their decision.

Another critique focuses on XAI methods that use a background sample data set, such as LIME or SHAP. (Slack, Hilgard, Jia, Singh, & Lakkaraju, 2020) demonstrated the possibility of deliberately altering the outcome of XAI methods by changing the input data distribution. The performed adversarial attacks show that the data distribution was biased but the post-hoc explanations looked innocuous. Regulators, auditors, and risk managers should be aware of this risk and consider not only the ML models but also the XAI methods in the course of their model audits or validations. Another alternative could be the use of XAI methods which take the complete data set into account, such as the calculation of Shapley values. A disadvantage of this alternative is the considerable computational effort. (Huang & Marques-Silva, 2023) claim in their recent paper that for some specific use cases even Shapley Values, calculated with the whole data set, identify variables as important which do not have any importance for the model's prediction at all.

Future Research should take the above-mentioned risks into account and further develop inherently explainable models or combine and extend XAI methods to provide complete explanations and increase their robustness against manipulation.

References

Abellán, J., & Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications*, *73*, 1–10.

Ahelegbey, D. F., Giudici, P., & Hadji-Misheva, B. (2019). Factorial network models to improve p2p credit risk management. *Frontiers in Artificial Intelligence*, *2*, 8.

Altman, E. I., & Heine, M. L. (2006). Default recovery rates and lgd in credit risk modeling and practice: an updated review of the literature and empirical evidence.

Altmann, A., Toloi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. doi: 10.1093/bioinformatics/btq134

Anderson, R. (2007). *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*. Oxford University Press.

Armstrong, C. S., Core, J. E., Taylor, D. J., & Verrecchia, R. E. (2011). When does information asymmetry affect the cost of capital? *Journal of accounting research*, *49*(1), 1–40. Retrieved from `https://www.jstor.org/stable/20869861`

BaFin. (2023, May). *Rundschreiben 05/2023 (ba) - mindestanforderungen an das risikomanagement - marisk.*

BCBS. (2000, September). *Principles for the management of credit risk.* https://www.bis.org/publ/bcbs75.pdf.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social networks*, *29*(4), 555–564.

Bonanno, G., Caldarelli, G., Lillo, F., & Mantegna, R. N. (2003). Topology of correlation-based minimal spanning trees in real and model markets. *Physical Review E*, *68*(4), 046130.

Bracke, P., Datta, A., Jung, C., & Shayak, S. (2019). Machine learning explainability in finance: an application to default risk analysis. *Staff Working Paper No. 816*. Retrieved from `https://www.bankofengland.co.uk/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis`

Breuer, W., Mueller, T., Rosenbach, D., & Salzmann, A. (2018). Corporate social responsibility, investor protection, and cost of equity: A cross-country comparison. *Journal of Banking and Finance*, *96*, 34–55.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, *30*(1-7), 107–117.

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, *57*, 203–216. doi: 10.1007/s10614-020-10042-0

Cao, L. (2022). AI in finance: challenges, techniques, and opportunities. *ACM Computing Surveys (CSUR)*, *55*(3), 1–38. doi: 10.48550/arXiv.2107.09051

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Claessens, S., Frost, J., Turner, G., & Zhu, F. (2018). Fintech credit markets around the world: size, drivers and policy issues. *BIS Quarterly Review September*.

Croxson, K., Bracke, P., & Jung, C. (2019, May). Explaining why the computer says 'no'. *FCA - Insight*.

D'Amato, D., Droste, N., Allen, B., Kettunen, M., Lähtinen, K., Korhonen, J., et al. (2017). Green, circular, bio economy: A comparative analysis of sustainability avenues. *Journal of Cleaner Production*, *168*, 716–734. doi: 10.1016/j.jclepro.2017.09.053

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837–845.

Desender, K. A., LópezPuertas-Lamy, M., Pattitoni, P., & Petracci, B. (2020). Corporate social responsibility and cost of financing–the importance of the international corporate governance system. *Corporate Governance: An International Review*, *28*, 207–234. doi: 10.1111/corg.12312

Dhaliwal, D. S., Li, O. Z., Tsang, A., & Yang, Y. G. (2011). Voluntary non-financial disclosure and the cost of equity capital: The initiation of corporate social responsibility reporting. *The Accounting Review*, *86*(1), 59–100.

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, *20*(1), 134–144. doi: 10.1080/07350015.1995.10524599

Dorfleitner, G., Halbritter, G., & Nguyen, M. (2015). Measuring the level and risk of corporate responsibility–an empirical comparison of different ESG rating approaches. *Journal of Asset Management*, *16*, 450–466. doi: 10.1057/jam.2015.31

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

EBA. (2020, January). *Eba report on big data and advanced analytics.*

EC. (2013a, June). *Directive 2013/36/eu of the european parliament and of the council of 26 june 2013 on access to the activity of credit institutions and the prudential supervision of credit institutions and investment firms, amending directive 2002/87/ec and repealing directives 2006/48/ec and 2006/49/ec.* Journal of the European Union.

EC. (2013b, June). *Regulation (eu) no 575/2013 of the european parliament and of the council of 26 june 2013 on prudential requirements for credit institutions and investment firms and amending regulation (eu) no 648/2012.* Journal of the European Union.

EC. (2016). *Regulation (eu) 2016/679 - general data protection regulation (gdpr).*

EC. (2019a, April). *Ethics guidelines for trustworthy ai.*

EC. (2019b, November). *Regulation (eu) 2019/2088 of the european parliament and of the council of 27 november 2019 on sustainability-related disclosures in the financial services sector.*

EC. (2021). *Proposal for a regulation laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts.*

ECB. (2019). *Guide to the internal models approach (ima).*

ECB. (2020, November). *Guide on climate-related and environmental risks.*

Eldomiaty, T. I., Al Qassemi, T. B. F., Mabrouk, A. F., & Abdelghany, L. S. (2016). Institutional quality, economic freedom and stock market volatility in the MENA region. *Macroeconomics and Finance in Emerging Market Economies*, *9*(3), 262–283. doi: 10.1080/17520843.2015.1093011

El Ghoul, S., Guedhami, O., Kwok, C. C., & Mishra, D. R. (2011). Does corporate social responsibility affect the cost of capital? *Journal of Banking & Finance*, *35*(9), 2388–2406.

Embong, Z., Mohd-Saleh, N., & Sabri Hassan, M. (2012). Firm size, disclosure and cost of equity capital. *Asian Review of Accounting*, *20*(2), 119–139. doi: 10.1108/13217341211242178

Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending. *Applied Economics*, *47*(1), 54–70.

Eom, K., & Nam, G. (2017). Effect of entry into socially responsible investment index on cost of equity and firm value. *Sustainability*, *9*(5), 717. doi: 10.3390/su9050717

FED. (2011). *Supervisory policy on model risk management.*

Fisher-Vanden, K., & Thorburn, K. S. (2011). Voluntary corporate environmental initiatives and shareholder wealth. *Journal of Environmental Economics and Management*, *62*(3), 430–445.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of Statistics*.

Fryer, D., Strümke, I., & Nguyen, H. (2021). Shapley values for feature selection: The good, the bad, and the axioms. *IEEE Access*, *9*, 144352–144360. doi: 10.48550/arXiv.2102.10936

FSB. (2017, November). *Artificial intelligence and machine learning in financial services - market developments and financial stability implication* (Tech. Rep.). Financial Stability Board.

Gan, L., Wang, H., & Yang, Z. (2020). Machine learning solutions to challenges in finance: An application to the pricing of financial products. *Technological Forecasting and Social Change*, *153*, 119928. doi: 10.1016/j.techfore.2020.119928

García-Sánchez, I.-M., Hussain, N., Khan, S.-A., & Martínez-Ferrero, J. (2021). Do markets punish or reward corporate social responsibility decoupling? *Business & Society*, *60*(6), 1431–1467. doi: 10.1177/0007650319898839

Ghoddusi, H., Creamer, G. G., & Rafizadeh, N. (2019). Machine learning in energy economics and finance: A review. *Energy Economics*, *81*, 709–727. doi: 10.1016/j.eneco.2019.05.006

Giudici, P. (2005). *Applied data mining: statistical methods for business and industry.* John Wiley & Sons.

Giudici, P. (2018). Financial data science. *Statistics and Probability letters*, 160-164.

Giudici, P., Hadji-Misheva, B., & Spelta, A. (2019a). Correlation network models to improve p2p credit risk management. *Artificial Intelligence in Finance.*

Giudici, P., Hadji-Misheva, B., & Spelta, A. (2019b). Network based credit risk models. *Quality Engineering*, 1–13.

Giudici, P., & Misheva, B. H. (2018). P2p lending scoring models: Do they predict default? *Journal of Digital Banking*, *2*(4), 353–368.

Giudici, P., & Raffinetti, E. (2020). Lorenz model selection. *Journal of Classification*, *37*(3), 754–768. doi: 10.1007/s00357-019-09358-w

Giudici, P., & Raffinetti, E. (2021). Shapley-Lorenz explainable artificial intelligence. *Expert systems with applications*, *167*, 114104.

Giudici, P., Sarlin, P., & Spelta, A. (2017). The interconnected nature of financial systems: direct and common exposures. *Journal of Banking & Finance.*

Gramegna, A., & Giudici, P. (2021). SHAP and LIME: an evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence*, *4*, 752558. doi: 10.3389/frai.2021.752558

Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly*, 497–530.

Grira, J., Hassan, M. K., Labidi, C., & Soumaré, I. (2019). Equity pricing in Islamic banks: International evidence. *Emerging Markets Finance and Trade*, *55*(3), 613–633. doi: 10.1080/1540496X.2018.1451323

GSIA. (2018). 2018 global sustainable investment review.

Guegan, D., Hassani, B., et al. (2017). *Regulatory learning: Credit scoring application of machine learning* (Tech. Rep.). HAL.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., et al. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1–42.

Gunning, D., & Aha, D. (2019). Darpa's explainable artificial intelligence (xai) program. *AI magazine*, *40*(2), 44–58.

Gupta, K. (2018). Environmental sustainability and implied cost of equity: International evidence. *Journal of Business Ethics*, *147*(2), 343–365. Retrieved from `http://www.jstor.org/stable/45022380`

Hand, D., Smyth, D., Hand, P., Mannila, H., & Smyth, P. (n.d.). *Principles of data mining.* MIT Press.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263–1284.

He, W. P., Lepone, A., & Leung, H. (2013). Information asymmetry and the cost of equity capital. *International Review of Economics & Finance*, *27*, 611–620. doi: 10.1016/j.iref.2013.03.001

Huang, X., & Marques-Silva, J. (2023). The inadequacy of shapley values for explainability. *arXiv preprint arXiv:2302.08160.*

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Joseph, A. (2019a). Parametric inference with universal function approximators. *Staff Working Paper No. 784.* Retrieved from `https://www.bankofengland.co.uk/working-paper/2019/shapley-regressions-a-framework-for-statistical-inference-on-machine-learning-models`

Joseph, A. (2019b, March). *Shapley regressions: a framework for statistical inference on machine learning models* (Research report No. 784). Bank of England.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, *18*(1), 39–43.

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

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, *46*(5), 604–632.

Koshevoy, G. (1995). Multivariate Lorenz majorization. *Social Choice and Welfare*, 93–102. Retrieved from `https://www.jstor.org/stable/41106114`

Koshevoy, G. A., & Mosler, K. (1996). The lorenz zonoid of a multivariate distribution. *Journal of the American Statistical Association*, *91*(434), 873–882.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with shapley-value-based explanations as feature importance measures. In *International conference on machine learning* (pp. 5491–5500).

Lau, C.-K. (2019). The economic consequences of business sustainability initiatives. *Asia Pacific Journal of Management*, *36*(4), 937–970.

Leondes, C. T. (2001). *Expert systems: the technology of knowledge management and decision making for the 21st century.* Elsevier.

Lerman, R. I., & Yitzhaki, S. (1984). A note on the calculation and interpretation of the Gini index. *Economics Letters*, *15*(3-4), 363–368. doi: 10.1016/0165-1765(84)90126-5

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136.

Lin, B., & Bai, R. (2022). Machine learning approaches for explaining determinants of the debt financing in heavy-polluting enterprises. *Finance Research Letters*, *44*, 102094.

Lin, J. (2018). Using weighted Shapley values to measure the systemic risk of interconnected banks. *Pacific Economic Review*, *23*(2), 244–270. doi: 10.1111/1468-0106.12155

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18.

Liu, L., Chen, C., & Wang, B. (2022). Predicting financial crises with machine learning methods. *Journal of Forecasting*, *41*(5), 871–910. doi: 10.1002/for.2840

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, *2*(1), 56–67. doi:

10.48550/arXiv.1905.04610

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*. Retrieved from `https://doi.org/10.48550/arXiv.1705.07874` doi: 10.48550/arXiv.1705.07874

Mantegna, R. N., & Stanley, H. E. (1999). *Introduction to econophysics: Correlations and complexity in finance.* Cambridge: Cambridge University Press.

Mariani, M., Pizzutilo, F., Caragnano, A., & Zito, M. (2021). Does it pay to be environmentally responsible? investigating the effect on the weighted average cost of capital. *Corporate Social Responsibility and Environmental Management*, *28*(6), 1854–1869.

Merton, R. C. (1987). A simple model of capital market equilibrium with incomplete information. *The Journal of Finance*, *42*(3), 483–510.

Michaels, A., & Grüning, M. (2017). Relationship of corporate social responsibility disclosure on information asymmetry and the cost of capital. *Journal of Management Control*, *28*(3), 251–274.

Milne, A., & Parboteeah, P. (2016). *The business models and economics of peer-to-peer lending* (resreport No. 17). ECRI.

Molnar, C. (2020). *Interpretable machine learning.* Lulu.com.

Mosler, K. (1994). Majorization in economic disparity measures. *Linear Algebra and its applications*, *199*, 91–114.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080.

Newman, M. (2018). *Networks.* Oxford university press.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, *103*(23), 8577–8582.

Ortmann, K. M. (2016). The link between the shapley value and the beta factor. *Decisions in economics and finance*, *39*, 311–325.

Pecora, N., Kaltwasser, P. R., & Spelta, A. (2016). Discovering sifis in interbank communities. *PloS one*, *11*(12), e0167781.

Pencina, M. J., D'Agostino Sr, R. B., D'Agostino Jr, R. B., & Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond. *Statistics in medicine*, *27*(2), 157–172.

Perra, N., & Fortunato, S. (2008). Spectral centrality measures in complex networks. *Physical Review E*, *78*(3), 036107.

Pinto, J. E. (2020). *Equity asset valuation.* John Wiley & Sons.

Polak, P., Nelischer, C., Guo, H., & Robertson, D. C. (2020). Intelligent finance and treasury management: what we can expect. *AI & Society*, *35*, 715–726. doi: 10.1007/s00146-019-00919-6

Reinsel, D., Gantz, J., & Rydning, J. (2017). Data age 2025: The evolution of data to life-critical. *Don't Focus on Big Data*, *2*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206–215.

Serrano-Cinca, C., & Gutiérrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (p2p) lending. *Decision Support Systems*, *89*, 113–122.

Shad, M. K., Lai, F.-W., Shamim, A., & McShane, M. (2020). The efficacy of sustainability reporting towards cost of debt and equity reduction. *Environmental Science and Pollution Research*, *27*, 22511–22522. doi: 10.1007/s11356-020-08398-9

Shapley, L. (1953). A value for n-person games. *Contributions to the Theory of Games*, 307-317.

Shapley, L. S. (1953). Stochastic games. *Proceedings of the national academy of sciences*, *39*(10), 1095–1100.

Simonian, J. (2019). Portfolio selection: a game-theoretic approach. *The Journal of Portfolio Management*, *45*(6), 108–116. doi: 10.3905/jpm.2019.1.095

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the aaai/acm conference on ai, ethics, and society* (pp. 180–186).

Sokol, K., & Flach, P. (2021). Explainability is in the mind of the beholder: Establishing the foundations of explainable artificial intelligence. *arXiv preprint arXiv:2112.14466*.

Spelta, A., & Araújo, T. (2012). The topology of cross-border exposures: beyond the minimal spanning tree approach. *Physica A: Statistical Mechanics and its Applications*, *391*(22), 5572–5583.

Spelta, A., Flori, A., & Pammolli, F. (2018). Investment communities: Behavioral attitudes and economic dynamics. *Social Networks*, *55*, 170–188.

Stigler, G. J. (1958). The economies of scale. *The Journal of Law & Economics*, *1*, 54–71.

Strumbelj, E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, *11*, 1–18.

Thomas, L., Edelman, D., & Crook, J. (1997). Credit scoring and its applications. *SIAM Monographs*.

Tron, A., Dallocchio, M., Ferri, S., & Colantoni, F. (2023). Corporate governance and financial distress: Lessons learned from an unconventional approach. *Journal of Management and Governance*, *27*(2), 425–456. doi: 10.1007/s10997-022-09643-8

Tseng, C.-y., & Demirkan, S. (2021). Joint effect of CEO overconfidence and corporate social responsibility discretion on cost of equity capital. *Journal of Contemporary Accounting & Economics*, *17*(1), 100241. doi: 10.1016/j.jcae.2020.100241

Wang, K. T., Kartika, F., Wang, W. W., & Luo, G. (2021). Corporate social responsibility, investor protection, and the cost of equity: Evidence from east asia. *Emerging Markets Review*, *47*, 100801. doi: 10.1016/j.ememar.2021.100801

Weber, P., Carl, K. V., & Hinz, O. (2023). Applications of explainable artificial intelligence in finance—a systematic review of finance, information systems, and computer science literature. *Management Review Quarterly*, 1–41. doi: 10.1007/s11301-023-00320-0

Weller, A. (2019). Transparency: motivations and challenges. In *Explainable ai: interpreting, explaining and visualizing deep learning* (pp. 23–40). Springer.

Widyawati, L. (2020). A systematic literature review of socially responsible investment and environmental social governance metrics. *Business Strategy and the Environment*, *29*(2), 619–637. doi: 10.1002/bse.2393

Wong, W. C., Batten, J. A., Mohamed-Arshad, S. B., Nordin, S., Adzis, A. A., et al. (2021). Does ESG certification add firm value? *Finance Research Letters*, *39*, 101593. doi: j.frl.2020.101593

Yu, E. P.-y., Guo, C. Q., & Luu, B. V. (2018). Environmental, social and governance transparency and firm value. *Business Strategy and the Environment*, *27*(7), 987–1004. doi: 10.1002/bse.2047

Yu, E. P.-y., Tanda, A., Luu, B. V., & Chai, D. H. (2021). Environmental transparency and investors' risk perception: Cross-country evidence on multinational corporations' sustainability practices and cost of equity. *Business Strategy and the Environment*, *30*(8), 3975–4000. doi: 10.1002/bse.2852

Yu, L., Yue, W., Wang, S., & Lai, K. K. (2010). Support vector machine based multiagent ensemble learning for credit risk evaluation. *Expert Systems with Applications*, *37*(2), 1351–1360.