INDEX

INTRODUCTION.....P.4

1. THEORETICAL MODELS FOR CREDIT RISK CLASSIFICATION AND THE ROLE OF ESG FACTORS.....P. 7

1.1 DEFINITION OF CREDIT RISK…….....P. 7

1.2 THE EVALUATION OF CREDIT RISK…….....P. 9

1.3 THE ALTMAN Z-SCORE MODEL…….....P.12

1.4 MACHINE LEARNING MODELS FOR CREDIT RISK CLASSIFICATION…….....P. 15

1.4.1 INTRODUCTION TO MACHINE LEARNING IN CREDIT RISK MODELLING…….……P.16

1.4.2 REGRESSION AND CLASSIFICATION…….P.18

1.4.3 MODEL TRAINING AND VALIDATION…….....…P.20

1.4.4 REGULARIZATION TECHNIQUES: RIDGE VS LASSO…….P.25

1.4.5 MULTINOMIAL LOGISTIC REGRESSION WITH LASSO PENALIZATION……P.30

1.4.6 LINEAR DISCRIMINANT ANALYSIS……...P.33

1.4.7 CLASSIFICATION TREES……...P.36

1.4.8 RANDOM FOREST MODEL……...P.40

1.4.9 MODEL EVALUATION METRICS…….…P.45

1.4.10 ESG FACTORS IN CREDIT RISK CLASSIFICATION……P.59

2. EMPYRICAL ANALYSIS: DEVELOPMENT AND COMPARISON OF PREDICTIVE MODELS.....P.67

2.1 INTRODUCTION TO THE EMPIRICAL ANALYSIS.....P.67

2.2 EXTRACTION OF THE DATA AND DATA CLEANING…….....P.68

2.3 DEPENDENT VARIABLE – ALTMAN Z - SCORE: CREDIT RISK CLASSES……....P.69

2.4 EXPLANATORY VARIABLES……P.71

2.5 TRAINING – TEST SET SPLIT……P.78

2.6 EXPLORATORY ANALYSIS – CORRELATION MATRIX……...P.80

2.7 DEVELOPMENT OF THE MODELS AND RESULTS……...P.82

2.7.1 MULTINOMIAL LOGISTIC REGRESSION……P.82

2.7.2 LINEAR DISCRIMINANT ANALYSIS……P.90

2.7.3 CLASSIFICATION TREE……..P.94

2.7.4 RANDOM FOREST……...P.98

2.7.5 S.A.F.E. A.I. METRICS……P.102

CONCLUSIONS.....P.109

BIBLIOGRAPHY.....P.111

INTRODUCTION

Credit risk represents one of the main sources of vulnerability for the economic system. It reflects the probability that a firm will be unable to meet its financial obligations and constitutes an important element in assessing corporate solidity and determining the cost of capital. Therefore, understanding and measuring creditworthiness is essential, not only for financial institutions and investors but also for regulators and for companies themselves, for whom it is crucial to ensure the long-term economic sustainability of their activities. Traditionally, the assessment of a firm's creditworthiness has been primarily based on accounting and financial indicators, which summarize management performance and indicate a company's ability to generate profitability and maintain balance sheet stability. However, in today's economic environment, these tools do not always capture all dimensions of risk.

In recent years, the international economic and regulatory landscape has increasingly focused on Environmental, Social, and Governance (ESG) issues, driven by global challenges such as climate change and the green transition. Consequently, the growing attention to corporate sustainability has extended the use of ESG criteria beyond the scope of social responsibility alone: governments, institutional investors, and asset managers have progressively integrated ESG metrics into investment decision-making processes and into the assessment of financial risk and corporate creditworthiness.

Within this context, this analysis aims to evaluate and compare the performance of different machine learning models used to classify the creditworthiness of publicly listed European companies, particularly those included in the EURO STOXX 600 index. The analysis employs the Altman

Z-Score, which is a statistical index developed by E. Altman in 1968 to assess a company's financial health, as a reference measure, here involved as a proxy for credit rating. The goal is to classify this metric into three risk categories: *Safe*, *Grey Zone*, and *Distress*.

In addition to comparing the predictive performance of the classification models employed, particular attention is devoted to the relevance of ESG variables within the models, to understand whether such factors provide meaningful informational value in determining a firm's creditworthiness, beyond traditional accounting and financial indicators.

The empirical analysis is conducted using the R programming language, with monthly data covering the past five years, extracted from the Bloomberg database.

The analysis is structured into three main sections.

1. Theoretical framework: which presents the conceptual background, introducing credit risk, describing the Altman Z-Score metric, and outlining the machine learning models applied in the analysis. It also discusses the growing role of ESG factors and their impact on firms' economic and reputational strength.

2. Empirical analysis: which describes the dataset composition, selected variables, and the methodology developed for classifying Altman's Z-Score. The previously introduced models are implemented and their ability to correctly classify the credit risk level of major European listed companies is evaluated.

3. Results interpretation and conclusions: which presents and interprets the obtained results from an economic and financial perspective. The predictive performance of the models is compared and the importance

and statistical significance of ESG variables in determining creditworthiness is assessed, identifying the factors that contribute most to credit risk classification.

1. THEORETICAL MODELS FOR CREDIT RISK CLASSIFICATION AND THE ROLE OF ESG FACTORS

1.1 DEFINITION OF CREDIT RISK

Credit risk represents the possibility that a debtor, such as a company or an individual, may be unable, in whole or in part, to meet their financial obligations, including the repayment of principal, the payment of any interest, or other contractual debts. However, this risk is not limited to the simple possibility of default; even a deterioration in the creditworthiness of the counterparty should be considered a manifestation of this type of risk. Credit risk is therefore influenced by a variety of factors, which can be related to the specific situation of the debtor or not. For instance, economic and financial health reflects a factor linked to the debtor itself, whereas the overall economic cycle represents an external aspect. It is thus evident that an unfavourable period of economic recession can further deteriorate creditworthiness.

However, it should be noted that risk should not be understood as a purely negative eventuality. High creditworthiness can, in fact, be a sign of financial solidity and a good investment opportunity: one of the fundamental principles of finance is precisely the search for a balance between risk and return, with investors aiming to achieve a certain level of return while minimising their exposure to risk as much as possible. This relationship, known as the *risk-return trade-off*, constitutes the cornerstone of portfolio decisions and the evaluation of investment opportunities.

Traditionally, the assessment of creditworthiness has relied on financial and accounting indicators, such as liquidity ratios (e.g., *Current Ratio*),

profitability metrics (e.g., *Return on Assets (ROA)*, *Return on Equity (ROE)*), and leverage measures, which evaluate the degree of financial leverage and debt sustainability, including the *Debt-to-Equity Ratio* and the *Interest Coverage Ratio*. These indicators provide a concise representation of a firm's ability to generate revenues, manage debt, and maintain balance sheet stability, making them highly informative for determining creditworthiness.

In recent years, however, the concept of credit risk has evolved beyond purely financial dimensions. The growing attention to *Environmental, Social, and Governance* (ESG) factors reflects the recognition that elements such as environmental management, stakeholder relations, and governance quality can significantly impact both reputational risk and a firm's resilience and creditworthiness.

1.2 THE EVALUATION OF CREDIT RISK

Given the fundamental importance of credit risk, its evaluation represents a crucial process for financial institutions, investors, and regulatory authorities, as it allows them to determine a debtor's capability to meet financial obligations. Proper measurement of credit risk is essential not only to ensure the stability of the financial system but also to promote the efficiency of capital markets and the long-term sustainability of firms.

The assessment of credit risk relies on both qualitative and quantitative aspects. Qualitative analyses consider factors such as the quality of management and the board of directors, which are linked to the credibility of future projects and the objectives that the company intends to pursue, the reliability and competitive position of the company, and the prospects for the sector. Quantitative analyses, on the other hand, focus on economic and financial aspects including financial statement analysis, financial health indicators, return on capital, cash flows and the ability to generate resources and income. In addition, quantitative credit risk analyses are also based on the construction of statistical models that estimate the probability of insolvency using financial and balance sheet indicators. A particularly representative example is *Altman's Z-Score*, which will be examined in detail in the following section. Developed by E. Altman in 1968, this statistical index combines several accounting variables to predict the probability of corporate insolvency within a two-year horizon.

A central role in credit risk evaluation is played by credit rating agencies, such as Standard & Poor's, Moody's, and Fitch. These agencies assign assessments of the financial soundness and creditworthiness of companies or countries, expressed through standardized alphabetic or alphanumeric rating scales.

Characterization of debt and issuer (source: Moody's)		Rating			Linear transformations	
		S&P	Moody's	Fitch	Scale 21	Scale 17
Highest quality	Investment grade	AAA	Aaa	AAA	21	17
High quality		AA+	Aa1	AA+	20	16
		AA	Aa2	AA	19	15
		AA-	Aa3	AA-	18	14
Strong payment capacity		A+	A1	A+	17	13
		A	A2	A	16	12
		A-	A3	A-	15	11
Adequate payment capacity		BBB+	Baa1	BBB+	14	10
		BBB	Baa2	BBB	13	9
		BBB-	Baa3	BBB-	12	8
Likely to fulfil obligations, ongoing uncertainty	Speculative grade	BB+	Ba1	BB+	11	7
		BB	Ba2	BB	10	6
		BB-	Ba3	BB-	9	5
High credit risk		B+	B1	B+	8	4
		B	B2	B	7	3
		B-	B3	B-	6	2
Very high credit risk		CCC+	Caa1	CCC+	5	1
		CCC	Caa2	CCC	4	
		CCC-	Caa3	CCC-	3	
Near default with possibility of recovery		CC	Ca	CC	2	
				C		
Default		SD	C	DDD	1	
		D		DD		
				D		

Figure 1: S&P, Moody's and Fitch rating systems and linear transformations. Reference: António Afonso, Pedro Gomes and Philipp Rother (2007) What "Hides" Behind Sovereign Debt Ratings?

Credit risk assessment requires a large amount of both qualitative and quantitative data and information. It does not focus solely on past performance but also considers future expectations for the entity being assessed. Due to their importance, credit ratings exert a direct influence on investor confidence, serving as a key benchmark for financial markets and contributing to the overall stability of the economic system.

Over time, credit risk assessment systems have evolved and become increasingly refined, shifting from traditional approaches based on accounting indicators to more sophisticated, data-driven methodologies. Nowadays, credit risk modelling is based on robust quantitative approaches that use large amounts of data and advanced machine learning techniques. These models apply statistical algorithms of varying degrees of complexity to identify relationships between variables, determine which factors exert the

greatest influence and to what extent, with the goal of classifying each entity into a risk category and, consequently, assigning a credit rating.

In this context, credit risk evaluation is approached through a comparative analysis of several machine learning models for risk classification, using *Altman Z-Score* as a proxy for credit risk. The analysis integrates both financial and ESG variables, with the objective of identifying the model that achieves the *best predictive performance* and of assessing the *significance and informational contribution of ESG factors* in determining creditworthiness of companies.

1.3 THE ALTMAN Z-SCORE MODEL

The Altman Z-Score is one of the most well-known and established methods for assessing corporate bankruptcy risk. Developed in 1968 by economist Edward Altman, the model was designed to provide a statistical tool capable of predicting the probability that a company would fail within two years. Altman's approach is based on Multiple Discriminant Analysis (MDA), a multivariate statistical technique that combines several accounting indicators into a single synthetic index capable of distinguishing between healthy and distressed companies. Altman identified a set of financial variables related to profitability, leverage, liquidity, and activity, which represent the main dimensions of corporate performance, and combined them into a linear model to estimate the probability of default for publicly listed companies.

Here is provided Bloomberg's definition of Altman Z-Score with the related calculation methodology:

VM001 - Z-Score (ALTMAN_Z_SCORE): Indicates the probability of a company filing for bankruptcy within the next two years. The higher the value, the lower the probability of bankruptcy. A score below 1.8 indicates bankruptcy is imminent. A score above 3 indicates bankruptcy is unlikely. Altman's Z-Score is only available on publicly listed companies with all the requisite fundamentals for the model. Calculated as:

Altman's Z-Score = 1.2 * (Working Capital / Tangible Assets) + 1.4 * (Retained Earnings / Tangible Assets) + 3.3 * (EBIT / Tangible Assets) + 0.6 * (Market Value of Equity / Total Liabilities) + (Sales / Tangible Assets)[1]

Over the years, Altman has continued to reevaluate his Z-score. From 1969 until 1975, Altman looked at 86 companies in distress, then 110 from 1976

[1] Bloomberg Terminal (2025)

to 1995, and finally 120 from 1996 to 1999, finding that the Z-score had an accuracy of between 82% and 94%. In 2012, he released an updated version called the Altman Z-score Plus that can be used to evaluate public and private companies, manufacturing and non-manufacturing companies, and U.S. and non-U.S. companies.[2]

Historically, the predictive power of the Z-Score has been evident. In 2007, the credit ratings assigned to certain asset-backed securities were significantly higher than their actual warranted risk levels. However, the Altman Z-Score clearly signalled a sharp deterioration in corporate financial health, indicating that many firms were approaching a potential state of insolvency. According to Altman's estimates, the median Z-Score of companies in 2007 was 1.81, corresponding to a B credit rating. This implied that roughly half of the firms should have received lower ratings, as they were highly vulnerable and faced a substantial risk of bankruptcy. Based on these findings, Altman anticipated the onset of a financial crisis and a forthcoming collapse in the credit market. He initially believed that the turmoil would originate from corporate defaults; however, the crisis that unfolded in 2008 began in the market for mortgage-backed securities (MBS). Nonetheless, by 2009, corporate defaults surged dramatically, reaching the second-highest rate in history.

Thus, investors can use Altman Z-score Plus to evaluate corporate credit risk. A score below 1.8 signals the company is likely headed for bankruptcy, while companies with scores above 3 are not likely to go bankrupt. Investors may consider purchasing a stock if its Altman Z-Score value is closer to 3 and selling, or shorting, a stock if the value is closer to 1.8. According to the

[2]https://www.investopedia.com/terms/a/altman.asp#citation-7

Z-score value the firms can be classified into three categories of risk based on the following thresholds:

- $Z > 2.99 \rightarrow$ *Safe Zone*: company in financial health.
- $1.81 < Z < 2.99 \rightarrow$ *Grey Zone*: area of uncertainty where the company's financial situation is not fully defined.
- $Z < 1.81 \rightarrow$ *Distress Zone*: high probability of insolvency in the short term.

The following section illustrates the machine learning models that have been implemented with the purpose of classifying the level of risk, represented by the Altman Z-Score, associated with listed companies belonging to the EURO STOXX 600 index.

1.4 MACHINE LEARNING MODELS FOR CREDIT RISK CLASSIFICATION

In recent years, the growing availability of financial data and the progress in computational techniques have significantly transformed the methods used to assess corporate credit risk. Machine learning techniques allow for a flexible and data-driven modelling of complex relationships between variables, enhancing the predictive power and adaptability of credit risk analysis. This section introduces the main supervised learning algorithms applied to the classification of credit risk, where the target variable corresponds to the Altman Z-Score risk category (*Safe, Grey Zone, or Distress*). The section begins with the fundamental concepts of machine learning and the difference between regression and classification problems, followed by an explanation of model training, validation, and regularization techniques. The main models considered, Multinomial Logistic Regression (with LASSO penalization), Linear Discriminant Analysis (LDA), Classification Trees and Random Forest are then presented and compared in terms of their theoretical foundations, interpretability, and suitability for credit risk prediction. The section concludes with an overview of the key performance metrics used to evaluate model accuracy and discriminative ability, like ROC Curves and Confusion Matrices, setting the basis for the empirical analysis conducted in the following chapter.

1.4.1 INTRODUCTION TO MACHINE LEARNING IN CREDIT RISK MODELLING

In recent years, the increasing availability of financial data and advances in computational techniques have profoundly transformed the field of credit risk analysis. Machine learning represents one of the most innovative and rapidly developing fields within quantitative finance and risk management. It encompasses a set of statistical and computational techniques that enable models to identify patterns and relationships in data and to make predictions or classifications based on observed information. The growing availability of large and complex datasets and the increasing computational tools have made machine learning an essential component in financial analytics, particularly in areas such as asset pricing, portfolio optimization, fraud detection, and credit risk assessment.

In the context of credit risk modelling, machine learning is very effective to analyse complex relationships among multiple variables. It allows for the systematic integration of a wide range of quantitative and qualitative factors such as financial indicators, market data, and ESG metrics, enhancing the model's ability to capture the multidimensional nature of corporate creditworthiness. By learning from historical patterns and adapting to different types of data, machine learning techniques can improve the accuracy and consistency of credit risk classification and support data-driven decision-making processes.

Machine learning algorithms can be mainly classified into two categories: *supervised* and *unsupervised learning models*.

- In supervised learning, the model is trained on a labelled dataset in which both the explanatory variables (features) and the target variable (label) are known. The goal is to learn a mapping function that can

predict the target variable for new and unseen data. For each observation of the predictor measurements x_i, i = 1,…,n, there is an associated response measurement y_i. We wish to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference).[3] Beyond predictive performance, traditional supervised learning methods offer two significant advantages:

1. *Explainability*: it can be understood why a model makes a prediction, especially vital in regulated fields.

2. *Efficiency*: classical algorithms often require fewer resources and data, making them ideal for small-scale to medium-scale deployments or embedded decision systems.[4]

- Unsupervised learning, by contrast, describes the somewhat more challenging situation in which for every observation i = 1,…,n, we observe a vector of measurements x_i, but no associated response y_i. In this setting we are, in some sense, working blind; the situation is referred to as *unsupervised* because we lack a response variable that can supervise our analysis. A sort of statistical analysis is possible by seeking to understand the relationships between the variables or between the observations.[5] The model analyses data without predefined labels, aiming to uncover hidden structures or clusters.

[3] Gareth James, Trevor Hastie, Daniela Witten, Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R Second Edition,* Springer, 2023, p.26
[4] IBM, https://www.ibm.com/think/topics/classification-vs-regression
[5] Gareth James, Trevor Hastie, Daniela Witten, Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R Second Edition,* Springer, 2023, p.26

This approach is often used for exploratory analysis or to identify homogeneous groups of firms based on financial similarity.

In this analysis the focus is on supervised learning models, where the objective is to predict a *target variable*. In this case, the level of credit risk is classified based on a set of known explanatory variables, both financial and ESG-related. The target labels are derived from the Altman Z-Score, which classifies firms into three risk categories: *Safe*, *Grey Zone*, and *Distress*. Through this approach, the model learns from historical data how different financial and ESG characteristics are associated with each level of credit risk, enabling the classification of new observations. Supervised learning methods are therefore particularly well-suited to credit risk assessment, as they allow the model to generalize from past observations and to provide predictive insights about the likelihood of default or financial distress.

1.4.2 REGRESSION AND CLASSIFICATION

Within supervised learning, predictive models can be divided into two fundamental categories depending on the nature of the target variable: *regression* and *classification*. They rely on labelled data to learn the relationships between input variables (*features*) and output variables (*targets*).

- Regression models aim to predict a continuous variable; the output is quantitative. In the field of credit risk, this might correspond to estimating a company's *probability of default (PD)*, the *numerical value of the Altman Z-Score*, or other continuous measures of financial stability. Regression models are typically evaluated through error-

based metrics such as the *Mean Squared Error (MSE), Root Mean Squared Error (RMSE),* or *R-squared (R^2)*, which measure the deviation between predicted and actual values.

- Classification models, on the other hand, are designed to predict discrete categories; here the output is qualitative. Each observation is assigned to one of several predefined classes according to its characteristics. In credit risk applications, classification models are used to categorize companies into distinct levels of risk. Classification models are typically assessed through *Accuracy*, the proportion of correctly classified observations, *Confusion Matrices*, which consist of tabular representation of predicted versus actual classes and *ROC Curves*, to evaluate overall discrimination ability and predictive quality.

The choice between regression and classification depends on the research objective and the characteristics of the response variable. Since the aim of this analysis is not to estimate the continuous value of the Z-Score but rather to identify the risk class associated with each firm, the classification approach is the most appropriate. This formulation allows for a direct comparison between different machine learning algorithms in terms of their accuracy in predicting categorical risk levels. By directly learning from labelled data (Z-Score thresholds), these models allow a consistent comparison of machine learning algorithms based on their ability to accurately classify corporate credit risk.

1.4.3 MODEL TRAINING AND VALIDATION

A crucial phase in developing machine learning models is the training and validation process, which ensures that the model not only fits the available data but can also generalize effectively to unseen observations. To achieve this goal, the dataset is divided into two subsets: a *training set*, used to estimate the model parameters, and a *test set*, used to evaluate predictive performance on new data. This separation helps to detect issues of *overfitting*, a common problem where the model performs well on training data but fails to generalize to new cases.

Model performance is often evaluated using the *Mean Squared Error (MSE)* for regression models and *accuracy or loss metrics* for classification models, computed separately on the training and test sets. Typically, as model flexibility increases, the *training MSE* (or equivalently, the *training loss*) decreases monotonically, while the *test MSE (or test loss)* follows a U-shaped pattern. As model flexibility increases, the training MSE will decrease, but the test MSE may not.[6] Initially, both improve as the model captures meaningful patterns in the data, but beyond a certain point, the model starts to fit random noise rather than the true underlying data structure. Overfit models typically have very low error rates on the training data but significantly higher ones on a separate test dataset. In this situation, the algorithm has become too specialized in the training data, identifying spurious relationships that do not generalize to unseen observations. As a result, although the model performs very well on known data, its predictive reliability deteriorates on new firms. The main indicators of overfitting are the following:

[6] Gareth James, Trevor Hastie, Daniela Witten, Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R Second Edition,* Springer, 2023, p.32

1. Excellent performance on training data but fails the test data: this is the classic sign of overfitting. The model achieves a great performance on training data but fails on the test data.

2. Huge gap between training and test error: if training error is very low, but test error is way high.

3. Overly sensitive to change: overfit models often have a ton of parameters and features, making them overly complex and sensitive to any tiny change in the data.[7]

The opposite situation is *underfitting,* the key characteristics are the following:

1. Missing the point: underfit models fail to grasp the relationships between features and target variables.

2. Too much simplicity: when the model is too simple, it won't be able to capture the complexity of the data. The model will underfit, causing it to perform poorly.

3. Bad performance all around: unlike overfitting, where the model at least does well on the training data, underfitting leads to poor performance on both the training and test data.

4. High error rates: both training and test errors will be high, indicating that the model isn't capturing the underlying patterns in the data.

[7] https://www.cudocompute.com/blog/overfitting-and-underfitting-in-machine-learning-causes-indicators-and-how

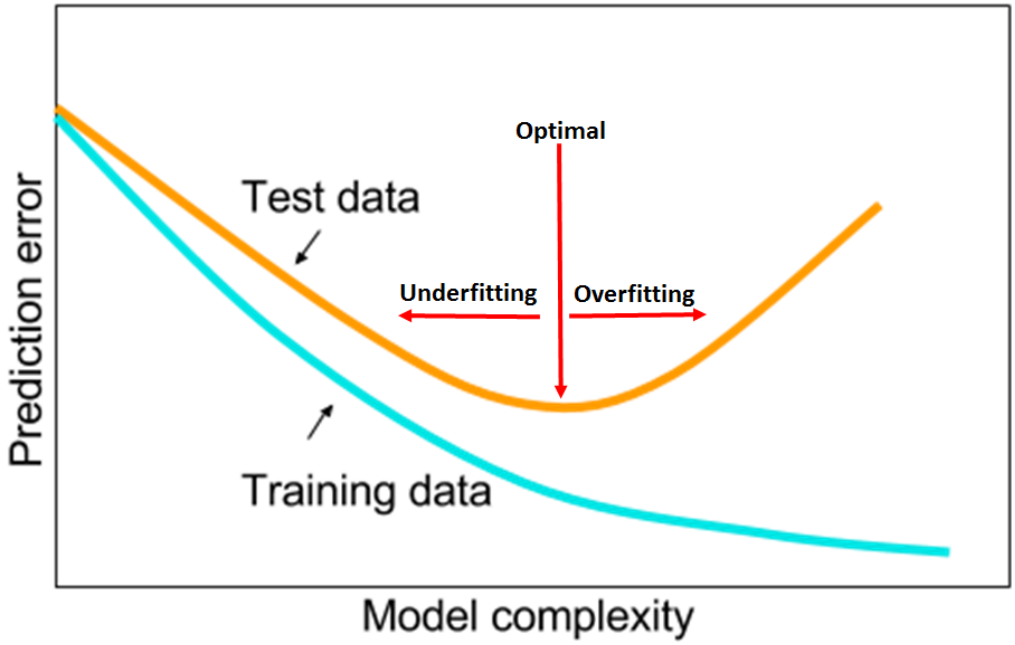

Figure 2: Pictorial explanation of the trade-off between underfitting and overfitting

There are several procedures that can be useful to address these problems, underfitting can be addressed by increasing model complexity, adding more features and parameters. While overfitting, by contrast, can be addressed by simplifying the features and reducing the parameters of the model, but more importantly by applying *resampling methods*, like *Cross-validation* and *regularization techniques*, which will be illustrated as follows.

Cross-validation is a resampling technique used to assess how the results of a statistical model will generalize to an independent dataset. The most common form, *k-fold cross-validation*, divides the dataset into k equal parts (or "folds"): the model is trained on k–1 folds and tested on the remaining one, repeating the process k times so that each subset serves once as a validation set. The average performance across all folds provides a reliable estimate of the model's generalization ability.

Cross-validation can be applied both to regression and classification models. For quantitative response variable problems, the Mean Squared Error, MSE_1, is computed on the observations in the held-out fold. This procedure is repeated k times; each time, a different group of observations is treated as a validation set. This process results in k estimates of the test error, MSE_1,MSE_2,...,MSE_k. The k-fold CV estimate is computed by averaging these values: $CV(k) = \frac{1}{k}\sum_{i=1}^{k} MSE_i$.

For classification problems, where the outcome Y is qualitative, Cross-validation works just as described earlier, except that rather than using MSE to quantify test error, we instead use the number of misclassified observations.[8]

Cross-validation is also essential for *hyperparameter tuning*, that is, selecting the optimal values of model parameters that are not directly learned from the data.

For example, in regularized regression models such as LASSO (Least Absolute Shrinkage and Selection Operator), which will be explored further in the next section, Cross-validation is used to determine the ideal value of a parameter. This usage will be implemented in the Multinomial Logistic Regression model to evaluate the optimal value of the parameter λ involved in the LASSO penalization, which controls the degree of coefficient shrinkage and variable selection.

Resampling and Cross-validation are tied to an important conceptual framework which is the *bias-variance trade-off*. Models with high complexity may fit training data very accurately (low bias) but exhibit poor

[8] Gareth James, Trevor Hastie, Daniela Witten, Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R Second Edition,* Springer, 2023, p.203-206

performance on new data (high variance), leading to overfitting. Conversely, overly simple models may generalize well (low variance), but fail to capture underlying relationships (high bias), leading to underfitting in this case. Therefore, finding the right balance between model complexity, bias and variance is crucial for achieving good generalisation results. Cross-validation helps mitigate this trade-off by providing a systematic way to assess how different model configurations perform across multiple data partitions.

In this analysis, model validation is performed using a time-based train–test split. The dataset covers the period from October 2020 to September 2025, with data up to December 2024 used for training and the 2025 data reserved to the test set. This temporal division ensures that the models are evaluated on genuinely unseen future data, reflecting a realistic forecasting scenario and avoiding information leakage from the test period into the training phase. In addition, *k-fold cross-validation* (with k = 10) is applied within the training set to fine-tune hyperparameter λ involved in the LASSO penalization applied to the Multinomial Logistic Regression model to optimize the bias-variance trade-off. There is a bias-variance trade-off associated with the choice of k in k-fold cross-validation. Typically, k-fold cross-validation is performed using k = 5 or k = 10, as these values have been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance.[9] This combined approach helps to prevent overfitting and to ensure the reliability and generalizability of the results.

[9] Gareth James, Trevor Hastie, Daniela Witten, Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R Second Edition,* Springer, 2023, p.206

1.4.4 REGULARIZATION TECHNIQUES: RIDGE VS LASSO

Regularization is a fundamental technique in machine learning and statistical modelling used to prevent overfitting, improving the model's ability to generalize to new data. The core idea behind regularization is to introduce a penalty term into the loss function, discouraging the model from assigning excessively large weights to the predictors. By constraining the magnitude of the coefficients, regularization helps to control model complexity, reducing variance without significantly increasing bias.

Mathematically, regularization modifies the standard loss function (such as the Mean Squared Error or the negative log-likelihood) by adding a penalty term that depends on the values of the coefficients of the model parameters. The general form of the *regularized loss function* can be written as:

$$L(\beta) = \text{Loss}(\beta) + \lambda\, P(\beta)$$

where $\lambda \geq 0$ is the regularization parameter, which controls the strength of the penalty, and $P(\beta)$ represents the penalty function applied to the model coefficients. Two of the most used regularization methods are Ridge regularization (L2 penalty) and LASSO regularization (L1 penalty).

- Ridge regularization (L2 penalty) adds the *squared magnitude* of the coefficients as a penalty term:

$$P(\beta) = \sum_{j=1}^{p} \beta_j^2$$

Ridge regularization specifically mitigates the effects of *multicollinearity* in regression analysis. This is useful when machine learning models with a large number of parameters are developed, particularly if those parameters also

have high weights. Ridge regularization can be applied both to linear and logistic regression.

Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated, meaning they contain overlapping information about the response variable. In such cases, the estimates of the regression coefficients β become unstable and highly sensitive to small changes in the data. This instability leads to large variances in the estimated coefficients, making the model unreliable and difficult to interpret.[10]

The L2 regularization term, $\lambda \sum_{j=1}^{p} \beta_j^2$, shrinks the coefficients toward zero, but never exactly to zero. This shrinkage reduces the variance of the estimates and mitigates the effects of multicollinearity, leading to more stable and robust parameter estimates even when predictors are highly correlated. The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates. When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates. However, as $\lambda \rightarrow \infty$, the impact of the shrinkage penalty grows and the ridge regression coefficient estimates will approach zero. Unlike least squares, which generates only one set of coefficient estimates, ridge regression will produce a different set of coefficient estimates, $\widehat{\beta_\lambda^R}$, for each value of λ. Selecting a good value for λ is critical, it can be found by applying Cross-validation, as explained in previous paragraph.

[10] https://www.ibm.com/it-it/think/topics/multicollinearity

Specifically, the ridge regression coefficient estimates $\widehat{\beta^R}$ are the values that minimize:[11]

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2.$$

- LASSO regularization (L1), instead, introduces a penalty based on the *absolute values* of the regression coefficients. The penalty term is defined as:

$$P(\beta) = \sum_{j=1}^{p} | \beta_j |$$

The key feature of LASSO regularization is its ability to perform both shrinkage and variable selection simultaneously. Like Ridge regression, LASSO aims to prevent overfitting by constraining the magnitude of the coefficients, however, it can drive some of them *exactly to zero*. This property induces *sparsity* in the model, effectively excluding non-informative or redundant predictors from the final specification. As a result, LASSO not only improves model generalization but also enhances interpretability, since only the most relevant variables remain active. The L1 regularization term, $\lambda \sum_{j=1}^{p} | \beta_j |$, serves to control the degree of shrinkage applied to the coefficients. When $\lambda = 0$, the penalty has no effect and the model reduces to the standard least squares estimation. As λ increases, the penalty grows stronger, shrinking the coefficients toward zero and setting some of them exactly equal to zero when their contribution to the model's predictive power is minimal. This makes LASSO

[11] Gareth James, Trevor Hastie, Daniela Witten, Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R Second Edition,* Springer, 2023, p.237-238

particularly suitable when dealing with high-dimensional datasets or with correlated predictors, where feature selection becomes essential to reduce redundancy and avoid overfitting. The LASSO coefficients, $\widehat{\beta_\lambda^L}$, minimize the quantity:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p} | \beta_j | = RSS + \lambda\sum_{j=1}^{p} | \beta_j |$$

We can see that the LASSO and ridge regression have similar formulations. The only difference is that the β_j^2 term in the ridge regression penalty has been replaced by $| \beta_j |$ in the LASSO penalty. In statistical parlance, the LASSO uses an L1 penalty instead of an L2 penalty. The L1 norm of a coefficient vector is given by: $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$; while the L2 norm is given by: $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p}\beta_j^2}$.[12] The tuning parameter λ again plays a central role: a small λ yields a model close to the unpenalized regression, while a large λ enforces stronger regularization, increasing sparsity. The optimal value of λ is typically determined through k-fold Cross-validation, as previously described, by selecting the parameter that minimizes the validation error.

Thus, LASSO regularization is ideal for predictive modelling problems; its ability to perform automatic variable selection can simplify models and improve predictive accuracy. This feature is especially useful in *handling high-dimensional datasets*: a dataset is considered high-dimensional when the number of predictor variables is much greater than the number of observations. LASSO regression

[12] Gareth James, Trevor Hastie, Daniela Witten, Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R Second Edition,* Springer, 2023, p.237-242

can help reduce dimensionality within such datasets by shrinking some coefficient weights to zero and eliminating least important predictors from the model. [13]

LASSO regularization is applied within the Multinomial Logistic Regression (MLR) model to control model complexity and enhance feature selection. The use of L1 penalization allows the model to automatically exclude variables with negligible predictive power, yielding a parsimonious and interpretable model structure. Moreover, the regularization strength, which depends on λ, is optimized via 10-fold Cross-validation, ensuring an effective balance between bias and variance and improving the model's ability to generalize to unseen data. The following sections provide a detailed overview of the classification models employed in this analysis, outlining their theoretical foundation and practical relevance to credit risk prediction.

[13] https://www.ibm.com/it-it/think/topics/lasso-regression

1.4.5 MULTINOMIAL LOGISTIC REGRESSION WITH LASSO PENALIZATION

The first model employed in the analysis is the Multinomial Logistic Regression (MLR), a widely used technique for multi-class classification problems. Since the response variable (Altman Z-Score class) is categorical, the most appropriate modelling framework is classification, which aims to assign each observation to one of several predefined risk categories. Because the problem involves a qualitative dependent variable, a traditional linear regression model would not be suitable. Linear regression assumes a continuous outcome and can produce predicted values outside the [0, 1] range, making its estimates meaningless for categorical responses. In contrast, logistic regression is specifically designed to model the probability of class membership, $Pr(Y \mid X)$, by linking the predictors to the log-odds of belonging to a particular category. When extended to the multinomial setting, logistic regression becomes capable of handling multiple risk classes simultaneously, making it a natural and interpretable choice for credit risk assessment based on the Altman Z-Score classification.

Logistic regression is a probabilistic classification model commonly used when the dependent variable is categorical. It models the probability of class membership through the *log-odds* (or logit) transformation, linking the linear combination of predictors to the probability that an observation belongs to a given class. For a binary outcome $Y \in \{0,1\}$, the model can be expressed as:

$$log\left(\frac{P(Y = 1 \mid X)}{P(Y = 0 \mid X)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

This formulation estimates the log-odds of being in the positive class as a linear function of the predictors. When a predictor X_j increases by one unit, log odds change by the value of the associated β_j coefficient.

When the response variable includes more than two categories, as in the current case of credit risk classification into three risk categories (Safe, Grey Zone, Distress), the model can be generalized to the Multinomial Logistic Regression. The model parameters are estimated by maximizing the likelihood of the observed class memberships, which corresponds to minimizing the negative log-likelihood function. In this setting, the probability of an observation i belonging to class k is modelled using the *softmax* function:

$$P(Y_i = k \mid X_i) = \frac{\exp\left(\beta_{0k} + X_i'\beta_k\right)}{\sum_{l=1}^{K} \exp\left(\beta_{0l} + X_i'\beta_l\right)}, \text{for } k = 1, \ldots, K$$

where K is the number of categories and β_k represents the vector of coefficients associated with class k. The coefficients β_k represent the change in the log-odds of belonging to class k relative to the other categories for a one-unit increase in the corresponding predictor, holding all other variables constant. The softmax function can be viewed as a generalization of the sigmoid function used in binary logistic regression, extending it to the case of K mutually exclusive classes. This probabilistic formulation ensures that all class probabilities are non-negative and sum to one. Softmax coding is widely used in machine learning applications, rather than selecting a baseline class, all K classes are treated symmetrically, thus, rather than estimating coefficients for $K - 1$ classes, coefficients are estimated for all K classes.[14] To prevent overfitting and handle potential multicollinearity among predictors, a regularization term is introduced into the loss function, which is constituted by the negative log-likelihood and the penalty term. LASSO

[14] Gareth James, Trevor Hastie, Daniela Witten, Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R Second Edition,* Springer, 2023, p.141

regularization (L1 norm) is applied, which adds a penalty proportional to the absolute values of the coefficients:

$$Loss\ (\beta) = -\ (1/N) \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik}\ log\ P(Y_i = k|X_i) + \lambda \sum_{j=1}^{p} |\beta_j|$$

where y_{ik} is an indicator variable that equals 1 if observation i belongs to class k, and λ is the tuning parameter controlling the degree of penalization. This L1 penalty term encourages sparsity by shrinking some coefficients exactly to zero, effectively performing automatic variable selection within the multinomial context.

Given the large number of financial and ESG predictors, LASSO regularization was preferred over Ridge, as it not only mitigates the effects of multicollinearity but also simplifies the model by selecting the most relevant features, enhancing interpretability. The optimal value of λ is determined through Cross-validation, ensuring a balance between model complexity and predictive performance, as discussed in Section 1.4.3.

From an interpretative perspective, the Multinomial Logistic Regression with LASSO penalization provides a robust framework for credit risk analysis, as it combines the probabilistic interpretability of logistic models with the feature selection capability of LASSO. This approach allows identifying the most influential financial and ESG factors associated with each risk category, improving both predictive accuracy and explainability which are crucial elements in financial decision-making.

1.4.6 LINEAR DISCRIMINANT ANALYSIS

The second model employed is the Linear Discriminant Analysis (LDA). While the Multinomial Logistic Regression involves directly modelling $P(Y_i = k \mid X_i)$ using the softmax function $\frac{\exp(\beta_{0k}+X_i'\beta_k)}{\sum_{l=1}^{K}\exp(\beta_{0l}+X_i'\beta_l)}$, a multiclass generalization of the logit function, $log\left(\frac{P(Y=1|X)}{P(Y=0|X)}\right) = \beta_0 + \sum_{j=1}^{p}\beta_j X_j$, which is used for the case of two response classes; Linear Discriminant Analysis (LDA) is based on Fisher's linear discriminant, a statistical method developed by Sir Ronald Fisher in the 1930s and later simplified by C. R. Rao as a multi-class version. Fisher's method reduces dimensions by separating classes of projected data. Separation means maximizing the distance between the projected means and minimizing the projected variance within classes. LDA works by identifying a linear combination of predictors, known as discriminant functions, that separates or characterizes two or more classes. LDA does this by projecting data with two or more dimensions into one dimension so that it can be more easily classified. The technique is, therefore, sometimes referred to as dimensionality reduction. This versatility ensures that LDA can be used for multi-class data classification problems.[15] Unlike logistic regression model, which models the conditional probability directly, LDA assumes a generative model for the predictors and then applies Bayes' theorem to infer class membership probabilities.

This represents a less direct approach to estimate the posterior probabilities. From modelling the conditional distribution of the response *Y* given the predictors *X*, this new approach models the distribution of the predictors *X* separately in each of the three response classes, for each value of *Y*, then

[15] https://www.ibm.com/think/topics/linear-discriminant-analysis

estimates for posterior probabilities $P(Y_i = k \mid X_i)$ are computed applying the Bayes' theorem:

$$P(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

where π_k is the prior probability of class k, and $f_k(x)$ is the multivariate normal density function for that class.

In this application, where there are multiple predictors, we will assume that $X = (X_1, X_2, \ldots, X_p)$ is drawn from a multivariate Gaussian (or normal) distribution, with a class-specific mean vector and a common covariance matrix. This distribution assumes that each individual predictor follows a one-dimensional normal distribution, $X \sim \mathcal{N}(\mu_k, \sigma_k^2)$, the Gaussian density is defined as:

$$f_k(x \mid \mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)$$

where μ_k and σ_k^2 are the mean and the variance parameters for the k^{th} class, with some correlation between each pair of predictors. In the multivariate case there is a p-dimensional random variable *X,* where p is the number of predictors. Formally, $X \sim \mathcal{N}(\mu, \Sigma)$, μ is the mean of the vector *X*, containing the *p* predictors, and Σ is the *p x p* covariance matrix of *X*, the multivariate Gaussian density is defined as:

$$f(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)' \Sigma^{-1} (x-\mu)\right)$$

The LDA classifier assumes that the k^{th} class observations follow a multivariate Gaussian distribution $\mathcal{N}(\mu_k, \Sigma)$, where μ_k is a class-specific mean vector and Σ is a covariance matrix that is common to all classes. By

plugging the density function for the k^{th} class, $f_k(X = x)$, into the Bayes formula the *discriminant function* can be obtained, the Bayes classifier assigns an observation $X = x$ to the class for which the discriminant function is largest. The discriminant function for the multivariate case is the following:

$$\delta_k(x) = x' \, \Sigma^{-1} \, \mu_k - ½ \, \mu_k' \, \Sigma^{-1} \, \mu_k + log \, \pi_k$$

where μ_k is the vector of predictor means for class k, Σ is the common covariance matrix shared across all classes, π_k is prior probability of class k, x is the vector of observed predictor values.[16]

The model is called Linear Discriminant Analysis because the discriminant function, $\delta_k(x)$, is a *linear function of the predictor variables* x. This linearity arises from the assumption that all classes share the same covariance matrix Σ, which leads to linear decision boundaries between classes.

The key advantages of applying Linear Discriminant Analysis are the following:

1. Simplicity and efficiency of computation: LDA is a simple yet powerful algorithm. It's relatively easy to understand and implement, its efficient computation ensures quick results.

2. Handle multicollinearity: LDA can address multicollinearity, the presence of high correlations between different features. It transforms the data into a lower-dimensional space while maintaining information integrity.[17]

[16] Gareth James, Trevor Hastie, Daniela Witten, Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R Second Edition,* Springer, 2023, p.146

[17] https://www.ibm.com/think/topics/linear-discriminant-analysis

However, LDA also has limitations, since it constructs linear boundaries between classes, its flexibility is limited when the true class separation is nonlinear or when the assumptions of normality and equal covariance matrices are strongly violated. Furthermore, LDA is not suitable for unlabelled data: since it is applied as a supervised learning algorithm, it classifies or separates labelled data. In contrast, for unlabelled data, others dimensionality reduction techniques, which ignore class labels and preserves variance, can instead be applied.

1.4.7 CLASSIFICATION TREES

After discussing linear models, Multinomial Logistic Regression and Linear Discriminant Analysis, there is a transition to tree-based methods application, a fundamentally different class of models. A Decision Tree classifier creates an upside-down tree to make predictions, starting at the top with a question about an important feature in data, then branches out based on the answers.[18] A Classification Tree is a non-parametric supervised learning method that recursively partitions the feature space into a set of homogeneous and disjoint regions. Unlike parametric models such as logistic regression or LDA, classification trees do not assume any specific functional form between predictors and the response variable. Instead, they learn the structure of the data by performing recursive hierarchical binary splits in the predictor space to form distinct and homogeneous regions. This flexibility allows trees to naturally capture *non-linear relationships* and *high-order interactions* among variables, which linear models cannot easily accommodate. A classification tree can be visualized as a flowchart-like structure, where each internal node

[18] https://medium.com/data-science/decision-tree-classifier-explained-a-visual-guide-with-code-examples-for-beginners-7c863f06a71e

represents a test on a predictor variable, each branch denotes the outcome of that test, and each terminal node (leaf) corresponds to a predicted class. Thus, it can be defined as a machine learning technique that classifies features iteratively, based on specific criteria, consisting of a root node, internal nodes and leaf nodes.[19]

Specifically, a classification tree predicts the class of an observation by following a sequence of decision rules based on the predictor variables. Formally, the feature space is partitioned into J distinct and non-overlapping regions $R_1, R_2, \dots, R_J$, and for an observation falling in region R_m, the model assigns as predicted class the most frequent class label within that region:

$$\hat{y}_i = argmax_k \widehat{p_{mk}}$$

where $\widehat{p_{mk}}$ is the proportion of training observations in region R_m belonging to class k.

This recursive partitioning is performed through *binary splitting*, where at each step the algorithm selects a predictor and a split point that best separates the classes, creating subsets that are as homogeneous as possible. The splitting process continues until a stopping criterion is met, such as a minimum node size or maximum tree depth. To determine the best splits, trees rely on *impurity measures*, which quantify the heterogeneity of the classes within a node. The simplest impurity measure is the *misclassification error*:

$$E_m = 1 - \max_k(\widehat{p_{mk}})$$

which represents the proportion of observations in node m that do not belong to the majority class. While intuitive, misclassification error is less sensitive

[19] https://www.sciencedirect.com/topics/computer-science/decision-tree-classifier

to changes in class proportions, so other measures are often preferred when growing the tree, the *Gini index* is one of the most widely used criteria:

$$G_m = 1 - \sum_{k=1}^{K} (\widehat{p_{mk}})^2$$

It quantifies the probability that a randomly chosen observation would be incorrectly classified if labelled according to the class distribution in node *m*. Another option is *cross-entropy* (or *deviance*):

$$D_m = -\sum_{k=1}^{K} \widehat{p_{mk}} \log(\widehat{p_{mk}})$$

which, like the Gini index, captures node purity but gives more weight to low-probability classes. At each step, the algorithm chooses the split that *maximally reduces impurity*, a process known as *recursive binary splitting*, as previously mentioned.

Although a fully grown tree can perfectly classify training data, it usually presents overfitting, since it captures noise rather than meaningful patterns, leading to poor test set performance. A simpler tree, with fewer splits, might lead to lower variance and better interpretation at the cost of a little bias, thus, to improve generalization, *tree pruning* is applied. In cost-complexity pruning, a regularization term α penalizes tree size:

$$C_\alpha(T) = \sum_{m=1}^{|T|} Q_m(T) + \alpha|T|$$

This is the cost-complexity function where: $Q_m(T)$ represents the impurity measure in node m of the tree *T*, α is the *complexity (tuning) parameter*, that balances model fit and simplicity. $|T|$ is the number of terminal (leaf) nodes.

As α increases, the penalty for tree size becomes stronger, encouraging simpler trees with fewer terminal nodes and reducing overfitting. The optimal tree size is typically selected via k-fold Cross-validation, which evaluates performance for trees of different sizes and identifies the one that minimizes *cross-validated deviance*. This regularization process mirrors the role of the LASSO penalty in the Multinomial Logistic Regression, as both methods aim to control model complexity and avoid overfitting while preserving predictive performance.[20]

Classification trees offer several advantages: they are highly *interpretable*, as the tree structure resembles human decision-making and produces intuitive rules. They naturally handle *non-linear relationships* and *interactions* between predictors, and they can accommodate *mixed data types*, numerical and categorical predictors. Trees are also relatively *robust to outliers and missing values*, as splits depend only on the ordering of observations rather than exact values. However, trees also have limitations. They tend to be *high-variance models*, where small changes in the data can produce very different trees. Additionally, while trees capture global non-linearities, their decision boundaries are *piecewise constant*, which can limit smoothness and flexibility in classification.[21]

Hence, Classification trees provide an intuitive and flexible framework for multi-class credit risk assessment. By recursively partitioning the feature space, they uncover complex, non-linear relationships and interactions that linear models cannot easily detect. At the same time, pruning and Cross-validation ensure that these models generalize well to unseen data, achieving a balance between interpretability and predictive accuracy.

[20] Gareth James, Trevor Hastie, Daniela Witten, Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R Second Edition,* Springer, 2023, p. 335-336
[21] https://www.ibm.com/think/topics/decision-trees

1.4.8 RANDOM FOREST MODEL

After introducing single-tree models, there is a progression to *ensemble methods*, which aim to improve predictive accuracy and stability by combining the outputs of multiple base models. One of the earliest and most influential ensemble techniques is *bootstrap aggregation*, or *Bagging*, a resampling-based method designed to reduce model variance. Bagging works by generating multiple training datasets through *bootstrap sampling* which consists of repeatedly drawing samples *with replacement* from the original dataset, each of the same size as the training set. Each sample, thus, represents a slightly different version of the data, on which an individual decision tree is trained. The predictions produced by the individual trees must then be aggregated to obtain the final model output.
In regression problems, this is done by averaging the numerical predictions across all trees, which reduces variance and stabilizes the result.
In classification problems, each tree votes for a class label and Bagging assigns to each observation the class receiving the majority of votes among all trees. This aggregation process stabilizes the predictions and mitigates the tendency of single trees to overfit, as the variability of individual models is averaged out. Building on this idea, Random Forests extend the bagging framework by introducing an additional layer of randomness: at each split within a tree, a random subset *m* of *p* predictors is selected, typically m = $\sqrt{p}$ is chosen, so the number of predictors considered at each split is approximately equal to the square root of the total number of predictors and only those features are considered for the splitting rule.[22] This random feature selection *decorrelates* the trees in the forest, addressing the bias-variance trade-off, ensuring that no single strong predictor dominates the model, and

[22] Gareth James, Trevor Hastie, Daniela Witten, Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R Second Edition,* Springer, 2023, p.343

further enhances generalization performance.[23] As a result, Random Forests combine the strengths of Bagging and Decision Trees to produce an ensemble model that is both powerful and robust, capable of capturing complex, non-linear relationships in the data. Mathematically, suppose that $\hat{f}^{(b)}(x)$ is the prediction from the bth tree, trained on a bootstrap sample of the data. The Random Forest prediction for a new observation x is given by:

$$\widehat{f_{RF}}(x) = \frac{1}{B}\sum_{b=1}^{B} \widehat{f^{(b)}}(x)$$

where B is the total number of trees in the ensemble. For classification problems, the final predicted class corresponds to the majority vote among all individual tree predictions. This averaging (or voting) reduces the model's variance without increasing bias, providing smoother and more generalizable decision boundaries.

From a theoretical standpoint, Random Forests improve predictive accuracy by simultaneously increasing the strength of individual learners and reducing their correlation. Leo Breiman formalized this idea by showing that the generalization error of the Random Forest is bounded by:

$$\text{Generalization Error} \leq \frac{\rho(1-s^2)}{s^2}$$

where s denotes the strength in terms of predictive accuracy of the individual trees and ρ the average correlation among them. Reducing ρ through random feature selection is therefore fundamental to improving ensemble performance.[24] The expected test error of a Random Forest can be decomposed as a function of the strength of individual trees and their

[23] https://www.sciencedirect.com/topics/computer-science/random-forest-classifier
[24] Leo Breiman, *Random Forests*. *Machine Learning*, Springer (Kluwer Academic Publishers), 2001.

correlation: reducing correlation among trees while maintaining strong predictive ability for each tree leads to better ensemble performance.

In practical terms, Random Forests offer several advantages: they handle high-dimensional data, non-linear relationships, and interactions automatically; they are less sensitive to outliers and provide internal estimates of prediction error through the *out-of-bag (OOB) samples*, which serve a similar role to Cross-validation. More precisely, the OOB mechanism arises directly from the bootstrap sampling procedure used to construct each tree. In every bootstrap sample, on average only about two-thirds of the original observations are selected, the remaining one-third are *left out* for that tree. These excluded observations constitute the out-of-bag sample. For each observation, the Random Forest aggregates predictions coming only from the trees for which that observation was OOB, producing an unbiased estimate of its prediction error. Aggregating these OOB errors across all observations yields an internal and cross-validated assessment of model performance without the need for a separate validation set. This OOB framework is not only useful for model evaluation but also supports the computation of *variable importance measures*. Two main families of importance metrics are commonly used. The first is the *Mean Decrease Impurity (MDI)*, which sums the weighted impurity decreases (typically based on the Gini index) for all nodes in which a feature is used for splitting, averaged across all trees. The second is the *Permutation Importance*, or *Mean Decrease Accuracy (MDA)*, which relies directly on OOB samples: the values of a predictor are randomly permuted in the OOB set, and the resulting increase in prediction error quantifies its importance. Both methods provide rankings of relative variable importance, though their results may differ due to the random feature selection at each split and bootstrap sampling used to construct the trees. By permuting a predictor in the OOB samples and measuring the corresponding

increase in prediction error, Random Forests quantify the extent to which each variable contributes to predictive accuracy.

This property is especially valuable in credit risk modelling, where interpretability and understanding feature relevance are crucial. In the context of this analysis, variable importance scores are used to assess which financial indicators and which ESG dimensions (environmental, social, governance subcomponents and other ESG related scores) play the most significant role in predicting credit risk outcomes. This enables a transparent comparison between traditional financial determinants and sustainability-related factors, helping determine whether ESG information offers incremental predictive value beyond conventional credit risk drivers.

Random Forests offer several notable advantages that make them an excellent choice for predictive modelling, including credit risk assessment:

1. Reduced Risk of Overfitting: single decision trees often overfit the training data, capturing noise as if it were signal. In contrast, Random Forests build multiple decorrelated trees and aggregate their predictions, which lowers the overall variance and stabilizes the model, making overfitting much less likely.

2. Flexibility and Robustness: Random Forests can handle both regression and classification tasks with high accuracy. They naturally accommodate mixed data types, outliers, and missing values, and can automatically capture complex non-linear relationships and interactions between variables.

3. Feature Importance Assessment: a key strength of Random Forests is the ability to quantify variable importance, helping identify which features contribute most to prediction accuracy. Techniques such as

Mean Decrease Impurity (MDI) and Permutation Importance (Mean Decrease Accuracy, MDA) provide rankings of predictor relevance. This property is particularly valuable in credit risk modelling, where understanding whether financial metrics or ESG dimensions to drive outcomes is critical for interpretability.[25]

4. Minimal Preprocessing Requirements: Random Forests generally require little data preparation. They perform well without extensive normalization or encoding, and are robust to missing values and noise, making them practical for large-scale, real-world datasets.

Despite these advantages, several limitations must be considered:

1. Computationally Intensive: training a Random Forest can be very resource demanding, especially with large datasets or a high number of trees. Each tree must be grown independently, and predictions require aggregating outputs across all trees, which increases processing time.

2. Limited Interpretability: while overall feature importance can be assessed, the individual decision-making process is obscured compared to a single tree. This reduced transparency can be a challenge in applications where explaining specific predictions is necessary.

3. Higher Memory and Storage Requirements: because Random Forests maintain multiple deep trees in memory, they require more computational resources for storage and processing, particularly for large-scale or high-dimensional datasets.

[25] https://www.ibm.com/think/topics/random-forest#684929713

4. Slower Prediction Speed: in real-time applications, generating predictions requires passing the input through all trees and combining results, making Random Forests slower than simpler models such as single decision trees or linear models.[26]

Overall, the strengths of Random Forests, especially their robustness, predictive accuracy, and ability to evaluate variable importance, generally outweigh the drawbacks, making them a powerful tool in complex modelling tasks, including multiclass credit risk assessment where financial and ESG features are simultaneously considered.

1.4.9 MODEL EVALUATION METRICS

This section is devoted to explaining the evaluation metrics used to assess the predictive performance of the models involved in the analysis. Evaluating the performance of a classification model requires a comprehensive set of metrics capable of capturing not only overall prediction accuracy but also the distribution of different types of errors. This is particularly important in credit risk classification, where misclassifying a risky firm as safe may carry far more severe consequences than the opposite. Since the empirical analysis involves three credit risk classes, each model will be evaluated through a *One-vs-All* strategy, in which each class is assessed against the remaining two. This approach is standard when extending binary evaluation metrics to multiclass settings and allows a clearer understanding of how well each model distinguishes each rating category individually.

[26] https://medium.com/data-science/random-forest-explained-a-visual-guide-with-code-examples-9f736a6e1b3c

The metrics employed to evaluate the predictive performance of all models include *Accuracy and Classification Loss*, the *Confusion Matrix* and its derived indicators, and finally *ROC curve*s and the *Area Under the Curve (AUC)* as measures of discriminative ability.

Accuracy is the most immediate measure of classification performance. It represents the proportion of correctly classified observations out of the total and offers a simple snapshot of how well the model performs overall. A high accuracy score indicates that the model is making a large proportion of correct predictions, whereas a low score indicates that the model is making too many incorrect predictions. It is defined as:

$$Accuracy = \frac{Total\ Number\ of\ Observations}{Correct\ Predictions} = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

where TP is the number of true positive instances, TN is the number of true negative instances, FP is the number of false positive instances, and FN is the number of false negative instances.[27] Despite intuitive, accuracy may obscure important differences in model behaviour, particularly when the classes are *imbalanced* or when certain misclassification types are more costly than others. For instance, in credit risk assessment, correctly predicting high-risk entities is more important than correctly identifying low-risk ones. For this reason, accuracy is always reported together with its complement, the *classification loss* or *misclassification rate*:

$$Loss = 1 - Accuracy$$

Loss highlights the proportion of observations the model fails to classify correctly, offering a clearer indication of predictive shortcomings.

[27] https://medium.com/@mlmind/evaluation-metrics-for-classification-fc770511052d

A deeper understanding of model performance is obtained through the *confusion matrix*, a tool that compares the model's predictions with the actual class labels. Unlike aggregate measures such as accuracy, which compress all errors into a single number, the confusion matrix reveals how the model is making mistakes and which classes tend to be confused.

For each model, a *single 3×3 confusion matrix* is produced, where rows correspond to the predicted classes and columns to the correct ones. The diagonal elements indicate correct classifications, while off-diagonal cells reveal systematic confusions between classes, for example, whether firms classified as medium risk are often mistaken for low-risk ones, or whether high-risk cases are under-detected. The *One-vs-All* framework is applied: the analysis temporarily focuses on a single class at a time by treating it as the "positive" category, while grouping the remaining two classes together as "negative". Repeating this procedure for each class yields interpretable, class-specific assessment measures while maintaining a consistent evaluation structure across models.

Within this binary perspective, four fundamental quantities are defined:

1. *True Positives (TP)*: cases belonging to the target class that the model correctly identifies,
2. *False Positives (FP)*: cases incorrectly flagged as belonging to the target class,
3. *True Negatives (TN)*: cases outside the target class correctly recognised as such,
4. *False Negatives (FN)*: cases belonging to the target class that the model fails to identify,

which correspond to the same quantities involved in the calculation of Accuracy previously displayed.

To evaluate the discriminative power of each model, *Receiver Operating Characteristic (ROC)* curves are employed. In the One-vs-All setting, a curve is generated for each class by plotting the *True Positive Rate* ($\frac{\mathrm{TP}}{\mathrm{TP+FN}}$) against the *False Positive Rate* ($\frac{\mathrm{FP}}{\mathrm{FP+TN}}$) at varying classification thresholds.

The True Positive Rate measures the proportion of correctly identified positive instances, while the False Positive Rate measures the proportion of negative instances incorrectly classified as positive. Rather than relying on a single probability cut-off, the ROC curve visualises the model's performance across all possible classification thresholds, offering a comprehensive view of how effectively each model separates one class from the others under different decision criteria.

The *Area Under the Curve (AUC)* provides a summary of ROC performance. AUC values closer to 1 indicate that the model distinguishes well between a given class and the others, while values around 0.5 suggest performance no better than random guessing. In a multiclass context, analysing AUC scores for each One-vs-All curve provides insights into which classes are easier or harder for the model to discriminate.

AUC can be interpreted as the probability that a randomly chosen positive instance possesses a higher predicted probability than a randomly chosen negative one.[28] Graphically, the ideal ROC curve is the one closest to the top-left corner of the plot, representing high TPR and low FPR. A curve that bends strongly toward this corner indicates a model with strong discriminative capability, whereas a curve lying close to the diagonal line corresponds to near-random performance.

[28] https://towardsdatascience.com/unlock-the-power-of-roc-curves-intuitive-insights-for-better-model-evaluation/

In addition to conventional performance measures, such as Accuracy and Classification Loss, Confusion Matrix and ROC curve, the analysis also adopts SAFE AI metrics to evaluate accuracy and predictor contribution, which provide a detailed framework for assessing predictive performances of machine learning models, especially in sensitive applications as credit risk classification.

Artificial Intelligence (AI) can bring great opportunities, as it is concerned with building machines and algorithms able to perform several tasks that typically require human intelligence, furthermore, machine learning methods are boosting the applications of AI in all human activities. Although, differently from ordinary computer software and applications, AI not only converts inputs into outputs, but can also change the surrounding environment, with the risk of creating harms for individuals, organisations and the environment, thus it can generate very relevant risks, such as cyber risks and model risks. This is the reason why policy makers, regulators and standard bodies around the world are issuing regulations and recommendations that AI developers, deployers and users should follow to manage the risks arising from the adoption of AI methods to make them sustainable. A reference model is the *European Artificial Intelligence Act* (*EU-AI Act*), which requires risk management of high-risk AI applications.[29] European Parliament's priority was to make sure that AI systems used in the EU are safe, transparent, traceable, non-discriminatory and environmentally friendly. AI systems should be overseen by people, rather than by automation, to prevent harmful outcomes. EU Parliament also wanted to establish a technology-neutral, uniform definition for AI that could be applied to future AI systems. The new rules establish obligations for providers and users

[29] Golnoosh Babaei, Paolo Giudici, Emanuela Raffinetti, *A Rank Graduation Box for SAFE AI*, Expert Systems with Applications, Volume 259, 2025

depending on the level of risk of AI risk qualification. While many AI systems pose minimal risk, they need to be assessed.[30] The AI Act introduces a risk-based approach to AI applications, defining an AI risk taxonomy with four risk categories: unacceptable risk, high risk, limited risk, and minimal risk. The requirements established for high-risk applications include *sustainability*, *accuracy*, *fairness and explainability*, which need a set of metrics that can establish not only whether but also how much the requirements are satisfied over time.[31]

SAFE AI metrics are based on the compliance to the main regulatory principles, clearly specified into the EU-AI Act, which are crucial for AI trustworthy:

- "S" for Sustainability, AI should be sustainable (robust to outlier data);
- "A" for Accuracy, AI should lead to accurate and realistic predictions;
- "F" for Fairness, AI should not discriminate by age, ethnicity, gender or other population groups;
- "E" for Explainability, AI should be explainable and interpretable in terms of its drivers.

SAFE AI metrics, thus, aim to ensure that machine learning models are not only accurate but also robust, transparent, and equitable; integrating SAFE AI in model evaluation allows to quantify aspects such as the contribution of individual predictors, the handling of outliers and potential biases in

[30] https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
[31] Paolo Giudici, Emanuela Raffinetti, *SAFE Artificial Intelligence in finance*, Finance Research Letters, Volume 56, 2023

predictions, all of which are critical when ESG factors are considered in credit risk assessment.

“Sustainability”, “Accuracy”, “Fairness” and “Explainability” are assessed by a structured set of performance metrics based on the *Rank Graduation Box* (*RGB*) methodology, which is based on the notion of concordance between two cumulative distributions, in particular the basic instruments are the Lorenz Curve, the related Dual Lorenz Curve and the Concordance curve, statistical tools widely used to summarise the distribution of income and wealth, which can be adapted to assess the distribution of predicted scores or class membership probabilities in credit risk classification.

If Y* and Y** are two statistical distributions (continuous, ordinal or binary), each defined on a set of n data points:

- the Lorenz curve, L_{Y^*}, is built by arranging the Y* values in a *non-decreasing sense*. For i = 1,…,n, the Lorenz curve is defined by the set of points $(i/n, \frac{\Sigma_{j=1}^{i} y_{r_j^*}^*}{n\overline{y^*}})$, where r_j^* indicates the non-decreasing ranks of Y* and $\overline{y^*}$ indicates the mean of Y;

- the same Y* values can be ordered in a *non-increasing sense* to build the dual Lorenz curve, L'_{Y^*}. For i = 1,…,n, the dual Lorenz curve is defined by the pairs: $(i/n, \frac{\Sigma_{j=1}^{i} y_{r_{n+1-j}^*}^*}{n\overline{y^*}})$, where r_{n+1-j}^* indicates the non-increasing ranks of Y*;

- The Concordance curve can be derived ordering the Y* values with respect to the ranks of the Y** values. It is defined by the pair of points $(i/n, \frac{\Sigma_{j=1}^{i} y_{r_j^{**}}^*}{n\overline{y^*}})$, where r_j^{**} indicates the non-decreasing ranks of Y**. In this framework, the Concordance curve compares the ranks

induced by the model's predicted scores with the ranks implied by the observed class labels.

Based on the behaviour of the Concordance curve, there are four main reference scenarios that occur in model comparison:

1. Full concordance (best case): the best case occurs when the ordering of the Y response variable values corresponds to the ordering of the predicted values ($r_j^{**} = r_j^*$), the Concordance curve perfectly overlapping the Lorenz curve, here the model perfectly aligns with observed ranks;

2. Full discordance (worst case): the worst case occurs when the ordering of the Y response variable values is in inverse correspondence with the ordering of the predicted values ($r_j^* = r_{n+1-j}^*$), the Concordance curve perfectly overlapping the dual Lorenz curve, here the ranks of the predicted classes are opposite to the ranks of the actual classes;

3. No association (random case): this scenario verifies when all the predicted classes are identical, the Concordance curve overlaps the 45-degree line, the model provides no useful ranking;

4. General case: the distance between the Concordance curve and the Lorenz curve measures how the ranks of Y** differ from the ranks of Y*, in terms of the ranked values, it quantifies the degree of agreement between predicted and actual ranks and how a model improves over random predictions.[32] [33]

[32] Golnoosh Babaei, Paolo Giudici, Emanuela Raffinetti, *A Rank Graduation Box for SAFE AI*, Expert Systems with Applications,Volume 259, 2025
[33] Raffinetti, E, *A Rank Graduation Accuracy measure to mitigate Artificial Intelligence risks*., *Qual Quant* **57** (Suppl 2), 131–150 (2023)

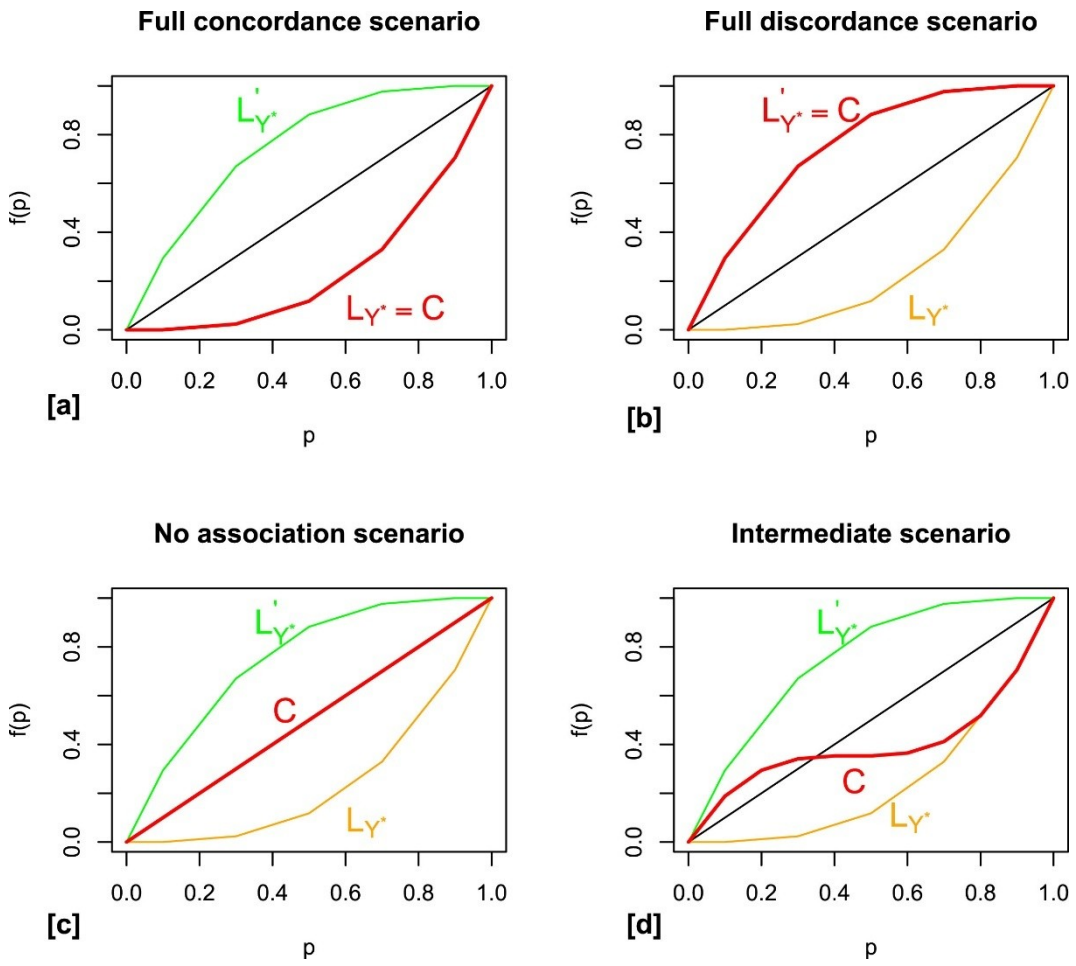

Figure 3: Relationship between Lorenz, dual Lorenz, and Concordance curves under the four ranking association scenarios. Reference: Golnoosh Babaei, Paolo Giudici, Emanuela Raffinetti, A Rank Graduation Box for SAFE AI, Expert Systems with Applications, Volume 259, 2025

Hence, based on the Rank Graduation Box approach, risk metrics for "Sustainability", "Accuracy", "Fairness" and "Explainability" can be derived.

The first metrics is related to "Accuracy", which requires that the output of an AI application is "close" to the observed (or expected) output. "Accuracy" can be assessed through the *Rank Graduation Accuracy measure* (RGA). Given a set of K predictors, RGA can be obtained by:

- considering the actual values of the target variable → y;
- considering the predicted values provided by the model fitted on the K predictors → $\hat{y}$;
- re-ordering the y values with respect to the non-decreasing ranks of the $\hat{y}$ values, denoted with $\hat{r}$.

The RGA is obtained by dividing the area between the Concordance curve and the dual Lorenz curve by its maximum attainable value, which corresponds to the area between the Lorenz curve and the dual Lorenz curve. RGA is defined as:

$$\text{RGA} = \frac{\sum_{i=1}^{n}\{\frac{1}{n\bar{y}}(\sum_{j=1}^{i} y_{r_{n+1-j}} - \sum_{j=1}^{i} y_{\hat{r}_j})\}}{\sum_{i=1}^{n}\{\frac{1}{n\bar{y}}(\sum_{j=1}^{i} y_{r_{n+1-j}} - \sum_{j=1}^{i} y_{r_j})\}} = \frac{\sum_{i=1}^{n} i y_{\hat{r}_i} - \sum_{i=1}^{n} i y_{r_{n+1-i}}}{\sum_{i=1}^{n} i y_{r_i} - \sum_{i=1}^{n} i y_{r_{n+1-i}}}$$

$0 \leq \text{RGA} \leq 1$, with RGA = 1 for a perfectly concordant model, RGA = 0 for a perfectly discordant model, RGA = 0.5 for random predictions (C coincides with the 45-degree line). The best scenario occurs when the predicted ranks of the response variable are equal to the observed ranks, with the Concordance curve perfectly overlapping the Lorenz curve, the area between the Concordance curve and the dual Lorenz curve equals the maximum attainable area, which is the area between the Lorenz and the dual Lorenz curves. In this framework, this corresponds to a scenario in which all observations are correctly ordered according to their true credit risk, implying that the model achieves the highest possible Rank Graduation Accuracy and therefore provides the most accurate ranking-based classification among the competing models. Rank Graduation Accuracy can thus be used as a comparative metric to evaluate which model produces the most accurate overall ranking of firms.

The second measure is about "Sustainability", which is related to the robustness of AI applications. The output of AI may be affected by extreme data, which can distort the inputs of an AI application and, consequently, its output. The measurement of robustness is usually conducted in terms of an appropriate distance between the model predictions and those obtained under data perturbations. To evaluate the capability of a model to be robust to extreme data and outliers, the RGA measure is re-formalised, giving rise to the *Rank Graduation Robustness* (RGR) metric. RGR can be obtained by:

- considering the predicted values computed by applying the model on data without perturbations → $\hat{y}$;
- considering the predicted values provided by the model fitted on perturbed data → $\hat{y}^p$;
- re-ordering the $\hat{y}$ values with respect to the non-decreasing ranks of the $\hat{y}^p$ values, denoted with r^p.

Based on RGA, RGR is defined as:

$$\mathrm{RGR} = \frac{\sum_{i=1}^{n}\{\frac{1}{n\bar{\hat{y}}}(\sum_{j=1}^{i}\hat{y}_{r_{n+1-j}} - \sum_{j=1}^{i}\hat{y}_{r_j{}^p})\}}{\sum_{i=1}^{n}\{\frac{1}{n\bar{\hat{y}}}(\sum_{j=1}^{i}\hat{y}_{r_{n+1-j}} - \sum_{j=1}^{i}\hat{y}_{r_j})\}} = \frac{\sum_{i=1}^{n} i\hat{y}_{r_i{}^p} - \sum_{i=1}^{n} i\hat{y}_{r_{n+1-i}}}{\sum_{i=1}^{n} i\hat{y}_{r_i} - \sum_{i=1}^{n} i\hat{y}_{r_{n+1-i}}}$$

$0 \leq \mathrm{RGR} \leq 1$, with RGR = 1 for a full robust model, RGR = 0 for a full perturbed model, RGR = 0.5 if the perturbations lead to a random model. The case of maximum robustness occurs when the ranks of the response predicted values correspond to the ranks of the predicted values obtained using the perturbed data, with the Concordance curve C perfectly overlapping the Lorenz curve.

Now the measurement of "Explainability" is considered, which is involved to evaluate the contribution of each predictor to the explanation of the response variable. The associated evaluation metric is *the Rank Graduation Explainability* (RGE). Given a set of K predictors, RGE can be obtained by:

- considering the predicted values provided by the full model (with all the K predictors) → $\hat{y}$;
- considering the predicted values provided by the reduced model (excluding the k-th predictor under evaluation) → $\hat{y}^{(-X_k)}$;
- re-ordering the $\hat{y}$ values with respect to the non-decreasing ranks of the $\hat{y}^{(-X_k)}$ values, denoted with r^{-X_k}.

RGE is defined as:

$$\text{RGE} = 1 - \frac{\sum_{i=1}^{n}\{\frac{1}{n\bar{\hat{y}}}(\sum_{j=1}^{i}\hat{y}_{r_{n+1-j}} - \sum_{j=1}^{i}\hat{y}_{r_j - \text{Xk}})\}}{\sum_{i=1}^{n}\{\frac{1}{n\bar{\hat{y}}}(\sum_{j=1}^{i}\hat{y}_{r_{n+1-j}} - \sum_{j=1}^{i}\hat{y}_{r_j})\}} = 1 - \frac{\sum_{i=1}^{n} i\hat{y}_{r_i - \text{Xk}} - \sum_{i=1}^{n} i\hat{y}_{r_{n+1-i}}}{\sum_{i=1}^{n} i\hat{y}_{r_i} - \sum_{i=1}^{n} i\hat{y}_{r_{n+1-i}}}$$

$0 \leq \text{RGE} \leq 1$, with RGE = 1 if the k-th predictor provides the maximum explanation, RGE = 0 if the k-th predictor does not contribute to the explanation of the response, RGE = 0.5 if the model without the k-th predictor corresponds to the random model (C coincides with the 45-degree line). The stronger is the effect of a variable X_k on explaining Y, the larger is the divergence between the ranks of the full model predicted values and those associated with the reduced model. So, the case of maximum explainability occurs when the ranks of the $\hat{Y}$ response predicted values is in inverse correspondence with the ranks of the $\hat{Y}^{(-X_k)}$; predicted values (obtained by fitting the model on the data without the k-th predictor), with the Concordance curve C perfectly overlapping, instead, the dual Lorenz curve in this case. In the context of this analysis, the RGE measure is employed to

assess the relative contribution of each predictor to the quality of the credit risk ranking produced by the model, allowing an explicit evaluation of whether ESG scores provide a meaningful and incremental explanatory role alongside traditional financial variables in ordering firms from lower to higher credit risk. In this way, the relevance of sustainability information in driving the predictive ordering of credit risk can be directly assessed.

Ultimately, the measurement of "Fairness" is considered. Fairness is typically assessed by comparing the outputs generated for different population groups. These comparisons, when made without taking into account for other input variables, may fail to explain the underlying reasons for any observed disparities. In some cases, a model may appear fair overall yet exhibit unfairness when considered conditionally on relevant factors. For example, a credit score may seem fair with respect to the borrower's gender across the entire sample, but not conditionally on the level of income. In this case the evaluation measure is the *Rank Graduation Fairness* metric (RGF). Given a set G of group (dummy) variables, the RGF can be obtained by:

- considering the predicted values provided by the full model (including all the G group variables as predictors) → $\hat{y}$;
- considering the predicted values provided by the reduced model (excluding the g-th group under evaluation) → $\hat{y}^{(-X_g)}$;
- re-ordering the $\hat{y}$ values with respect to the non-decreasing ranks of the $\hat{y}^{(-X_g)}$ values, denoted with r^{-X_g}.

RGF is defined as:

$$\mathrm{RGF} = \frac{\sum_{i=1}^{n}\{\frac{1}{n\bar{\hat{y}}}(\sum_{j=1}^{i}\hat{y}_{r_{n+1-j}} - \sum_{j=1}^{i}\hat{y}_{r_j - \mathrm{Xg}})\}}{\sum_{i=1}^{n}\{\frac{1}{n\bar{\hat{y}}}(\sum_{j=1}^{i}\hat{y}_{r_{n+1-j}} - \sum_{j=1}^{i}\hat{y}_{r_j})\}} = \frac{\sum_{i=1}^{n} i\hat{y}_{r_i - \mathrm{Xg}} - \sum_{i=1}^{n} i\hat{y}_{r_{n+1-i}}}{\sum_{i=1}^{n} i\hat{y}_{r_i} - \sum_{i=1}^{n} i\hat{y}_{r_{n+1-i}}}$$

$0 \leq RGF \leq 1$, with RGF = 1 in the case of maximum fairness, RGF = 0 in the case of maximum unfairness, RGF = 0.5 if the model without the g-th group variable corresponds to the random model. The maximum fairness scenario arises when the ranks of the response predicted values $\hat{Y}$ correspond to the ranks of the $\hat{Y}^{(-X_g)}$ predicted values (obtained by fitting the model without the g-th group variable), with the Concordance curve C perfectly overlapping the Lorenz curve.[34] [35]

SAFE AI metrics will be employed in the empirical analysis, along with the traditional evaluation metrics, to evaluate and compare the predictive performance of the models, specifically RGA and RGE will be used to assess the overall ranking accuracy and the contribution of ESG predictors.

[34] Golnoosh Babaei, Paolo Giudici, Emanuela Raffinetti*, A Rank Graduation Box for SAFE AI*, Expert Systems with Applications,Volume 259, 2025
[35] Raffinetti, E*, A Rank Graduation Accuracy measure to mitigate Artificial Intelligence risks*., *Qual Quant* **57** (Suppl 2), 131–150 (2023)

1.4.10 ESG FACTORS IN CREDIT RISK CLASSIFICATION

In recent years, the growing attention to environmental, social, and governance issues has substantially reshaped the way credit risk is assessed. This evolution is closely linked to the EU Sustainable Finance Framework, through which the European Commission promotes the integration of sustainability considerations into financial decision-making, with the aim of fostering long-term economic resilience and redirecting capital flows toward sustainable activities. Within this framework, sustainability factors are no longer viewed as purely ethical concerns, but as financially material drivers of risk and return. In the EU's policy context, sustainable finance is understood to support economic growth while reducing environmental pressures, in line with the climate and environmental objectives of the European Green Deal, while also taking social and governance aspects into account. It further emphasizes transparency regarding ESG-related risks that may affect the financial system and the mitigation of such risks through appropriate governance of financial and corporate actors. Hence, by channelling private capital toward the transition to a climate-neutral, climate-resilient, and resource-efficient economy, sustainable finance plays a central role in achieving the EU's sustainability objectives.[36]

Beyond traditional financial indicators such as leverage, profitability, and liquidity, credit analysts and rating agencies increasingly incorporate non-financial sustainability dimensions into their evaluations. This shift is driven by the recognition that ESG performance can influence a firm's long-term resilience, cost of capital, exposure to regulatory sanctions, reputational vulnerability, and ultimately its creditworthiness. Furthermore, ESG information captures risks and opportunities that are not immediately visible

[36] https://finance.ec.europa.eu/sustainable-finance/overview-sustainable-finance_en

in standard accounting metrics, offering a broader perspective on corporate performance. This is particularly relevant in credit risk assessment, where the cost of misclassifying high-risk entities can be substantial.

The ESG framework is structured around three pillars: Environmental, Social, and Governance, which collectively capture the multidimensional non-financial impacts of corporate activity. These dimensions reflect the areas in which firms can generate positive or negative outcomes, either directly through their operations or indirectly through their value chains. A comprehensive ESG assessment therefore evaluates performance across all three pillars to provide a transparent picture of a company's sustainability profile and long-term risk exposure.

The Environmental pillar concerns a firm's interaction with the natural environment and its ability to manage ecological risks. Beyond general considerations such as carbon emissions, energy efficiency, and regulatory compliance, environmental criteria include several issues. The first issue is about direct and indirect Greenhouse gas emissions (Scope 1, 2, and increasingly Scope 3). Scope 1, 2 and 3 refer to the standard Greenhouse Gas Protocol classification, which consists of a set of standards and tools developed as a joint initiative of the *World Resources Institute and the World Business Council for Sustainable Development (WBCSD)*. It provides solutions for measuring and managing Greenhouse gas emissions from public and private sector operations, value chains and mitigation actions. Reducing Greenhouse gas emissions helps mitigate climate change, numerous companies and organisations are pursuing sustainability and climate action by setting targets to achieve *net zero emissions*. Measuring and reporting Scope 1, 2 and 3 emissions can help understand their contributions and

identify opportunities for their reduction.[37] Scope 1 includes direct emissions from sources owned or controlled by the company, Scope 2 covers indirect emissions from purchased electricity, heat, or steam, while Scope 3 comprises all other indirect emissions generated along the value chain, such as those from suppliers, business travel, logistics, or product use. The other issues are electricity and resource consumption, waste production and recycling policies, water usage and pollution reduction plans, impacts and dependencies related to biodiversity and ecosystems, exposure to transition and physical climate risks. Firms with robust environmental practices tend to be better equipped to navigate regulatory tightening, climate-related operational disruptions, and increasing stakeholder scrutiny, moreover, they are often better positioned to manage regulatory risks and operational challenges related to climate change.

The Social pillar evaluates how firms manage relationships with stakeholders across their entire value chain including employees, customers, suppliers, and local communities. Social criteria extend beyond broad notions of workforce management and encompass workplace health and safety indicators (e.g. accident frequency rates), training and professional development programmes, labour practices, employee rights, diversity and inclusion policies, supply-chain responsibility and human rights safeguards, customer protection, product quality, data privacy, engagement with and impact on local communities. Strong social performance can enhance a firm's reputation, operational resilience, strengthen employee retention and productivity, and reduce operational vulnerabilities associated with labour disputes or supply-chain disruptions.

[37] https://www.ibm.com/it-it/think/topics/scope-1-2-3-emissions

Governance pillar relates instead to the structures, processes, and controls through which a firm is directed and overseen. A good governance reduces agency conflicts and enhances transparency, thus lowering the likelihood of mismanagement or financial irregularities. Governance criteria include structure, independence and diversity of the Board of Directors, alignment of executive compensation with long-term performance, internal controls and transparency in financial reporting, shareholder rights and protection of minority interests, ethical conduct, anti-corruption policies, supply-chain integrity, relationships with external stakeholders and adherence to regulatory standards. High governance quality thus mitigates agency problems, reduces financial mismanagement, and enhances access to capital markets, it is also strongly linked to lower information asymmetry, reduced cost of capital, and greater investor confidence.

An ESG evaluation synthesises performance across all three pillars to provide a coherent and transparent representation of a firm's sustainability and ethical footprint. ESG analysis is very useful as a complement to traditional financial metrics: a company's financial results can be interpreted based on its environmental, social, and governance practices, allowing investors, lenders, and rating agencies to identify long-term risks and opportunities that may not be visible through financial statements. Thus, ESG assessments increasingly influence investment decisions, credit risk evaluations, and strategic corporate adjustments.[38]

Beyond the conceptual definition of Environmental, Social, and Governance factors, the practical implementation of ESG relies on a wide set of international reporting frameworks, standards, and rating methodologies. Over the past decade, ESG criteria have become a global reference for

[38] https://www.greenscope.io/it/esg#anchor-4

assessing the non-financial performance of companies and are now embedded in several regulatory and voluntary disclosure systems. Frameworks such as the *Global Reporting Initiative (GRI)*, along with standards issued by the *International Sustainability Standards Board (ISSB)* and the *Sustainability Accounting Standards Board (SASB)*, provide structured guidelines for companies to measure and communicate their sustainability performance. These frameworks support the harmonization of ESG reporting practices across industries. In addition to normative standards, numerous rating agencies produce ESG scores that evaluate corporate performance across the three pillars. These ratings aim to offer standardized and comparable assessments of a firm's sustainability profile and frequently inform investment decisions, supplier selection processes, and risk assessments.

The strategic relevance of ESG is also reinforced by the growing integration of sustainability criteria into investment practices. A rising share of global assets under management is allocated according to ESG considerations, driven by the view that more sustainable firms tend to be more resilient in the long run and better equipped to anticipate regulatory, environmental, and social risks. Several investor surveys indicate that institutional investors increasingly incorporate ESG factors when allocating capital, not only to align with responsible investment principles but also to enhance long-term risk-adjusted returns. Investors may also exclude controversial sectors or engage directly with firms to influence their sustainability practices, contributing to a broader transformation of corporate behaviour.

In addition, ESG reporting has become a regulatory requirement in many jurisdictions. In the European Union, the *Corporate Sustainability Reporting Directive (CSRD)* and the related *European Sustainability Reporting Standards (ESRS)* mandate detailed disclosures on environmental, social, and

governance aspects. The ESRS provide a comprehensive and standardised framework for disclosing sustainability information, ensuring that firms report material environmental, social, and governance impacts in a consistent and comparable manner across the European Union. These standards rely on the *double-materiality principle* which requires companies to disclose both how sustainability issues affect their financial performance and how their activities impact the society and the environment, strengthening the integration of ESG considerations into risk management and corporate reporting practices. Combined with frameworks such as the *Sustainable Finance Disclosure Regulation (SFDR)*, these regulations aim to increase transparency, facilitate comparability across companies, and support the integration of sustainability factors into financial decision-making. Reliable ESG measurement and reporting have become essential components of modern corporate transparency and risk assessment.[39] [40]

Also, there is a growing body of empirical research that supports the relevance of ESG indicators in credit risk assessment. Evidence suggests that firms with stronger ESG performance, particularly in governance, but also in environmental and social dimensions, tend to exhibit lower default probabilities, improved credit ratings, and reduced overall credit risk.[41] In highly regulated or environmentally sensitive sectors, environmental and social metrics become particularly salient, complementing traditional financial indicators in credit scoring models.[42] Furthermore, strong ESG practices, especially governance quality, can lower a firm's cost of debt and enhance access to capital markets, demonstrating that ESG factors provide

[39] https://www.greenscope.io/it/esg#anchor-8
[40] https://www.climatepartner.com/it/formazione/glossario/european-sustainability-reporting-standards-esrs
[41] Laura Bonacorsi, Vittoria Cerasi, Paola Galfrascoli, Matteo Manera, *ESG Factors and Firms' Credit Risk*, Journal of Climate Finance, Volume 6, 2024
[42] Patrycja Chodnicka-Jaworska P, *ESG as a Measure of Credit Ratings*, Risks 2021, 9, 226.

meaningful information beyond conventional financial fundamentals.[43] These findings underscore the importance of integrating ESG metrics into credit risk modelling, allowing for a more complete evaluation of firm resilience, sustainability, and long-term financial stability.

In this analysis, ESG metrics and scores are incorporated as explanatory variables alongside traditional financial and accounting measures. For each classification model employed, variable importance is analysed to assess the relative contribution of each ESG variable, such as Environmental, Social, Governance and ESG risk score, in predicting credit risk class.

In linear classification models, variable importance is inferred from the magnitude of the estimated coefficients. In Multinomial Logistic Regression, the relevance of each predictor is reflected in the size of the coefficients that remain after regularization, as larger absolute values indicate a stronger contribution to class separation. In Linear Discriminant Analysis, importance is instead evaluated through the discriminant loadings, which quantify how strongly each variable contributes to the formation of the discriminant functions used to differentiate between classes. In Classification trees, attention is given to which predictors are used for splits and how they influence the construction of the trees. For Random Forest model, variable importance can be visualised through plots that summarise each feature's contribution to predictive accuracy. By combining financial and ESG variables within the same classification framework, the analysis in addition to comparing overall predictive performance across models for firms included in the EURO STOXX 600 index, also aims to highlight which ESG factors have the greatest influence as predictors relative to traditional

[43] Egidio Palmieri, Greta Benedetta Ferilli, Yener Altunbas, Valeria Stefanelli, Enrico Fioravante Geretto, *Business model and ESG pillars: The impacts on banking default risk*, International Review of Financial Analysis, Volume 91,2024

financial metrics. Hence, this approach allows for an evaluation of the incremental value provided by information related to sustainability factors in the credit risk assessment context.

2. EMPIRICAL ANALYSIS: DEVELOPMENT AND COMPARISON OF PREDICTIVE MODELS

2.1 INTRODUCTION TO THE EMPIRICAL ANALYSIS

The empirical analysis conducted in this chapter aims to develop the models previously introduced and to evaluate whether the integration of ESG variables is relevant to the classification of corporate credit risk for the Euro STOXX 600 companies. Based on the theoretical foundations outlined in the first chapter, regarding credit risk modelling, evaluation metrics for assessing predictive performance of the models and the role of ESG factors, the empirical analysis puts these concepts into operation through the application of the machine learning models previously explained.

The analysis is conducted on a five-year dataset of the Euro STOXX 600 index sourced from Bloomberg, combining conventional accounting and financial ratios with ESG indicators, these predictors are used as explanatory variables. By applying the same dependent variable and set of predictors across all the models, the analysis ensures full comparability of results.

The empirical analysis involves two main objectives: assessing and comparing the classification performance of the models and examining the relative importance of financial and ESG predictors within each modelling framework. This approach allows for a substantial evaluation of whether ESG information provides meaningful contribution to credit risk classification beyond traditional financial variables.

2.2 EXTRACTION OF THE DATA AND DATA CLEANING

The analysis relies on a cross-sectional dataset sourced from the Bloomberg Terminal, which provides comprehensive and standardized financial and ESG data for listed companies. The sample is composed of all firms included in the EURO STOXX 600 index, which represents the largest capitalization companies across 17 European countries. It consists of a broad and diversified set of firms, suitable for analyzing credit risk dynamics and the role of ESG factors.

The dataset covers the period from October 2020 to September 2025, resulting in a five-year observation window. Data has been extracted at a monthly frequency, generating multiple observations for each firm over time.

Before model estimation and after importing the libraries required for the analysis, a preliminary preprocessing and data cleaning procedure has been implemented to ensure the consistency of the dataset.

```
# --- Load Required Libraries ---
rm(list=ls())
library(readxl)
library(MASS)
library(glmnet)
library(vip)
library(corrplot)
library(car)
library(dplyr)
library(ggplot2)
library(pROC)
library(class)
library(caret)
library(tree)
library(randomForest)
```

Code 1: Importing the libraries required for the analysis

First, the date variable has been converted to a standard character format to avoid inconsistencies in data handling. Observations containing missing

entries and information in the original dataset have been first converted into standard missing values (NA) in R and subsequently removed from the sample to ensure that the empirical analysis is conducted on a cleaned dataset exclusively composed of complete observations. Then, a non-informative column for the further steps, containing stocks tickers, has been excluded from the dataset.

```
# --- Data Cleaning ---
Data$DATE <- as.character.Date(Data$DATE)
Data[Data == "n.a."] <- NA
Data <- na.omit(Data)
Data <- Data[, -1]
```

Code 2: Data cleaning, removing missing and non-informative columns

2.3 DEPENDENT VARIABLE – ALTMAN Z - SCORE: CREDIT RISK CLASSES

The dependent variable employed in the empirical analysis is a categorical indicator of credit risk, constructed based on the Altman Z-Score values. As discussed in the previous chapter (Section 1.3), the Altman Z-Score represents one of the most widely used and empirically validated measures for assessing firms' financial distress and default risk. The analytical formulation of the Z-Score and a detailed explanation of its components are also provided in Section 1.3.

Following the thresholds reported by Bloomberg and consistent with Altman's original rules, firms are classified to one of the three discrete risk categories:

- *Safe Zone*: corresponding to financially healthy firms with low credit risk;

- *Grey Zone*: representing firms with intermediate financial conditions and uncertainty;
- *Distress Zone*: which identifies companies exposed to a high probability of financial distress.

These categories are constructed by applying the standard Z-Score cut-off values, as indicated by Bloomberg, to each firm observation in the dataset. In particular, observations with an Altman Z-Score value below 1.8 are classified as Distress, values between 1.8 and 3 are assigned to the Grey Zone and values greater than 3 are classified as Safe.

A new categorical variable, denoted as *Risk_Class*, is created to encode these three credit risk classes. The variable takes values 1 for Safe firms, 2 for Grey Zone firms and 3 for Distress firms. This numerical encoding allows to directly use *Risk_Class* as a dependent variable within the classification algorithms applied in the empirical analysis. After the generation of credit risk classes, the original continuous variable containing the Altman Z-Score values is removed from the dataset, ensuring that Z-Score itself is not exploited as explanatory variable by the models.

```
42 # --- Create Credit Risk Classes based on Altman Z-Score ---
43 Data$Risk_Class <- NA
44 Data$Risk_Class[Data$ALTMAN_Z_SCORE < 1.8] <- 3
45 Data$Risk_Class[Data$ALTMAN_Z_SCORE >= 1.8 & Data$ALTMAN_Z_SCORE <= 3] <- 2
46 Data$Risk_Class[Data$ALTMAN_Z_SCORE > 3] <- 1
47
48 Data <- Data[, -2]  # removing ALTMAN_Z_SCORE
```

Code 3: Creation of credit risk classes encoded in the Risk_Class variable and removal of the continuous Altman Z-Score variable

2.4 EXPLANATORY VARIABLES

The set of independent variables employed in the empirical analysis includes both traditional financial indicators and ESG-related metrics, combining financial performance measures and sustainability-related information, these variables can capture complementary dimensions of firms' credit risk profiles. All the variable definitions, construction methodologies and data sources are provided by Bloomberg and ensure consistency and comparability across the companies. The financial variables include indicators of leverage, profitability, liquidity, interest coverage and revenue growth, which have always been central determinants of creditworthiness. The ESG-related variables capture firms' environmental, social and governance performance, as well as sustainability ratings and ESG risk scores, which have been increasingly recognized as crucial drivers of credit risk.

The descriptions of the variables, sourced from Bloomberg terminal, are provided below:

SR001 - BESG ESG Score (ESG_SCORE): Provides the Bloomberg score evaluating the company's aggregated Environmental, Social and Governance (ESG) performance. The score is based on Bloomberg's view of ESG financial materiality. The score is a weighted generalized mean (power mean) of Pillar Scores, where the weights are determined by the pillar priority ranking. Values range from 0 to 10; 10 is best.

SR002 - BESG Environmental Pillar Score (ENVIRONMENTAL_SCORE): Provides the Bloomberg score evaluating the company's aggregated Environmental performance. The score is based on Bloomberg's view of financial materiality. The Pillar Score is a weighted generalized mean (power

mean) of Issues Scores, where the weights are determined by the Issue Priority ranking. Values range from 0 to 10; 10 is best.

SR003 - BESG Social Pillar Score (SOCIAL_SCORE): Provides the Bloomberg score evaluating the company's aggregated Social performance. The score is based on Bloomberg's view of financial materiality. The Pillar Score is a weighted generalized mean (power mean) of Issue Scores, where the weights are determined by the Issue Priority ranking. Values range from 0 to 10; 10 is best.

SR004 - BESG Governance Pillar Score (GOVERNANCE_SCORE): Provides the Bloomberg score evaluating the company's aggregated Governance performance. The score is based on Bloomberg's view of financial materiality. The Pillar Score is a weighted generalized mean (power mean) of Theme Scores, where the weights are determined by the Theme Priority rankings using a transformation function. Values range from 0 to 10; 10 is best.

X6495 - S&P Global ESG Rank (SP_R_ESG_RANK): Total sustainability percentile rank, converted from the total sustainability score, based on the S&P Global ESG Rank (formerly RobecoSAM Corporate Sustainability Assessment). A company's Total Sustainability Score is the sum of all question scores and ranges from 0-100. The Total Sustainability Score is based on individual questions that roll up into criteria, which in turn roll up into three dimensions - Economic, Environmental and Social. The types and weights of individual questions and criteria are adjusted for each industry-specific questionnaire to reflect the materiality of specific sustainability themes within each industry. The Total Sustainability Score can be defined as follows: Total Sustainability Score = (Number of Question points received x Question Weight x Criterion Weight).

X7278 - SA ESG Risk Score (SA_ESG_RISK_SCR): The company's overall score in the ESG Risk Rating. It applies the concept of risk decomposition to derive the level of unmanaged risk for a company, which is assigned to one of five risk categories. The score ranges from 0 and 100, with 0 indicating that risks have been fully managed (no unmanaged ESG risks) and 100 indicating the highest level of unmanaged risk. It is calculated as the difference between a company's overall exposure score and its overall managed risk score, or alternatively by adding the Corporate Governance unmanaged risk score to the sum of the company's issue unmanaged risk scores.

X5876 - ISS QualityScore (Governance) (ISS_QUALITYSCORE): Overall score assigned by Institutional Shareholder Services (ISS) to the company's governance practices. The score ranges from 1 for best to 10 for worst.

RR251 - Short and Long Term Debt (SHORT_AND_LONG_TERM_DEBT): Sum of short term and long term debt. Figure is reported in million; the Scaling Format Override (DY339, SCALING_FORMAT) can be used to change the display units for the field.

INDUSTRIALS, INSURANCE, UTILITIES & REITS

Calculated as: Short Term Debt + Long term Debt

Where: Short Term Debt is BS047, BS_ST_BORROW Long Term Debt is BS051, BS_LT_BORROW

If known long term debt is not disclosed, this field returns blank.

BANKS & FINANCIALS

Calculated as: Short Term Debt + Securities Sold With Repurchase Agreements + Long Term Debt

Where: Short Term Debt is BS047, BS_ST_BORROW Securities Sold With Repurchase Agreements is BS049, BS_SEC_SOLD_REPO_AGRMNT Long Term Debt is BS051, BS_LT_BORROW

RR028 - Return on Assets (RETURN_ON_ASSET): Indicator of how profitable a company is relative to its total assets, in percentage. Return on assets gives an idea as to how efficient management is at using its assets to generate earnings.

INDUSTRIALS, BANKS, FINANCIALS, UTILITIES, & REITS

Calculated as: (Trailing 12M Net Income / Average Total Assets) * 100

Where: Trailing 12M Net Income is RR813, TRAIL_12M_NET_INC Average Total Assets is the average of the beginning balance and ending balance of BS035, BS_TOT_ASSET

INSURANCE

Calculated as: ((Trailing 12M Net Income + Trailing 12M Policyholders' Surplus) / Average Total Assets) * 100

Where: Trailing 12M Net Income is RR813, TRAIL_12M_NET_INC Trailing 12M Policyholders' Surplus is RR713, TRAIL_12M_POLICY_HOLDER_SURPLUS Average Total Assets is the average of the beginning balance and ending balance of BS035, BS_TOT_ASSET

RX225 - EBITDA Margin (EBITDA_TO_REVENUE):

INDUSTRIALS, FINANCIALS, UTILITIES & REITS Measure, in percentage, calculates the relation of Earnings Before Interest, Taxes,

Depreciation and Amortization to Revenue. Calculated as: (EBITDA / Revenue) * 100

Where: EBITDA is RR009, EBITDA Revenue is IS010, SALES_REV_TURN

RR053 - Current Ratio (CUR_RATIO):

INDUSTRIALS, UTILITIES, REITS & MUNICIPAL REVENUE Ratio to indicate the company's ability to pay back its short-term liabilities with its short-term assets. Unit: Actual.

Calculated as: Current Assets / Current Liabilities

Where: Current Assets is BS015, BS_CUR_ASSET_REPORT Current Liabilities is BS050, BS_CUR_LIAB

RR060 - EBIT/Interest (INTEREST_COVERAGE_RATIO):

INDUSTRIALS, UTILITIES, REITS, & MUNICIPAL REVENUE Commonly known as Interest Coverage Ratio. Ratio used to determine how easily a company can pay interest on outstanding debt. EBIT (earnings before interest and taxes) is also commonly known as Operating Income. Unit: Actual. Calculated as: EBIT / Total Interest Incurred

Where: EBIT is RR002, EBIT Total Interest Incurred is RR011, TOT_INT_EXP

RR033 - Revenue Growth Year over Year (SALES_GROWTH): A percentage increase or decrease of sales revenue by comparing current period with same period prior year.

Calculated as: (Revenue from Current Period - Revenue from Same Period Prior Year) * 100 / Revenue from Same Period Prior Year

Where: Revenue is IS010, SALES_REV_TURN Revenue Growth is not computed if Revenue changes signs from prior year to current period.

Together, these variables constitute the common predictor set employed in all models, enabling the evaluation of whether ESG-related variables provide additional information beyond that captured by traditional financial measures of credit risk.

Specifically, Bloomberg ESG scores measure a company's management of financially material ESG issues. Financial materiality is defined as the issues that can have a negative or positive impact on a company's financial performance, such as revenue streams, operating costs, cost of capital, asset value and liabilities. Bloomberg identifies "financially material" issues based on proprietary research, which is shared transparently and based on an assessment of probability, magnitude and timing of the impact.

Below a specification regarding the methodology adopted by Bloomberg to structure ESG scores.

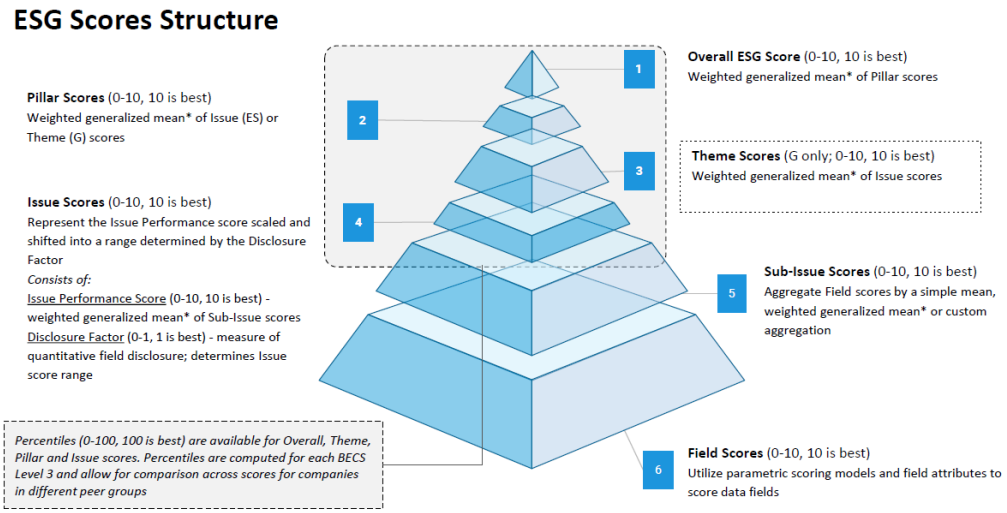

Figure 4: Bloomberg ESG Scores structure

Headline Scores: aggregate the scores of the E, S, and G pillars, based on an industry-specific weighting of E and S issues. For each industry, E, S and G pillars are ranked on a scale of 1 to 5, with 1 reflecting highest importance, then rankings are translated into a percentage weight.

Pillar Scores: derived from the Issue Scores using a generalized weighted mean (power mean); Issue Scores containing only binary fields have an 80% reduced weight.

Theme Scores: derived from the Governance Issue Scores using the same weighted mean.

Issue Scores, combine two dimensions:

- Issue Performance Score: weighted average of the Sub-Issue Scores, with reduced weights for binary fields.
- Issue Disclosure Factor (DF): weighted percentage which measures the quantitative and binary fields in the Issue.

Sub-Issue Scores: aggregated from the Field Scores using a weighted average, depending on the Fit/Quality attribute of fields in the sub-issue.

Field Scores: aggregated into a Sub-Issue score based on the Fit/Quality attribute. Input fields are given attributes based on their fit and quality and scored into Field Scores. Each field is scored using a quantitative methodology, taking into account normalization, polarity and the type of field. [44]

[44] Bloomberg Terminal. Environmental, Social and Governance (ESG) Scores (2023)

All explanatory variables are converted into a numeric structure to ensure compatibility with the machine learning algorithm employed and the dependent variable (*Risk_Class*) is converted into a 3-levels factor variable, reflecting its role as a multiclass classification target.

```
# --- Convert Variables to Numeric and Factor ---
numeric_vars <- c("ESG_SCORE","ENVIRONMENTAL_SCORE","SOCIAL_SCORE",
                  "GOVERNANCE_SCORE","SP_R_ESG_RANK","SA_ESG_RISK_SCR",
                  "ISS_QUALITYSCORE","SHORT_AND_LONG_TERM_DEBT",
                  "RETURN_ON_ASSET","EBITDA_TO_REVENUE","CUR_RATIO",
                  "INTEREST_COVERAGE_RATIO","SALES_GROWTH")

Data[numeric_vars] <- lapply(Data[numeric_vars], as.numeric)
Data$Risk_Class <- as.factor(Data$Risk_Class)
```

Code 4: Conversion of the explanatory variables into a numeric structure and target variable into a factor variable

2.5 TRAINING – TEST SET SPLIT

The next step is dividing the dataset into a training set and a test set to evaluate the predictive performance of the classification models. This step is essential as it allows to assess the predictive power of the models on unseen observations. In particular, the split is performed following a time-based criterion: all the observations from October 2020 until the end of 2024 are assigned to the training set, while the observations pertaining to the year 2025 are assigned to the test set, such that the most recent data are reserved for model validation. Hence, historical information is used to develop the models, while the recent information is used to test the predictions and verify the accuracy of the models employed.

Given to operational needs, in the code, the date variable is first converted into a proper date format, the dataset is the partitioned into training and test subsets based on the specified cut-off date, then the date column is removed from both the datasets, as it's no longer useful for further analysis.

```
61  # --- Split Data into Training and Test Sets ---
62  Data$DATE <- as.Date(Data$DATE, format = "%Y-%m-%d")
63  train <- subset(Data, DATE < as.Date("2025-01-01"))
64  test  <- subset(Data, DATE >= as.Date("2025-01-01"))
65  train <- train[, -1]
66  test  <- test[, -1]
```

Code 5: Train-test set splitting and removal of the date column

After the data cleaning and the temporal split, the resulting datasets contain the following number of observations: 18.065 observations (monthly frequency data) in the training set and 3.298 observations in the test set.

Below the first rows of the training dataset are displayed, to provide a concrete view of the dataset composition and the variables:

```
> head(train)
     ESG_SCORE ENVIRONMENTAL_SCORE SOCIAL_SCORE GOVERNANCE_SCORE SP_R_ESG_RANK SA_ESG_RISK_SCR ISS_QUALITYSCORE
5502      2.80                1.81         1.04             7.55             7          44.50                9
5564      4.12                5.73         2.24             6.93            45          20.97                1
5612      4.12                3.62         2.50             6.68            51          29.18                4
5901      5.85                5.10         5.59             7.19           100          19.42                4
5909      4.91                4.48         5.37             5.35            44          28.18                3
5920      3.61                2.46         3.54             5.49            44          28.77                6
     SHORT_AND_LONG_TERM_DEBT RETURN_ON_ASSET EBITDA_TO_REVENUE CUR_RATIO INTEREST_COVERAGE_RATIO SALES_GROWTH Risk_Class
5502                   1538.2          4.8249           20.1974    1.2889                 5.1906      10.0062          3
5564                    579.1          3.5324            3.5496    1.3922                18.4507       2.8101          1
5612                   2506.0          1.0495            6.8991    1.2674                 0.4208       5.9022          2
5901                   3736.0         -9.2071          -27.7497    1.1754                -5.8317       5.2562          3
5909                  15143.0          4.4847           20.6776    1.7119                10.5674      56.2059          2
5920                   1852.2          2.9691           12.5686    1.4553                 8.8545      -6.4748          2
```

Code 6: First rows of the training set

2.6 EXPLORATORY ANALYSIS – CORRELATION MATRIX

Before implementing the models, a preliminary exploratory analysis is conducted to examine the relationships among the explanatory variables. In particular, a correlation analysis to assess the level of association between traditional financial variables and ESG-related variables is performed. The main reasons are to identify potential multicollinearity, which occurs when two or more predictor variables in a regression model are highly correlated, that could affect the model effectiveness, especially the linear ones; consequently, to verify whether ESG variables provide overlapping information with conventional financial measures, confirming their potential incremental value in credit risk assessment.

The correlation matrix shows that there is almost no correlation between ESG and financial variables, suggesting that sustainability factors capture dimensions of risk that are not directly explained by standard accounting ratios. Stronger correlations emerge within the ESG dimension itself, in particular a positive correlation between the ESG score and its subcomponents, as expected, and a negative correlation between the ISS Quality Score and the ESG scores, indicating that ESG metrics provide useful additional and complementary information for the analysis.

A subset of the entire dataset is created (*num_vars*) to only select the numerical variables, which corresponds to all the explanatory variables.

Below the code relating to the correlation analysis and its output are provided:

```r
# --- Correlation Matrix between Numeric Variables ---
num_vars <- Data %>% select_if(is.numeric)
corr_matrix <- cor(num_vars, use = "pairwise.complete.obs")

corrplot(corr_matrix,
         method = "color",
         type = "upper",
         tl.cex = 0.8, tl.col = "black",
         addCoef.col = "black",
         number.cex = 0.7,
         title = "Correlation Matrix: ESG and Financial Variables")

corr_table <- round(corr_matrix, 3)
```

Code 7: Correlaton matrix between the predictior variables code

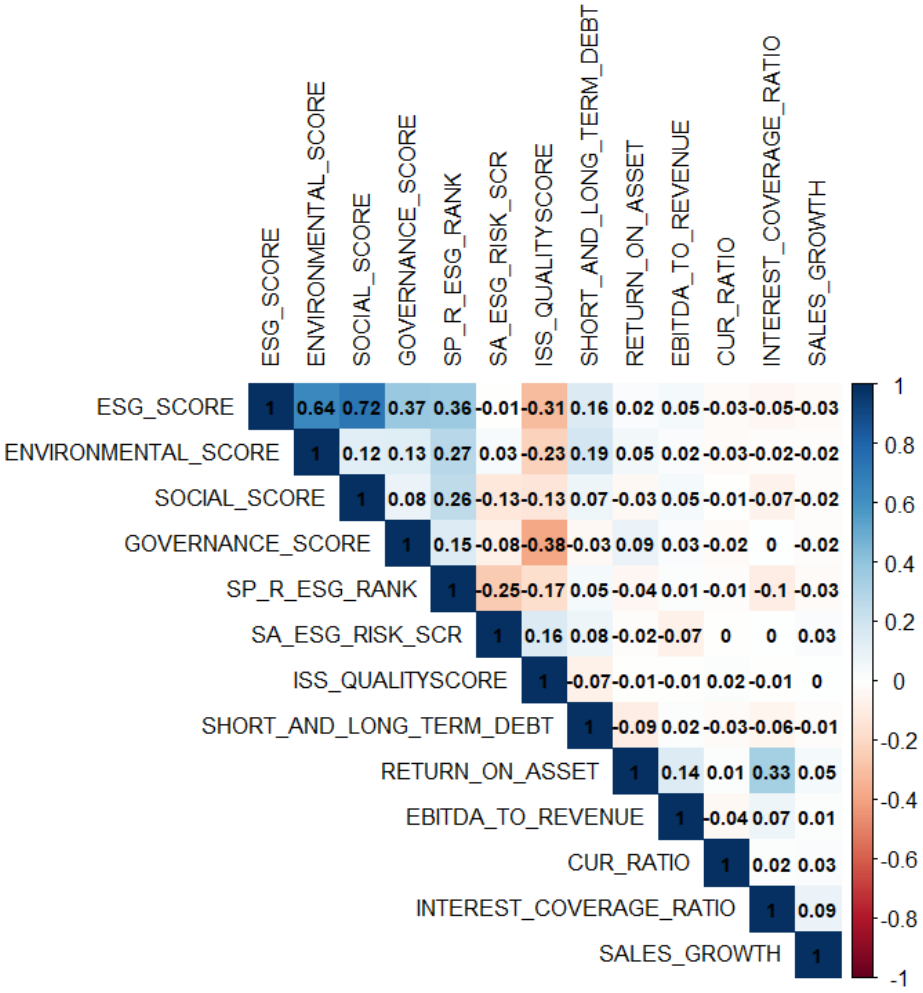

Figure 5: Correlation matrix

2.7 DEVELOPMENT OF THE MODELS AND RESULTS

This section outlines the predictive models employed in the credit risk classification. The aim is to assess the predictive performance of each model in classifying the firms into the correct credit risk class and evaluate the contribution of each predictor, with a focus on ESG-related variables. The estimation of results, performance metrics and graphical outputs are presented and discussed in the following subsections.

2.7.1 MULTINOMIAL LOGISTIC REGRESSION

The first model employed in the analysis is Multinomial Logistic Regression (MLR), which is a linear model particularly suitable for multiclass classification problems. Before estimating the model, the explanatory variables are organized in a matrix *X*, while the dependent variable is stored in a vector *y*, both for the training and the test set, as required by the *glmnet* framework in R.

```
# --- Define Predictors and Target Variable ---
x_train <- as.matrix(train[, 1:13])
y_train <- train[, 14]
x_test  <- as.matrix(test[, 1:13])
y_test  <- test[, 14]
```

Code 8: Defining the appropriate structure for the varaibles to fit the model

The model is estimated using *cv.glmnet*, implementing 10-fold Cross-validation and LASSO penalty: LASSO penalty, which is based on the L1 norm of the vector of the coefficients, shrinks less informative coefficients, eventually to zero, and remove variables that do not contribute to predictive performance while Cross-validation is used to identify the best level of

penalisation by selecting the value of the tuning parameter λ which minimizes the multinomial deviance.

```
 99  # --- Fit Model with Cross-Validation to Select Lambda ---
100  mlr_fit <- cv.glmnet(x_train, y_train,
101                       family = "multinomial",
102                       type.multinomial = "grouped",
103                       alpha = 1)
104
105  cat("Optimal lambda:", mlr_fit$lambda.min, "\n")
106  coef(mlr_fit)
107
108  # --- Plots ---
109  plot(mlr_fit)                              # Cross-validation curve
110  plot(mlr_fit$glmnet.fit, label=TRUE)       # Penalized coefficients
```

Code 9: MLR fitting, coefficients of predictor variables for each class, optimal lambda values, cross validation and LASSO regularization outputs

In this case, the selected λ value is very low, indicating a limited penalization degree, suggesting that several predictors carry relevant information for the credit risk classification.

```
> cat("Optimal lambda:", mlr_fit$lambda.min, "\n")
Optimal lambda: 0.0002205648
```

Code 10: Value of lambda which minimizes the multinomial deviance

The limited penalization effect is also reflected in the coefficients, as no coefficient is removed by the LASSO penalization. Each coefficient represents the change in the *log-odds* of belonging to a specific class associated with a one unit increase in the corresponding predictor, holding all the other predictors constant. A positive coefficient increases the log-odds, therefore the probability of belonging to a specific class, whereas a negative coefficient decreases the log-odds relative to that class, larger absolute values of the coefficients indicate stronger (positive or negative) effects. Below the estimated coefficients for each class are provided:

```
> coef(mlr_fit)
$`1`
14 x 1 sparse Matrix of class "dgCMatrix"
                                       1
(Intercept)                -5.755400e-02
ESG_SCORE                  -4.607614e-02
ENVIRONMENTAL_SCORE        -3.419686e-02
SOCIAL_SCORE               -5.370355e-02
GOVERNANCE_SCORE            7.370726e-02
SP_R_ESG_RANK               6.023537e-03
SA_ESG_RISK_SCR            -1.267104e-02
ISS_QUALITYSCORE            2.600684e-03
SHORT_AND_LONG_TERM_DEBT   -1.721795e-05
RETURN_ON_ASSET             1.033465e-01
EBITDA_TO_REVENUE          -3.103143e-03
CUR_RATIO                   8.526495e-02
INTEREST_COVERAGE_RATIO     3.318506e-04
SALES_GROWTH               -2.852376e-06

$`2`
14 x 1 sparse Matrix of class "dgCMatrix"
                                       1
(Intercept)                -2.849773e-01
ESG_SCORE                   1.663091e-01
ENVIRONMENTAL_SCORE        -5.385181e-02
SOCIAL_SCORE               -1.110766e-01
GOVERNANCE_SCORE           -1.164995e-01
SP_R_ESG_RANK              -1.586252e-03
SA_ESG_RISK_SCR             2.918894e-02
ISS_QUALITYSCORE            5.744139e-05
SHORT_AND_LONG_TERM_DEBT    3.618476e-06
RETURN_ON_ASSET             6.177034e-02
EBITDA_TO_REVENUE          -1.373547e-03
CUR_RATIO                   8.762998e-02
INTEREST_COVERAGE_RATIO     1.248603e-03
SALES_GROWTH               -3.534234e-05


$`3`
14 x 1 sparse Matrix of class "dgCMatrix"
                                       1
(Intercept)                 3.425313e-01
ESG_SCORE                  -1.202330e-01
ENVIRONMENTAL_SCORE         8.804868e-02
SOCIAL_SCORE                1.647801e-01
GOVERNANCE_SCORE            4.279227e-02
SP_R_ESG_RANK              -4.437285e-03
SA_ESG_RISK_SCR            -1.651789e-02
ISS_QUALITYSCORE           -2.658125e-03
SHORT_AND_LONG_TERM_DEBT    1.359948e-05
RETURN_ON_ASSET            -1.651168e-01
EBITDA_TO_REVENUE           4.476690e-03
CUR_RATIO                  -1.728949e-01
INTEREST_COVERAGE_RATIO    -1.580453e-03
SALES_GROWTH                3.819472e-05
```

Figure 6: Coefficients estimated by the MLR

The Cross-validation plot shows the relationship between the logarithm of the tuning parameter, log (λ), and the multinomial deviance, illustrating the trade-off between model complexity and accuracy. Each red point represents the average deviance across the folds, the left dotted line identifies the optimal value of λ, the one which minimizes the average multinomial deviance, numbers at the top indicates the number of non-zero coefficients for each value of log (λ).

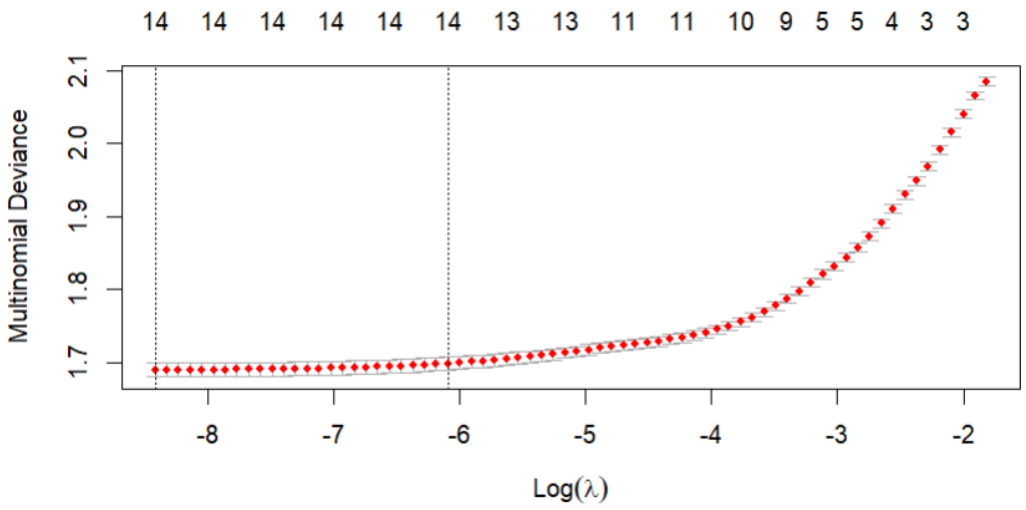

Figure 7: Cross-validation output

The regularization plots display the behaviour of the estimated coefficients related to the penalty parameter changes for each class. Each line describes the evolution of a predictor's coefficient to changes in the regularization strength, which is measured by the L1 norm. In the left side of the plot, which corresponds to low L1 norm values and large values of λ, all the coefficients are reduced to zero, moving right, as the L1 norm increases, regularization weakens and coefficients become part of the model, variables that present the largest growth has the strongest influence on the log-odds of the relative credit risk class.

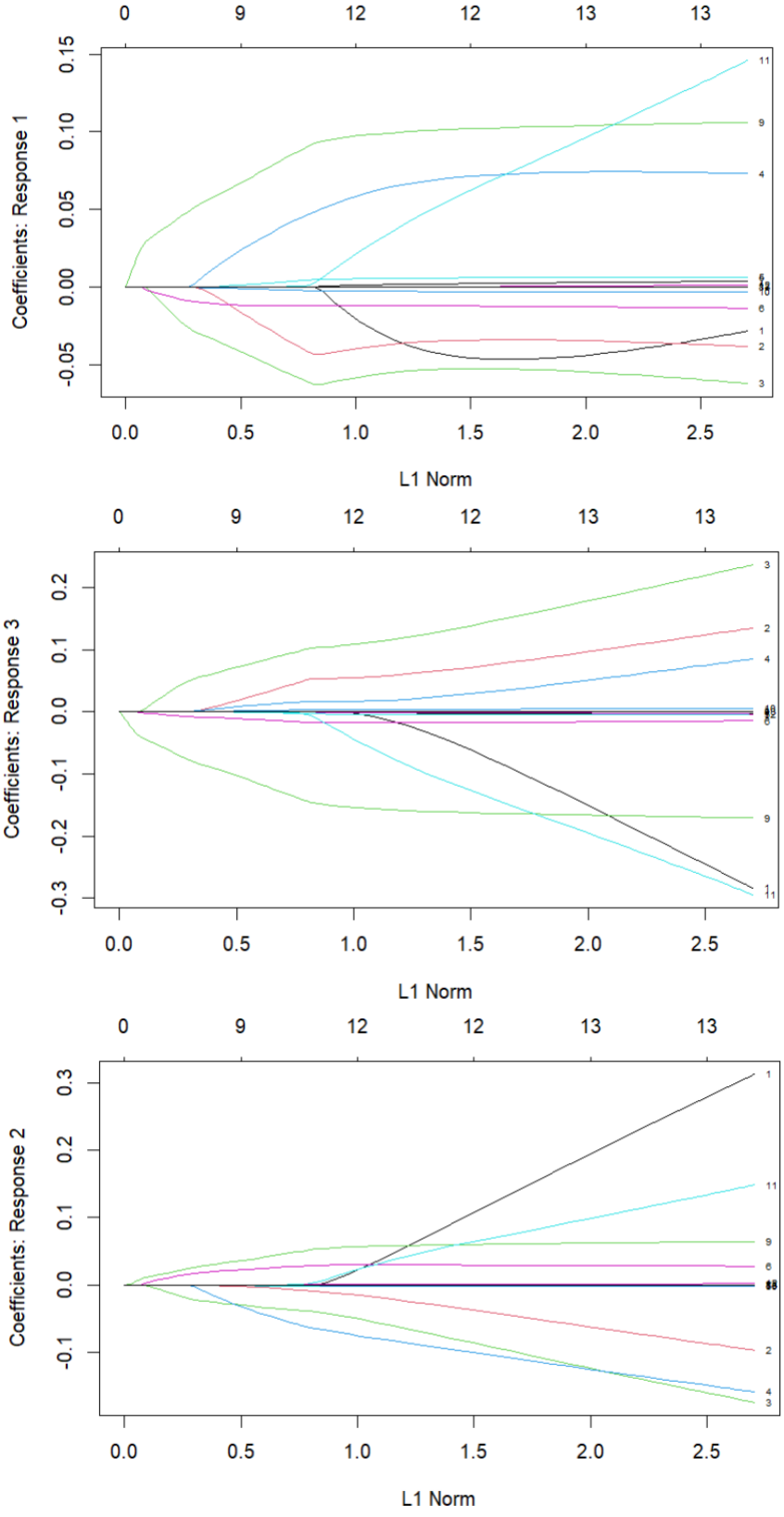

Figure 8: Regularization coefficient plots for each class

The Variable Importance plot ranks the predictors according to their overall contribution to the classification risk, the ranking reflects the coefficient regularization and optimal value of λ. It can be observed that *Current Ratio* and *Return on Asset* dominate with the highest importance indicating they have the strongest influence on classifying firms. ESG-related predictors emerge after, indicating they provide incremental information and a substantial contribution to the model predictions and classification of the firms.

```
112  # --- Variable Importance ---
113  vip(mlr_fit,
114      num_features = 13,
115      lambda = "lambda.min",
116      geom = "col",
117      aesthetics = list(fill = "steelblue"))
```

Code 11: Variable Importance

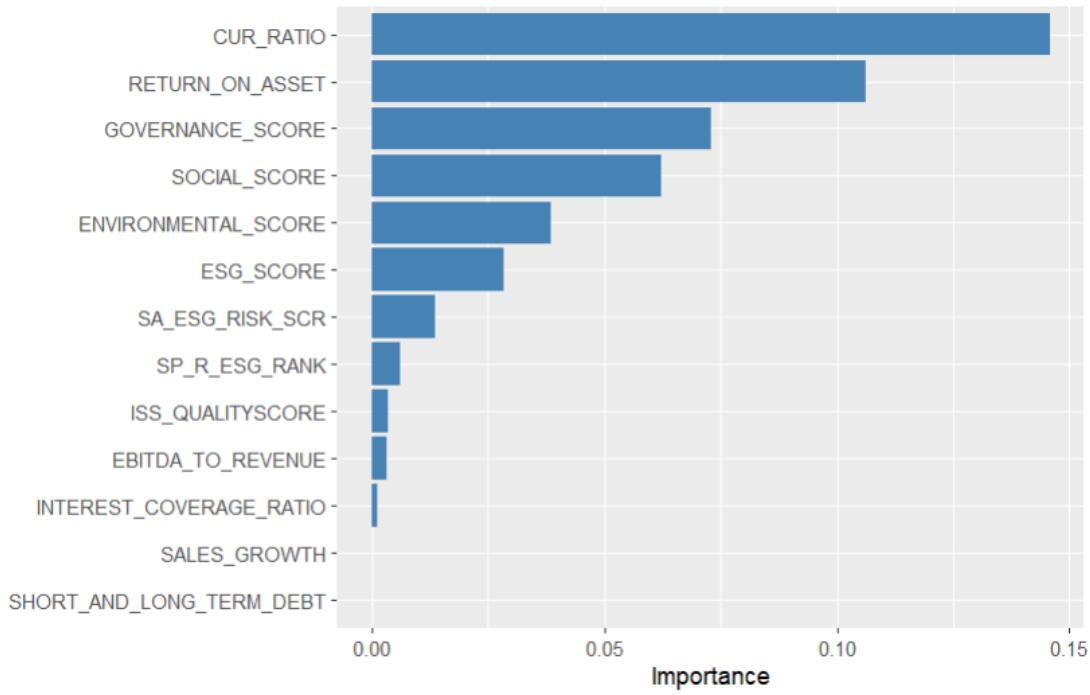

Figure 9: Variable Importance plot

After the estimation of the coefficients, the predictive performance of the model is evaluated on the test set to assess the capacity of the model to classify firms into the correct credit risk category. The predictive accuracy is computed as the ratio between the correctly classified observations and the total of the observations: Multinomial Logistic Regression achieves an overall accuracy of 62.58% and a test error rate of 37.42%, meaning that around two thirds of the observations in the test set are assigned to the correct class.

```
# --- Predictions and Performance Evaluation ---
mlr_pred <- predict(mlr_fit, newx = x_test, s = "lambda.min", type = "class")
confusionMatrix(as.factor(mlr_pred), as.factor(y_test))

# --- Accuracy and Error Rate ---
MLR_accuracy <- mean(mlr_pred == y_test)
MLR_test_error <- 1 - MLR_accuracy
```

```
> cat("MLR Accuracy:", round(MLR_accuracy, 4), "\n")
MLR Accuracy: 0.6258
> cat("MLR Test Error Rate:", round(MLR_test_error, 4), "\n")
MLR Test Error Rate: 0.3742
```

Code 12: MLR accuracy and test error rate

The comparison between predicted and observed classes is summarized by the confusion matrix, which provides a detailed view of classification performance across the three classes, the main diagonal contains the correct classifications. It can be noticed that the model is strongest on the "Safe" class with 1500 observations correctly predicted as safe, while the "Grey Zone" is the class that present the lowest accuracy, with only 89 observations correctly classified.

```
> confusionMatrix(as.factor(mlr_pred), as.factor(y_test))
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3
         1 1500  628  229
         2   72   89   67
         3   64  174  475
```

Code 13: MLR confusion matrix

The results are consistent with the ROC curves, which show a good discrimination for the “Safe” and “Distress” classes, with AUC values of 0.815 and 0.886, and an AUC value of 0.562 for the “Grey Zone” indicating a performance comparable to random classification and confirming that the model presents more difficulty to identify medium risk firms.

```
129  # --- ROC Curves (One-vs-All) ---
130  mlr_probs <- as.data.frame(predict(mlr_fit, newx = x_test,
131                                     s = "lambda.min", type = "response"))
132  mlr_roc1 <- roc(ifelse(y_test == 1, 1, 0), mlr_probs[, 1])
133  mlr_roc2 <- roc(ifelse(y_test == 2, 1, 0), mlr_probs[, 2])
134  mlr_roc3 <- roc(ifelse(y_test == 3, 1, 0), mlr_probs[, 3])
```

Code 14: MLR ROC curves

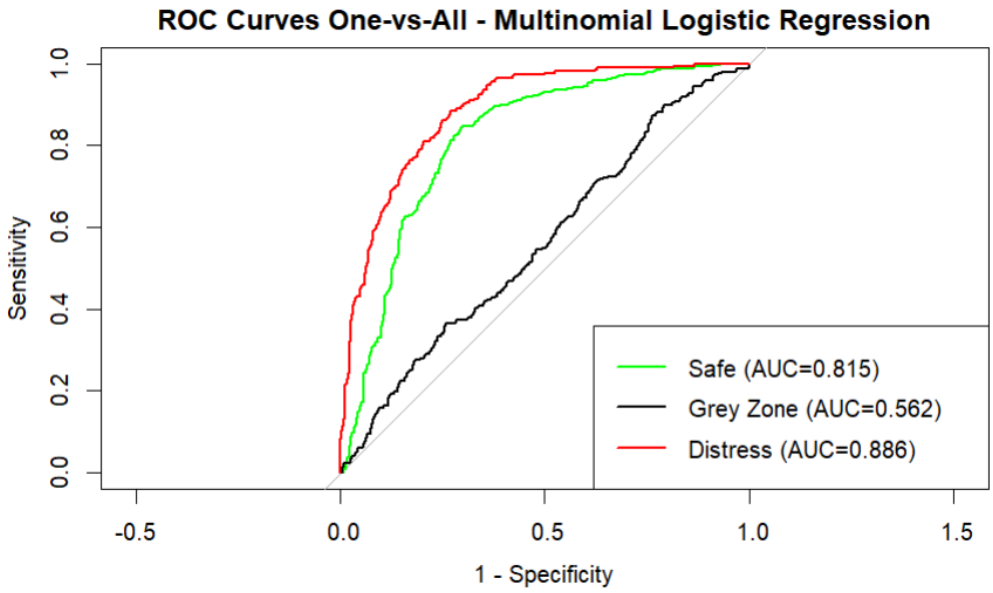

Figure 10: MLR ROC curves output and plot

2.7.2 LINEAR DISCRIMINANT ANALYSIS

This subsection presents the implementation of the second linear model, the Linear Discriminant Analysis (LDA). The model estimates the coefficients relative to the linear combinations (discriminant functions) of predictors that maximize the separation between the three classes. After the model fit, the predicted class labels and class probabilities for the test dataset are generated to evaluate the classification performance. The estimated coefficients represent the weights of each predictor in the linear combinations that define the two discriminant functions. Larger absolute value coefficients have a stronger contribution in the separation of the credit risk classes along that linear discriminant. Specifically, the first discriminant, LD1, mainly drives the separation between "Safe" and "Distress" classes, while LD2 mainly helps differentiate between "Grey Zone" class and the extremes. The sign of each coefficient indicates towards which direction of the LD axis the associated predictor shifts the observation. It can be observed that the coefficients associated to the ESG variables are the highest in absolute value, indicating that sustainability factors play an important role in class separation while financial metrics have a secondary role.

```
153 # --- Fitting the model ---
154 lda_fit <- lda(Risk_Class ~ ., data = train)
155 lda_fit
156 lda_pred <- predict(lda_fit, test)
157 lda_class <- lda_pred$class
```

```
Coefficients of linear discriminants:
                                LD1           LD2
ESG_SCORE                -3.922205e-01 -8.967824e-01
ENVIRONMENTAL_SCORE       2.032550e-01  2.140432e-01
SOCIAL_SCORE              3.473322e-01  4.047580e-01
GOVERNANCE_SCORE          1.041606e-01  5.847860e-01
SP_R_ESG_RANK            -1.137921e-02  6.437640e-03
SA_ESG_RISK_SCR          -4.069938e-03 -1.018212e-01
ISS_QUALITYSCORE         -3.610089e-02  5.338508e-03
SHORT_AND_LONG_TERM_DEBT  2.620923e-05 -4.819005e-06
RETURN_ON_ASSET          -6.844752e-02  4.863056e-03
EBITDA_TO_REVENUE         2.315438e-03 -1.297163e-03
CUR_RATIO                -6.747728e-03 -1.076663e-02
INTEREST_COVERAGE_RATIO   4.062934e-04 -1.718599e-03
SALES_GROWTH              1.442097e-03  1.595785e-03
```

Code 15: LDA fitting and coefficients

The importance of each predictor is also assessed and shown by the Variable Importance plots, which confirms that according to LDA model, ESG-related variables make the greatest contribution in the credit risk classification.

```
# --- Variable Importance (based on Discriminant Coefficients) ---
coef_df <- as.data.frame(abs(lda_fit$scaling))
coef_df$Variable <- rownames(coef_df)

ggplot(coef_df, aes(x = reorder(Variable, LD1), y = LD1)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(title = "Variable Importance (LDA - Discriminant 1)",
       y = "|Coefficient|")

ggplot(coef_df, aes(x = reorder(Variable, LD2), y = LD2)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(title = "Variable Importance (LDA - Discriminant 2)",
       y = "|Coefficient|")
```

Code 16: Variable Importance

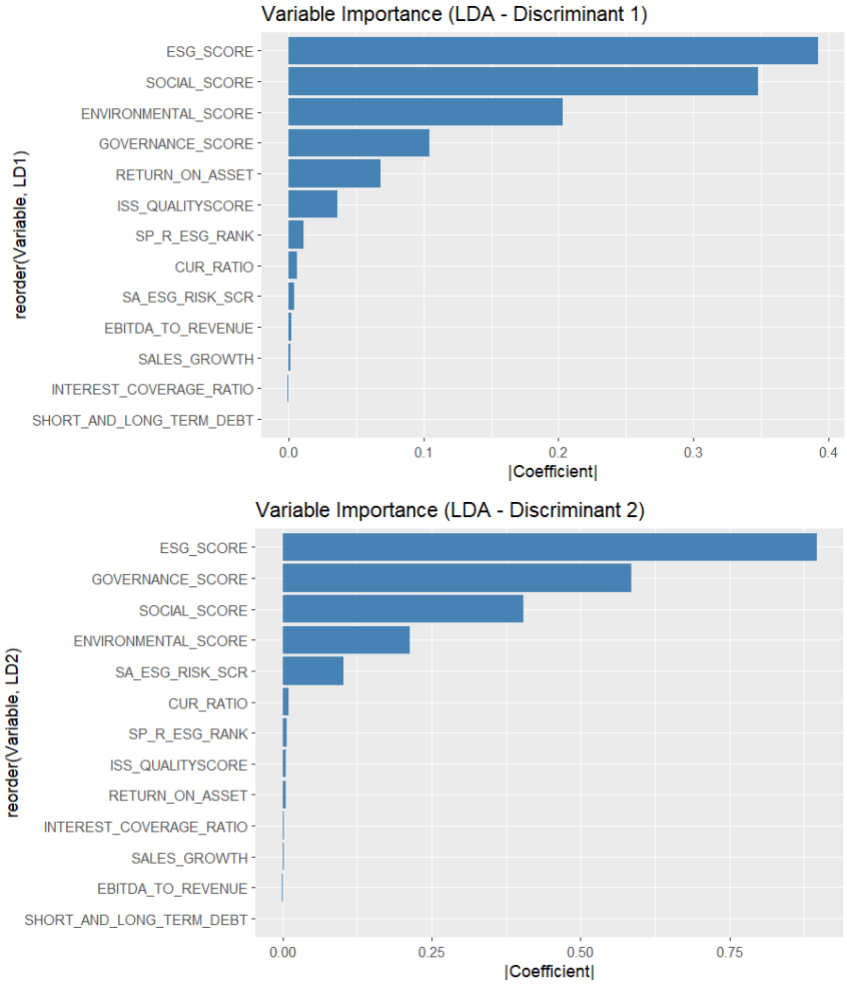

Figure 11: Variable Importance plots

As Multinomial Logistic Regression, LDA model accuracy is computed as the proportion of correctly classified observations in the test set. In this case, the model delivers a 57.1% accuracy and a test error rate of 42.9%, so it performs slightly worse than MLR. Linear Discriminant Analysis still captures structure in the data, but misclassifications remain substantial.

```
# --- Performance Evaluation ---
conf_matrix_lda <- confusionMatrix(as.factor(lda_class),
                                   as.factor(test$Risk_Class))
print(conf_matrix_lda)

# --- Accuracy and Error Rate ---
LDA_accuracy <- mean(lda_class == test$Risk_Class)
LDA_test_error <- 1 - LDA_accuracy
```

```
> cat("LDA Accuracy:", round(LDA_accuracy, 4), "\n")
LDA Accuracy: 0.571
> cat("LDA Test Error Rate:", round(LDA_test_error, 4), "\n")
LDA Test Error Rate: 0.429
```

Code 17: LDA accuracy and test error rate

The confusion matrix provides a detailed comparison of how the model classifies each risk class and identifies potential misclassifications. LDA also strongly favours for the "Safe" class with 1519 correct classifications, but many true "Grey Zone" and "Distress" observations are instead classified as "Safe", showing high sensitivity but low specificity and indicating a tendency to underestimate the credit risk. "Grey Zone" is again the most difficult class with only 85 correctly classified observations, it's slightly better for the "Distress" class with 279 correct classifications though many distressed firms are misclassified as "Safe", which is a major issue in credit risk rating.

```
> conf_matrix_lda <- confusionMatrix(as.factor(lda_class),
+                                    as.factor(test$Risk_Class))
> print(conf_matrix_lda)
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3
         1 1519  708  440
         2   52   85   52
         3   65   98  279
```

Code 18: LDA confusion matrix

ROC curves and AUC values indicate moderate discriminative ability, with a good performance at separating “Safe” and “Distress” observations, with an AUC equal to 0.738 and 0.791, though the performance is a bit weaker than MLR model. “Grey Zone” with an AUC value of 0.582 is again the most challenging class to accurately predict.

```
# --- ROC Curves (One-vs-All) ---
lda_probs <- lda_pred$posterior
lda_roc1 <- roc(ifelse(y_test == 1, 1, 0), lda_probs[, 1])
lda_roc2 <- roc(ifelse(y_test == 2, 1, 0), lda_probs[, 2])
lda_roc3 <- roc(ifelse(y_test == 3, 1, 0), lda_probs[, 3])
```

Code 19: LDA ROC Curves

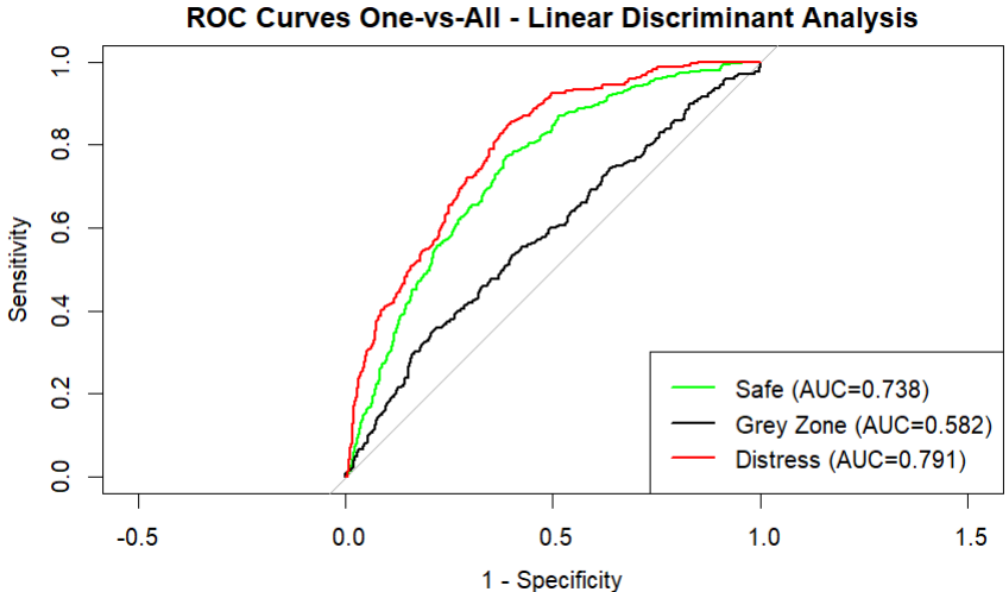

Figure 12: LDA ROC Curves output and plot

Overall, the two linear models provide evidence that both financial fundamentals and ESG-related variables are key drivers of credit risk, confirming that sustainability information contributes to credit risk classification. However, MLR and LDA models exhibit limited predictive accuracy, especially in distinguishing intermediate risk observations. These

results indicate that linear models may be unable to completely capture the complexity of firms' credit risk, these limitations motivate the move to non-linear models as Classification Trees and Random Forest to assess whether they can achieve better predictive performance and better handle the intermediate risk class.

2.7.3 CLASSIFICATION TREE

The first non-linear approach is based on Classification Trees; this model recursively partitions the predictor space into homogeneous regions and predicts the class of an observation by following a sequence of decision rules based on the explanatory variables. Unlike linear models, Classification Trees may be able to capture non-linear relationships and interaction effects among the predictors, leading to a better predictive accuracy. The predictors are selected through Cross-validation based on the misclassification error and the optimal tree size is determined by identifying the number of terminal nodes that minimizes the cross-validated deviance. The optimal fitted tree consists of 8 terminal nodes, a structure that balances interpretability and predictive performance. It provides a transparent and interpretable set of decision rules, highlighting how specific financial ratios and ESG indicators jointly contribute to the credit risk classification. The tree construction reveals that only a subset of predictors is effectively used in the binary splitting. The most important split is based on *Return on Assets*, indicating profitability as a key driver of credit risk, the consequent splits involve *Current Ratio*, *Short and Long-term Debt* and *EBITDA to Revenue*. Among ESG factors, *Social Score* emerges as the only sustainability variable directly used in the tree, indicating that sustainability information can play a

complementary role in credit risk assessment in combination with financial fundamentals.

```
# --- Fitting Pruned Tree via Cross-Validation ---
cv_tree <- cv.tree(tree_fit, FUN = prune.misclass)
plot(cv_tree$size, cv_tree$dev, type = "b",
     xlab = "Number of Terminal Nodes", ylab = "Deviance",
     main = "Cross-Validation for Tree Pruning")

best_size <- cv_tree$size[which.min(cv_tree$dev)]
pruned_tree <- prune.misclass(tree_fit, best = best_size)
summary(pruned_tree)
```

```
> summary(pruned_tree)

Classification tree:
tree(formula = Risk_Class ~ ., data = train)
Variables actually used in tree construction:
[1] "RETURN_ON_ASSET"          "CUR_RATIO"                "SOCIAL_SCORE"
[4] "SHORT_AND_LONG_TERM_DEBT" "EBITDA_TO_REVENUE"
Number of terminal nodes:  8
Residual mean deviance:  1.543 = 27870 / 18060
Misclassification error rate: 0.3408 = 6157 / 18065
```

Code 210: Classification Tree fitting and summary

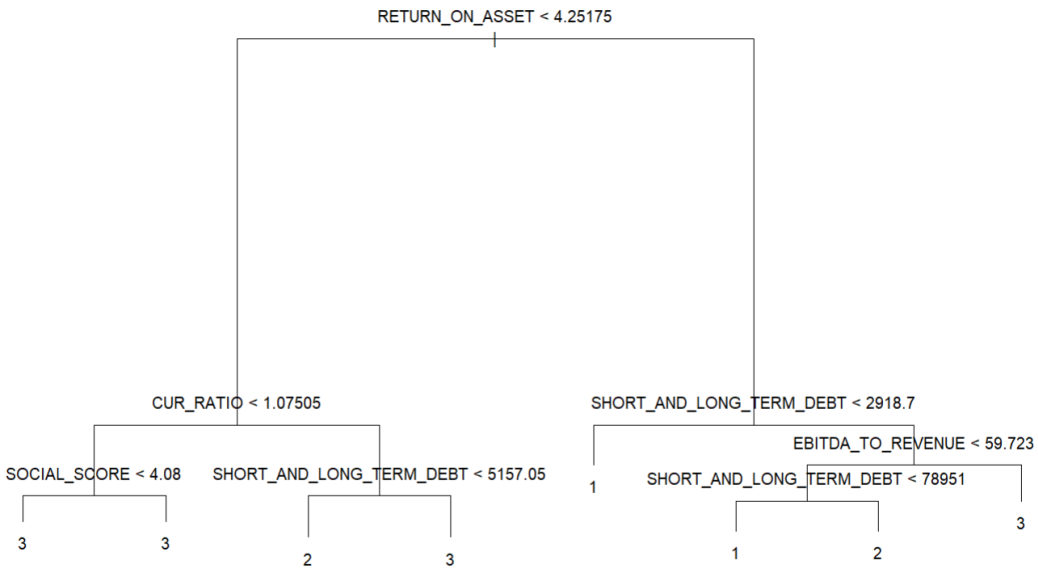

Figure 13: Classification Tree plot

The pruned tree achieves an overall accuracy of 61.55% on the test set and an error rate of 38.45%, a comparable, but not superior, level of performance obtained with the linear models previously considered.

```
# --- Prediction and Performance for Pruned Tree ---
pruned_tree_pred <- predict(pruned_tree, newdata = test, type = "class")
confusionMatrix(pruned_tree_pred, y_test)

tree_pruned_acc <- mean(pruned_tree_pred == y_test)
tree_pruned_err <- 1 - tree_pruned_acc
```

```
> cat("Pruned Tree Accuracy:", round(tree_pruned_acc, 4), "\n")
Pruned Tree Accuracy: 0.6155
> cat("Pruned Tree Test Error Rate:", round(tree_pruned_err, 4), "\n")
Pruned Tree Test Error Rate: 0.3845
```

Code 21: Classification Tree accuracy and test error rate

The Confusion matrix reveals that Classification Tree performs well in identifying firms in "Safe" and "Distress" categories, but the model struggles with classification of "Grey Zone" observations, as well as the linear models. This indicates that, although the Tree can capture some non-linear relationships in the data, it finds difficulty to clearly separate firms with intermediate risk profiles, with only 163 correct classifications.

```
> confusionMatrix(pruned_tree_pred, y_test)
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3
         1 1245  427   80
         2  235  163   69
         3  156  301  622
```

Code 22: Classification Tree confusion matrix

The ROC Curves further confirm this result. The model exhibits a strong discriminative performance for "Distress" firms with an AUC value of 0.870 and "Safe" observations with an AUC value of 0.776. The ROC Curve shows

a little improved “Grey Zone” discrimination surpassing the 0.6 threshold with an AUC of 0.607, but still a moderately low performance.

```
# --- ROC Curves (One-vs-All) ---
tree_probs <- predict(pruned_tree, newdata = test, type = "vector")

tree_prob_class1 <- tree_probs[, "1"]
tree_prob_class2 <- tree_probs[, "2"]
tree_prob_class3 <- tree_probs[, "3"]

tree_roc1 <- roc(ifelse(y_test == 1, 1, 0), tree_prob_class1)
tree_roc2 <- roc(ifelse(y_test == 2, 1, 0), tree_prob_class2)
tree_roc3 <- roc(ifelse(y_test == 3, 1, 0), tree_prob_class3)
```

Code 23: Classification Tree ROC Curves

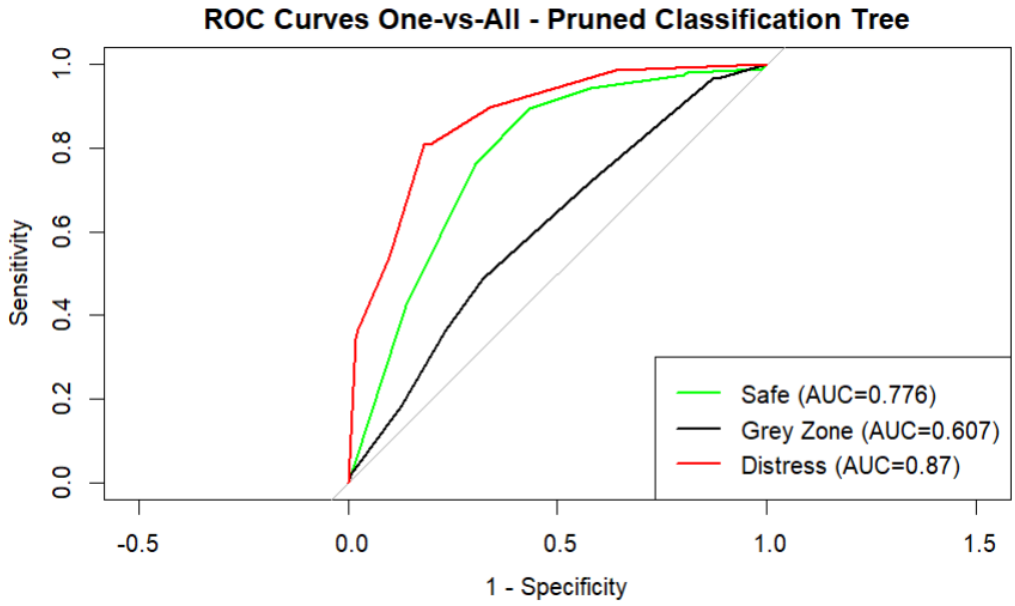

Figure 14: Classification Tree ROC Curves output and plot

Overall, despite the Classification Tree introduces non-linear decision boundaries and further interaction effects between financial and ESG variables, the model has not achieved a clear improvement in predictive performance with respect to the linear models, this result motivates the transition to more complex and flexible model like Random Forest.

2.7.4 RANDOM FOREST

Random Forest is employed to overcome the limitations of single Classification Tree and linear models. This model is an extension of Classification Tree as it aggregates a large number of decision trees: each tree is constructed based on a bootstrap selection of the training data and on a random sample of the predictors involved to tree fitting at each split. The error plot reports the evolution of the OOB error as the number of trees increases, the curves drop sharply within the first trees and remain stable reaching a very low error level, indicating that 500 trees are enough to ensure stability and absence of overfitting and adding more trees would not change performance or increase accuracy.

```
260 # --- Setting the seed and fitting the model ---
261 set.seed(100)
262 randfor_fit <- randomForest(Risk_Class ~ ., data = train, importance = TRUE)
263 plot(randfor_fit)
```

Code 24: Random Forest fitting

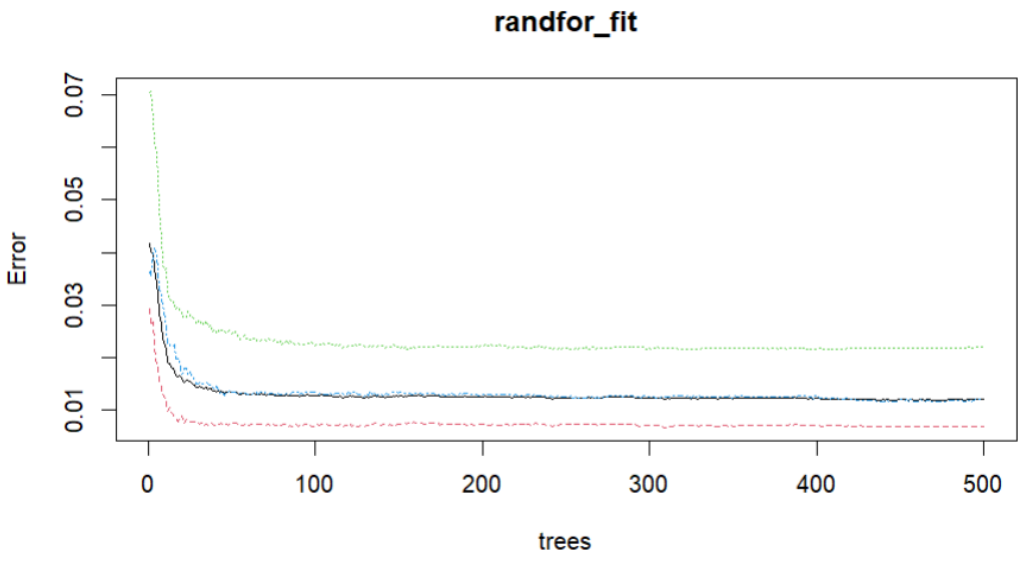

Figure 15: Random Forest Error plot and selection of the number of trees

The Variable Importance analysis shows which predictors are the drivers of the model performance. It is based on two measures: Mean Decrease Accuracy and Mean Decrease in Gini. According to both measures, financial indicators remain key determinants of credit risk classification, as *Return on Asset*, leverage measures, liquidity and profitability indicators ranks among the most influential variables. At the same time, ESG variables play a non-negligible role: in particular, the *Environmental Score* is a very relevant predictor, especially in terms of classification accuracy, ranking second among predictors. Overall, the other ESG-related variables also display positive importance, though generally lower than core financial variables. The coexistence of financial and ESG variables among the most important predictors indicate that ESG information provides complementary information that is effectively exploited by the Random Forest model.

```
265  # --- Variable Importance ---
266  varImpPlot(randfor_fit, main="Random Forest: Variable Importance")
```

Code 25: Random Forest variable importance

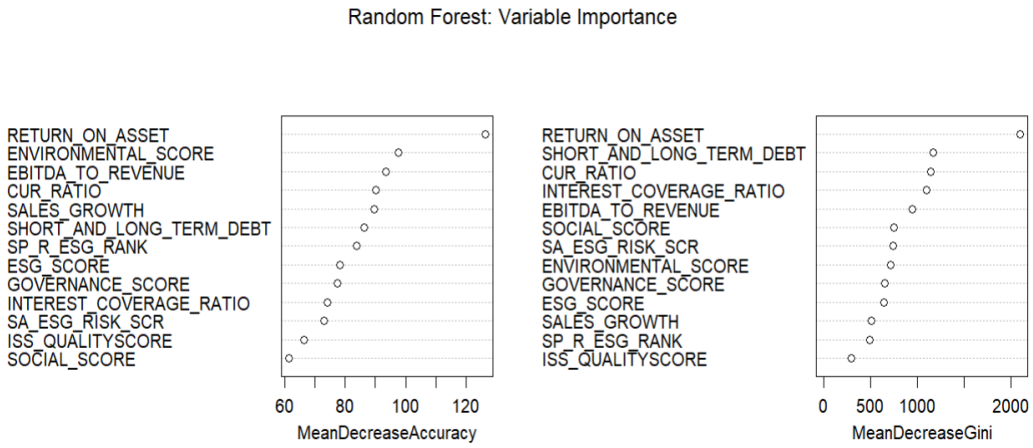

Figure 16: Random Forest variable importance output

In terms of predictive accuracy, the Random Forest significantly outperforms all the previously estimated models, delivering a substantial improvement with an accuracy of 89.57% and a test error rate of 10.43%.

```
# --- Prediction and Performance ---
rf_pred <- predict(randfor_fit, newdata = test)
confusionMatrix(rf_pred, y_test)

rf_acc <- mean(rf_pred == y_test)
rf_err <- 1 - rf_acc
cat("Random Forest Accuracy:", round(rf_acc,4), "\n")
cat("Random Forest Test Error Rate:", round(rf_err,4), "\n")
```

Code 26: Random Forest accuracy and test error rate

The confusion matrix shows enhanced classification performance across all the three credit risk categories, with a significant improvement for the "Grey Zone" risk class with 697 correct classifications, overcoming the limitations of the other models.

```
> confusionMatrix(rf_pred, y_test)
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3
         1 1533  123   11
         2   88  697   36
         3   15   71  724
```

Code 27: Random Forest confusion matrix

Graphically, the one-vs-all ROC Curves confirm the excellent discriminative ability of the Random Forest model, with an AUC value close to one for the "Distress" class and very high AUCs also for "Safe" and "Grey Zone" categories. These values indicate a strong ability of the model to correctly classify the observations, even in the intermediate risk category where the other models had more difficulty.

```
277  # --- ROC Curves (One-vs-All) ---
278  rf_probs <- predict(randfor_fit, newdata = test, type = "prob")
279  rf_roc1 <- roc(ifelse(y_test == 1,1,0), rf_probs[, 1])
280  rf_roc2 <- roc(ifelse(y_test == 2,1,0), rf_probs[, 2])
281  rf_roc3 <- roc(ifelse(y_test == 3,1,0), rf_probs[, 3])
```

Code 28: Random Forest ROC Curves

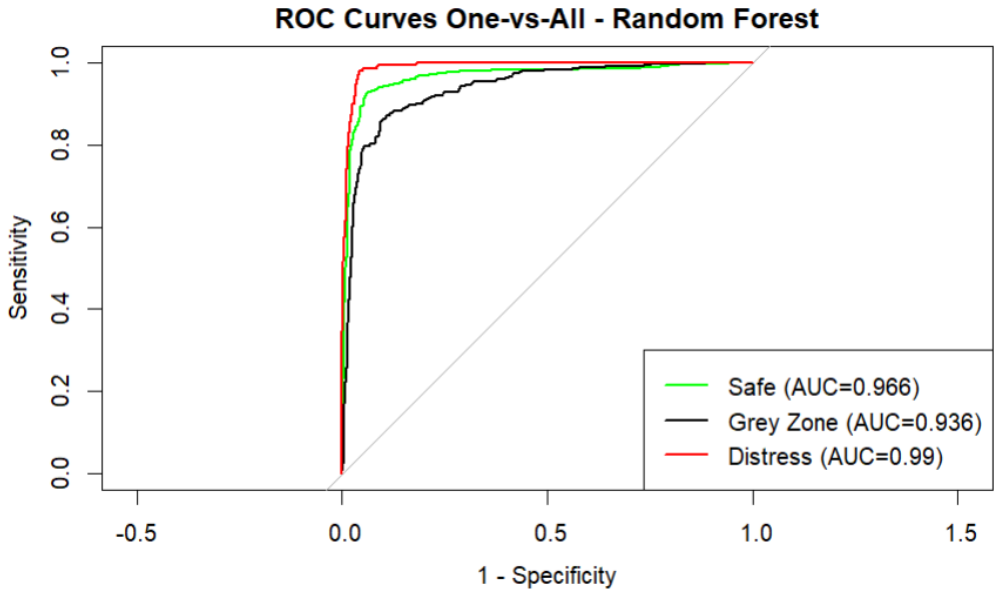

Figure 17: Random Forest ROC Curves plot and output

Overall, Random Forest results show that this model substantially improves credit risk classification performance by effectively capturing the non-linearity and the interaction effects among the predictors. These findings also indicate that ESG information are very informative, in particular environmental factors, which emerge as a relevant credit risk drivers alongside traditional financial indicators.

2.7.5 S.A.F.E. A.I. METRICS

In order to provide a deeper assessment of model performances beyond standard evaluation metrics, this section introduces SAFE AI metrics. While traditional evaluation metrics focus on the correct assignment of firms to three risk categories ("Safe", "Grey Zone" and "Distress"), these measures assess whether the predicted probabilities generate a meaningful ordering of firms in terms of underlying financial soundness. Rank Graduation Accuracy (RGA) is involved to evaluate the ranking accuracy of model outputs, using the continuous Altman Z-Score as a benchmark, in addition, Rank Graduation Explainability is employed to quantify the contribution of ESG variables to model predictions.

A second version of the dataset is constructed, keeping the continuous Altman Z-Score values. The same data cleaning and preprocessing of data has been adopted as for the first dataset.

```
# --- Creation of a second version of the dataset containing continuous ---
# --- Altman Z-Score for SAFE AI metrics computation ---
Data_2 <- as.data.frame(Dati_progetto)

# --- Data cleaning, numeric conversion and train-test split ---
Data_2$DATE <- as.character.Date(Data_2$DATE)
Data_2[Data_2 == "n.a."] <- NA
Data_2 <- na.omit(Data_2)
Data_2 <- Data_2[, -1]
numeric_vars_2 <- c("ALTMAN_Z_SCORE","ESG_SCORE","ENVIRONMENTAL_SCORE",
                    "SOCIAL_SCORE","GOVERNANCE_SCORE","SP_R_ESG_RANK",
                    "SA_ESG_RISK_SCR","ISS_QUALITYSCORE",
                    "SHORT_AND_LONG_TERM_DEBT","RETURN_ON_ASSET",
                    "EBITDA_TO_REVENUE","CUR_RATIO",
                    "INTEREST_COVERAGE_RATIO","SALES_GROWTH")

Data_2[numeric_vars_2] <- lapply(Data_2[numeric_vars_2], as.numeric)
Data_2$DATE <- as.Date(Data_2$DATE, format = "%Y-%m-%d")
train_2 <- subset(Data_2, DATE < as.Date("2025-01-01"))
test_2  <- subset(Data_2, DATE >= as.Date("2025-01-01"))
train_2 <- train_2[, -1]
test_2  <- test_2[, -1]
```

Code 29: Creation of the second dataset containing continuous Altman Z-Score values, dataset cleaning and preprocessing

The RGA measures the extent to which the predicted probabilities associated with each class generated by the models induce an ordering of the observations that is coherent with their Altman Z-Scores. RGA is computed separately for each model and for each risk class probability. The metric is calculated on the test-set observations, to be consistent with the traditional evaluation metrics. For the "Safe" class (Class 1), the Multinomial Logistic Regression achieves the highest RGA (0.886), followed by LDA (0.814) and Classification Tree (0.824). Random Forest, despite the very high classification accuracy, displays a lower RGA (0.709), indicating that linear models produce probability estimates that are more effective in ranking financially healthy firms, according to their Altman Z-Score, rather than classifying Safe firms. For the "Grey Zone" (Class 2), RGA values are lower across all the models, with Random Forest attaining the highest RGA (0.507), a result aligned with the classification accuracy, in which it is the only model capable of successfully discriminating intermediate risk companies. For "Distress" class (Class 3), RGA values are low for all the models, ranging between 0.08 and 0.173, indicating a limited ability of "Distress" class probabilities to generate a meaningful ranking within the riskiest firms, as the Altman Z-Score concentrates at low values and difference between firms are less informative from a ranking point of view.

Overall, the RGA highlights a distinction between classification accuracy framework and ranking: Random Forest dominates in terms of predictive accuracy across the three risk classes, but its probability outputs are less aligned with the continuous Altman Z-Score ranking, particularly for the "Distress" class; linear models, instead, despite lower classification accuracy, generate better probabilities for ranking purposes.

```
338
339 # --- RGA Computation ---
340 RGA_mlr_1  <- RGA(test_2$ALTMAN_Z_SCORE, mlr_probs[, 1])
341 RGA_lda_1  <- RGA(test_2$ALTMAN_Z_SCORE, lda_probs[, 1])
342 RGA_tree_1 <- RGA(test_2$ALTMAN_Z_SCORE, tree_probs[, "1"])
343 RGA_rf_1   <- RGA(test_2$ALTMAN_Z_SCORE, rf_probs[, 1])
344
345
346 RGA_mlr_2  <- RGA(test_2$ALTMAN_Z_SCORE, mlr_probs[, 2])
347 RGA_lda_2  <- RGA(test_2$ALTMAN_Z_SCORE, lda_probs[, 2])
348 RGA_tree_2 <- RGA(test_2$ALTMAN_Z_SCORE, tree_probs[, "2"])
349 RGA_rf_2   <- RGA(test_2$ALTMAN_Z_SCORE, rf_probs[, 2])
350
351
352 RGA_mlr_3  <- RGA(test_2$ALTMAN_Z_SCORE, mlr_probs[, 3])
353 RGA_lda_3  <- RGA(test_2$ALTMAN_Z_SCORE, lda_probs[, 3])
354 RGA_tree_3 <- RGA(test_2$ALTMAN_Z_SCORE, tree_probs[, "3"])
355 RGA_rf_3   <- RGA(test_2$ALTMAN_Z_SCORE, rf_probs[, 3])
```

```
323 # --- RGA Function ---
324 RGA <- function(y, yhat){
325   ryhat <- rank(round(yhat,4), ties.method="min")
326   support <- tapply(y, ryhat, mean)
327   rord <- numeric(length(y))
328   for(jj in 1:length(y)){
329     rord[jj] <- support[names(support)==ryhat[jj]]
330   }
331   ystar <- rord[order(yhat)]
332   I <- 1:length(y)
333   conc <- 2*sum(I*ystar)
334   dec <- 2*sum(I*sort(y,decreasing=TRUE))
335   inc <- 2*sum(I*sort(y))
336   (conc-dec)/(inc-dec)
337 }
```

```
> RGA_mlr_1
[1] 0.8858154
> RGA_lda_1
[1] 0.8139878
> RGA_tree_1
[1] 0.8244251
> RGA_rf_1
[1] 0.7093456
> RGA_mlr_2
[1] 0.4539549
> RGA_lda_2
[1] 0.4461977
> RGA_tree_2
[1] 0.3464451
> RGA_rf_2
[1] 0.5069636
> RGA_mlr_3
[1] 0.0802655
> RGA_lda_3
[1] 0.14496
> RGA_tree_3
[1] 0.1727606
> RGA_rf_3
[1] 0.1498687
```

Code 30: Computation and values of Rank Graduation Accuracy

After assessing the ranking accuracy through the RGA, now the focus shifts to the explainability, the objective is to evaluate the informational contribution of the main ESG indicators to the ranking, using the Rank Graduation Explainability (RGE). The analysis is focused on the four most relevant ESG variables: the overall ESG score, Environmental score, Social score and Governance score, following the evidence emerging from the Variable Importance analysis across the models. For each model and for each risk class, RGE is computed by comparing the ranking produced by the full model with the one obtained from a reduced model in which a single ESG variable is removed. The resulting metrics indicates the extent to which the removal of a specific ESG variable modifies the order of the predicted probabilities. Thus, RGE values that are close to zero indicate that removing the variable from the model has a very low impact on the ranking, while higher values, close to one, indicate an important contribution in terms of explainability and that the factor provides significant additional information. The results show very low RGE values for all the ESG variables, for all the models and the risk categories, meaning that the elimination of a single ESG score does not affect the overall ranking structure. This evidence does not imply that ESG factors are irrelevant in credit risk assessment, rather it indicates that contribution of ESG variables operates jointly with financial variables, rather than through a strong marginal effect of any single ESG indicator. Overall, RGE confirms the fact that ESG factors play a complementary role to standard financial indicators: they improve the quality of credit risk of the firms, but they are not sufficient to drive the ranking of firms' creditworthiness on their own. The conclusion is consistent with the results obtained from the classification models and demonstrates the multidimensionality of firms' credit risk which is tied both to their financial health and to sustainability factors.

```r
# --- RGE Function ---
RGEstar<-function(yhat,yhat_xk){
  ryhat_xk<-rank(round(yhat_xk,4),ties.method="min") # ranks of the predicted values
  support<-tapply(yhat,ryhat_xk,mean) # replace the observed target variable value
                                      # corresponding to the same predictive values
                                      # with their mean
  rord<-c(1:length(yhat))
  for(jj in 1:length(yhat))
  {
    rord[jj]<-support[names(support)==ryhat_xk[jj]]
  }
  yhatstar<-rord[order(yhat_xk)] # re-order the observed target variable values
                                 # with respect to the corresponding predicted values
  I<-1:length(yhatstar)
  conc<-2*sum(I*yhatstar) # first term of the RGEstar numerator (concordance)
  dec<-2*sum(I*sort(yhat,decreasing=TRUE)) # second term of the RGEstar numerator
                                           # and denominator (dual Lorenz)
  inc<-2*sum(I*sort(yhat)) # first term of the RGEstar denominator (Lorenz)
  RGEstar<-(conc-dec)/(inc-dec)
}

# --- RGE Computation ---
ESG_vars <- c("ESG_SCORE","ENVIRONMENTAL_SCORE","SOCIAL_SCORE","GOVERNANCE_SCORE")
models <- list(
  MLR  = mlr_probs,
  LDA  = lda_probs,
  Tree = tree_probs,
  RF   = rf_probs
)
classes <- c(1,2,3)  # Safe, Grey Zone, Distress

RGE_results <- data.frame(
  Model    = character(),
  Variable = character(),
  Class    = integer(),
  RGE      = numeric(),
  stringsAsFactors = FALSE
)

for (var in ESG_vars) {
  for (mod_name in names(models)) {
    mod_probs <- models[[mod_name]]
    for (cls in classes) {

# --- Fit reduced model on train without ESG var ---
      if (mod_name == "MLR") {
        x_train_red <- as.matrix(train[, setdiff(colnames(train)[1:13], var)])
        mlr_fit_red <- cv.glmnet(x_train_red, y_train, family="multinomial",
                                 type.multinomial="grouped", alpha = 1)
        yhat_red <- as.data.frame(predict(mlr_fit_red, newx = as.matrix(test[,
                                          setdiff(colnames(test)[1:13], var)]),
                                          s = "lambda.min", type = "response"))[, cls]
        yhat_full <- mod_probs[, cls]
      } else if (mod_name == "LDA") {
        lda_fit_red <- lda(as.formula(paste("Risk_Class ~ . -", var)), data = train)
        yhat_red <- predict(lda_fit_red, test)$posterior[, cls]
        yhat_full <- mod_probs[, cls]
      } else if (mod_name == "Tree") {
        tree_fit_red <- tree(as.formula(paste("Risk_Class ~ . -", var)), data = train)
        yhat_red <- predict(tree_fit_red, test, type = "vector")[, as.character(cls)]
        yhat_full <- mod_probs[, cls]
      } else if (mod_name == "RF") {
        rf_fit_red <- randomForest(as.formula(paste("Risk_Class ~ . -", var)), data = train)
        yhat_red <- predict(rf_fit_red, test, type = "prob")[, as.character(cls)]
        yhat_full <- mod_probs[, cls]
      }
```

```r

# --- Compute RGE ---
      RGEstar_val <- RGEstar(yhat_full, yhat_red)
      RGE_val <- 1 - RGEstar_val

      RGE_results <- rbind(RGE_results, data.frame(
        Variable = var,
        Model    = mod_name,
        Class    = cls,
        RGE      = RGE_val
      ))
    }
  }
}
```

	Variable	Model	Class	RGE
1	ESG_SCORE	MLR	1	0.0011039132
2	ESG_SCORE	MLR	2	0.0139678442
3	ESG_SCORE	MLR	3	0.0015381602
4	ESG_SCORE	LDA	1	0.0014648897
5	ESG_SCORE	LDA	2	0.0168638188
6	ESG_SCORE	LDA	3	0.0087240966
7	ESG_SCORE	Tree	1	0.0000000000
8	ESG_SCORE	Tree	2	0.0000000000
9	ESG_SCORE	Tree	3	0.0000000000
10	ESG_SCORE	RF	1	0.0009994684
11	ESG_SCORE	RF	2	0.0018063205
12	ESG_SCORE	RF	3	0.0011248169
13	ENVIRONMENTAL_SCORE	MLR	1	0.0012222611
14	ENVIRONMENTAL_SCORE	MLR	2	0.0081983981
15	ENVIRONMENTAL_SCORE	MLR	3	0.0037212870
16	ENVIRONMENTAL_SCORE	LDA	1	0.0055092533
17	ENVIRONMENTAL_SCORE	LDA	2	0.0090169638
18	ENVIRONMENTAL_SCORE	LDA	3	0.0178271024
19	ENVIRONMENTAL_SCORE	Tree	1	0.0000000000
20	ENVIRONMENTAL_SCORE	Tree	2	0.0000000000
21	ENVIRONMENTAL_SCORE	Tree	3	0.0000000000
22	ENVIRONMENTAL_SCORE	RF	1	0.0010946914
23	ENVIRONMENTAL_SCORE	RF	2	0.0019072840
24	ENVIRONMENTAL_SCORE	RF	3	0.0011149313
25	SOCIAL_SCORE	MLR	1	0.0027124715
26	SOCIAL_SCORE	MLR	2	0.0172551848
27	SOCIAL_SCORE	MLR	3	0.0069883502
28	SOCIAL_SCORE	LDA	1	0.0151873078
29	SOCIAL_SCORE	LDA	2	0.0188290514
30	SOCIAL_SCORE	LDA	3	0.0311475876
31	SOCIAL_SCORE	Tree	1	0.0080574405
32	SOCIAL_SCORE	Tree	2	0.0585382981
33	SOCIAL_SCORE	Tree	3	0.0102940751
34	SOCIAL_SCORE	RF	1	0.0014465247
35	SOCIAL_SCORE	RF	2	0.0028976979
36	SOCIAL_SCORE	RF	3	0.0014748990
37	GOVERNANCE_SCORE	MLR	1	0.0049741059
38	GOVERNANCE_SCORE	MLR	2	0.0302894650
39	GOVERNANCE_SCORE	MLR	3	0.0005984843

40	GOVERNANCE_SCORE	LDA	1	0.0063731396
41	GOVERNANCE_SCORE	LDA	2	0.0489488256
42	GOVERNANCE_SCORE	LDA	3	0.0062051426
43	GOVERNANCE_SCORE	Tree	1	0.0000000000
44	GOVERNANCE_SCORE	Tree	2	0.0000000000
45	GOVERNANCE_SCORE	Tree	3	0.0000000000
46	GOVERNANCE_SCORE	RF	1	0.0014090186
47	GOVERNANCE_SCORE	RF	2	0.0022422577
48	GOVERNANCE_SCORE	RF	3	0.0013657443

Code 31: Computation and values of Rank Graduation Explainability

CONCLUSIONS

The empirical analysis provides clear evidence on both the main objectives: the performance of the classification models and the role played by ESG factors in credit risk assessment.

The results obtained from the linear models show that both the financial fundamentals and the ESG-related variables contribute to credit risk classification, as the latter rank among the most important predictors, confirming that sustainability factors contain relevant information. However, the overall predictive accuracy of the linear models remains limited, especially in correctly discriminating “Grey Zone” firms, suggesting that linear models are unable to fully capture the complexity of credit risk.

These limitations motivate the adoption of non-linear models. Classification Trees improve flexibility but still show moderate performance, while Random Forest clearly emerges as the best performing model, achieving the highest classification accuracy across all the three risk categories, highlighting its ability to capture non-linear relationships and interactions among predictors. Random Forest also confirms the relevance of ESG information, with Environmental Score emerging as one of the most influential ESG variables, alongside the traditional financial indicators, particularly in terms of Mean Decrease Accuracy.

The SAFE AI metrics provide additional insights into model behaviour. Rank Graduation Accuracy (RGA) exhibits a distinction between classification accuracy and ranking accuracy: while Random Forest dominates in terms of predictive performance, its estimated probabilities are less aligned with the continuous Altman Z-Scores ranking, especially for the “Distress” class. Conversely, linear models, despite lower classification accuracy, generate probability estimates that are more suitable for ranking firms according to

their underlying financial health. Rank Graduation Explainability (RGE) confirms that individual ESG variables do not drive credit risk classification on their own. The low RGE values indicate that ESG factors play a complementary role: they enhance credit risk assessment when integrated with financial variables but are not sufficient to explain firms' creditworthiness alone.

Overall, the results support the conclusion that ESG factors are economically relevant and informative, but their contribution is substantial when they are combined with traditional financial indicators. From a predictive perspective, Random Forest represents the most effective model for credit risk classification in this framework, while linear models remain valuable for interpretability and ranking assessments.

All these findings demonstrate and underline the multidimensional nature of credit risk and the importance of integrating sustainability factors in credit risk evaluation.

BIBLIOGRAPHY

António Afonso, Pedro Gomes and Philipp Rother (2007), What “Hides” Behind Sovereign Debt Ratings?

Bloomberg L.P., Financial and ESG data for Euro STOXX 600 companies, October 2020 to September 2025, Retrieved on 06/10/2025 from Bloomberg Terminal

Bloomberg Terminal. Environmental, Social and Governance (ESG) Scores (2023)

Egidio Palmieri, Greta Benedetta Ferilli, Yener Altunbas, Valeria Stefanelli, Enrico Fioravante Geretto, Business model and ESG pillars: The impacts on banking default risk, International Review of Financial Analysis, Volume 91,2024

Gareth James, Trevor Hastie, Daniela Witten, Robert Tibshirani, An Introduction to Statistical Learning with Applications in R Second Edition, Springer, 2023

Golnoosh Babaei, Paolo Giudici, Emanuela Raffinetti, A Rank Graduation Box for SAFE AI, Expert Systems with Applications, Volume 259

Laura Bonacorsi, Vittoria Cerasi, Paola Galfrascoli, Matteo Manera, ESG Factors and Firms' Credit Risk, Journal of Climate Finance, Volume 6, 2024

Leo Breiman, Random Forests. Machine Learning, Springer (Kluwer Academic Publishers)

Paolo Giudici, Emanuela Raffinetti, SAFE Artificial Intelligence in finance, Finance Research Letters, Volume 56, 2023

Patrycja Chodnicka-Jaworska P, ESG as a Measure of Credit Ratings, Risks 2021, 9

Raffinetti, E, A Rank Graduation Accuracy measure to mitigate Artificial Intelligence risks., Qual Quant 57 (Suppl 2), 2023

https://finance.ec.europa.eu/sustainable-finance/overview-sustainable-finance_en

https://github.com/GolnooshBabaei/safeaipackage/blob/main/R_codes/Simulation_experiment_A.R

https://www.climatepartner.com/it/formazione/glossario/european-sustainability-reporting-standards-esrs

https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

https://www.greenscope.io/it/esg#anchor-4

https://www.greenscope.io/it/esg#anchor-8

https://www.ibm.com/it-it/think/topics/lasso-regression

https://www.ibm.com/it-it/think/topics/linear-discriminant-analysis

https://www.ibm.com/it-it/think/topics/multicollinearity

https://www.ibm.com/it-it/think/topics/scope-1-2-3-emissions

https://www.ibm.com/think/topics/classification-vs-regression

https://www.ibm.com/think/topics/decision-trees

https://www.ibm.com/think/topics/random-forest#684929713

https://www.investopedia.com/terms/a/altman.asp#citation-7

https://medium.com/@mlmind/evaluation-metrics-for-classification-fc770511052d

https://medium.com/data-science/decision-tree-classifier-explained-a-visual-guide-with-code-examples-for-beginners-7c863f06a71e

https://medium.com/data-science/random-forest-explained-a-visual-guide-with-code-examples-9f736a6e1b3c

https://www.sciencedirect.com/topics/computer-science/decision-tree-classifier

https://www.sciencedirect.com/topics/computer-science/random-forest-classifier

https://www.cudocompute.com/blog/overfitting-and-underfitting-in-

machine-learning-causes-indicators-and-how

https://towardsdatascience.com/unlock-the-power-of-roc-curves-intuitive-insights-for-better-model-evaluation/