Facultad de Ciencias Económicas y Empresariales

Explainability with NN2Poly in Neural Networks for Corporate Bankruptcy Prediction

Author: Claudia Valverde Hueso
Supervisor: Jenny Alexandra Cifuentes Quintero

MADRID| March 2026

Resumen

La predicción de la quiebra empresarial constituye una cuestión central en el análisis financiero debido a sus implicaciones económicas y sistémicas. La identificación temprana de empresas en situación de dificultad financiera resulta relevante para inversores, acreedores y reguladores, ya que la insolvencia no solo afecta a los accionistas, sino también a la estabilidad del sistema financiero en su conjunto. Aunque se han desarrollado numerosos modelos predictivos, las Redes Neuronales Artificiales han emergido como uno de los enfoques más precisos para la clasificación de quiebra, dada su capacidad para capturar relaciones complejas y no lineales entre indicadores financieros. No obstante, su limitada interpretabilidad, comúnmente denominada problema de la *black-box*, restringe su aplicabilidad en contextos de toma de decisiones financieras, donde la transparencia y la justificación económica son esenciales. Si bien se han propuesto diversas técnicas de interpretabilidad, muchas se basan en explicaciones post-hoc que no revelan directamente la estructura interna del modelo. En este contexto, el presente trabajo emplea NN2Poly, una técnica que reformula una red neuronal entrenada en una estructura polinómica simbólica, mejorando la interpretabilidad sin comprometer la capacidad predictiva.

La metodología seguida en esta investigación se estructuró en tres fases principales. En primer lugar, se preprocesó y analizó un conjunto de datos de empresas. Posteriormente, se entrenó una red neuronal de clasificación binaria para estimar el riesgo de quiebra, evaluando su rendimiento mediante métricas estándar de clasificación. Finalmente, se aplicó la técnica NN2Poly para transformar la red entrenada en un modelo polinómico, permitiendo un análisis simbólico de la contribución de las variables. Los resultados muestran que el modelo polinómico replica con fidelidad las predicciones de la red neuronal original, confirmando la validez de la transformación. La estructura polinómica extraída expone que la rentabilidad sostenida, capturada a través de ratios como el beneficio bruto sobre activos totales y medidas basadas en el EBITDA, junto con indicadores relacionados con el comportamiento de la deuda, como los días de rotación de pasivos, desempeñan un papel importante en la explicación del riesgo de quiebra. En contraste, los activos y el tamaño empresarial presentan efectos estabilizadores, en línea con la teoría financiera. Estos resultados muestran el potencial de NN2Poly para reducir la brecha entre capacidad predictiva e interpretabilidad en la modelización del riesgo financiero.

Abstract

Corporate bankruptcy prediction remains a central issue in financial analysis due to its economic and systemic implications. The early identification of financially distressed firms is essential for investors, creditors, and regulators, as firm failure affects not only shareholders but also broader financial stability. While numerous predictive models have been developed, Artificial Neural Networks have emerged as one of the most accurate approaches for bankruptcy classification, given their ability to capture complex and nonlinear relationships among financial indicators. Nevertheless, their limited interpretability, commonly referred to as the *black-box* problem, restricts their practical applicability in financial decision-making contexts, where transparency and economic justification are required. Although various interpretability techniques have been proposed, many rely on post-hoc explanations that do not directly uncover the internal structure of the network. In response, this study employs NN2Poly, a technique that reformulates a trained neural network into a symbolic polynomial structure, thereby enhancing interpretability without compromising predictive capacity.

The methodology followed in this research consisted of three main stages. First, a dataset of firms was preprocessed and analyzed. A binary classification neural network was then trained to estimate bankruptcy risk, with performance evaluated using standard classification metrics. Subsequently, the NN2Poly technique was applied to transform the trained network into a polynomial model, enabling a symbolic analysis of feature contributions. The results show that the polynomial model successfully replicates the predictions of the neural network, confirming the fidelity of the transformation. The extracted polynomial structure reveals that sustained profitability, captured by measures such as gross profit relative to total assets and EBITDA-based ratios, together with indicators related to debt behavior, including liabilities turnover days, play a central role in explaining bankruptcy risk. In contrast, retained earnings and firm size exhibit stabilizing effects, consistent with financial theory. These findings demonstrate the potential of NN2Poly to bridge the gap between predictive performance and interpretability in financial risk modeling.

Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Objectives 5

1.2.1 General Objective 5

1.2.2 Specific Objectives 5

1.3 Structure of the Document 6

2 Review of Predictive Models for Corporate Bankruptcy and Interpretability Approaches **7**

2.1 Machine Learning Techniques for Bankruptcy Prediction 7

2.2 Interpretability Techniques for Bankruptcy Prediction 10

3 Methodology for Predictive Modeling and Interpretability in Bankruptcy Prediction Environments **16**

3.1 Data Collection and Preprocessing 16

3.2 Model Training and Application 19

3.2.1 Neural Network Model Architecture 19

3.2.2 Training Process and Optimization 22

3.2.3 Hyperparameter Tuning 23

3.2.4 Model Evaluation 25

3.3 Model Interpretation via NN2Poly 27

3.3.1 NN2Poly Transformation 27

4 Empirical Analysis and Results **31**

4.1 Dataset Selection 31

4.2 Data Preprocessing and Descriptive Analysis 33

4.3 Implementation of the Neural Network Predictive Model 38

4.4 Application of the Interpretability Technique: NN2Poly 42

4.4.1 Final Neural Network Model 43

4.4.2 Polynomial Representation 44

5 Conclusions **50**

References **59**

List of Figures

1.1 Evolution of Global Business Insolvencies (2016–2026f). Own Elaboration. Data Source: (Allianz Trade, 2025). Note: Years marked with "f" indicate forecasted data. 2
1.2 Annual Change in Insolvencies by Major European Economies (2023–2025f). Own Elaboration. Data Source: (Allianz Trade, 2025). Note: Years marked with "f" indicate forecasted data. 3
1.3 Corporate Insolvencies in Spain (2019–2025f). Own Elaboration. Data Sources: (Informa D&B, 2025),(Informa D&B, 2024),(Informa D&B, 2023),(Informa D&B, 2022),(Informa D&B, 2021),(Solunion, 2025). Note: Years marked with "f" indicate forecasted data. 4

3.1 Data Analysis Methodology. Own Elaboration. 17
3.2 Structure of an MLP. Own elaboration. 21

4.1 Distribution of bankrupt vs non-bankrupt firms. Source: Own Elaboration. . 34
4.2 Top 10 variables with missing values. Source: Own Elaboration. 35
4.3 Boxplots of selected ratios before transformation. Source: Own Elaboration. 36
4.4 Histograms of selected financial ratios. Source: Own Elaboration. 37
4.5 Boxplots of key ratios by bankruptcy status. Source: Own Elaboration. . . 38
4.6 Correlation matrix of financial ratios. Source: Own Elaboration. 39
4.7 ROC Curve for the Final Neural Network Model on the Test Set. Source: Own elaboration. 41
4.8 Training history of neural network. Source: Own Elaboration. 43
4.9 Diagonal plot NN logits versus Poly logits. Source: Own Elaboration. . . . 46
4.10 Taylor approximation for the hidden layer. Source: Own Elaboration. . . . 46
4.11 Taylor approximation for the output layer. Source: Own Elaboration. 47
4.12 Top polynomial coefficient magnitudes and signs. Source: Own Elaboration. 48
4.13 Top 10 most influential polynomial terms. Source: Own Elaboration. . . . 48

List of Tables

2.1 Summary of selected studies on corporate bankruptcy prediction, highlighting data context, interpretability strategies, and application-level insights. . 14

4.1 Comparative characteristics of candidate bankruptcy datasets 32
4.2 Neural network hyperparameter configurations and corresponding AUC scores. Source: Own elaboration. 40
4.3 Confusion matrices for the final model using the standard threshold (0.5) and the optimal Youden threshold. Source: Own elaboration. 42
4.4 Confusion matrix for the evaluated model. Source: Own elaboration. 44
4.5 Confusion matrix NN vs Poly. Source: Own elaboration. 45

5.1 Description of financial variables used in the analysis 53

Acronyms

AI	Artificial Intelligence
ALE	Accumulated Local Effects
AUC	Area Under the ROC Curve
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GP	Genetic Programming
ICE	Individual Conditional Expectation
LIME	Local Interpretable Model-agnostic Explanations
LSTM	Long Short-Term Memory
MDA	Multiple Discriminant Analysis
ML	Machine Learning
MLP	Multilayer Perceptron
PDP	Partial Dependence Plots
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristics
ROSE	Random Over-Sampling Examples
SGD	Stochastic Gradient Descent
SHAP	Shapley Additive Explanations
SME	Small and Medium-sized Enterprises
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
XAI	Explainable Artificial Intelligence

Chapter 1

Introduction

1.1 Motivation

The current global business landscape is characterized by high economic uncertainty, persistent inflation, elevated interest rates, and growing geopolitical instability, all of which exert substantial financial pressure on firms across industries. Under these conditions, the early detection and prediction of corporate bankruptcy has become a central challenge for businesses, investors, and regulatory authorities. Anticipating potential defaults is not only relevant for firms seeking to mitigate financial distress and ensure continuity, but also for policymakers and financial institutions aiming to safeguard systemic stability and economic resilience.

According to the Global Insolvency Report by (Allianz Trade, 2025), global business insolvencies increased by approximately 10% in 2024 and are projected to rise by an additional 6% in 2025 and 3% in 2026, marking the fifth consecutive annual increase. This upward trend represents a normalization after the historically low insolvency rates observed during the pandemic, when extensive government support measures temporarily contained corporate defaults. As illustrated in Figure 1.1, global insolvencies have been increasing steadily since 2020, with Western Europe and North America showing particularly strong rebounds. These trends reflect the lingering consequences of post-pandemic economic normalization and the tightening of financial conditions. Liquidity constraints persist, particularly for highly leveraged firms, due to the delayed reduction of interest rates and limited credit availability. Allianz Trade estimates that a 1% decrease in credit supply can lead to a +3% increase in insolvencies in the United States, +2% in France, and +0.4% in Germany, illustrating the sensitivity of business solvency to financial tightening. (Allianz Trade, 2025) Spain has recorded a more moderate yet persistent increase of around 3%, whereas the Eurozone as a whole has experienced a sharper acceleration of 19%, driven mainly by France (+17%), Germany (+23%), and Italy (+45%). Construction, retail, and business services together account for more than half of total insolvencies, while Western Europe remains among the most affected regions. In 2025, this escalation is expected to threaten approximately 2.3

million jobs worldwide, with nearly 1.1 million located in Western Europe.

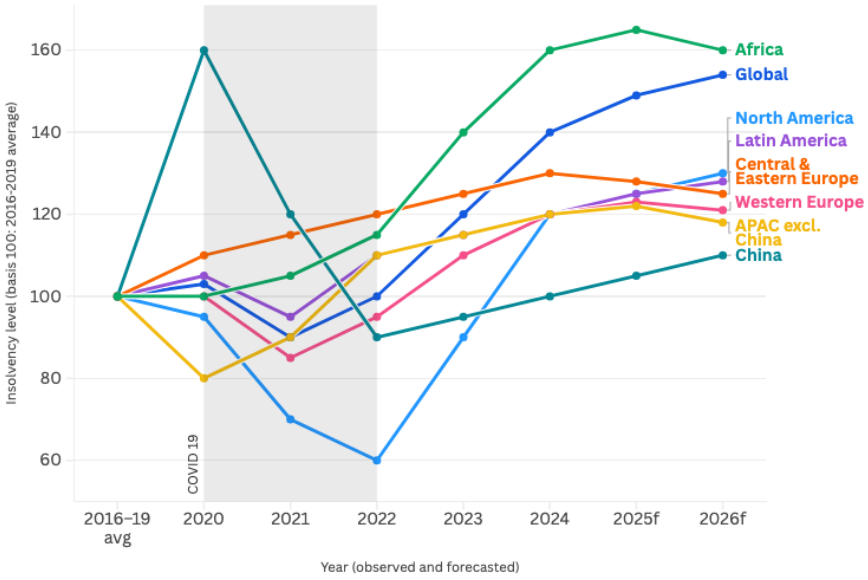

Figure 1.1: Evolution of Global Business Insolvencies (2016–2026f). Own Elaboration. Data Source: (Allianz Trade, 2025). Note: Years marked with "f" indicate forecasted data.

Figure 1.2 compares the annual change in insolvency rates across major European economies between 2023 and projections up to 2026. The data reveal significant cross-country heterogeneity in the evolution of corporate distress. France and Italy recorded sharp increases in insolvency rates in 2023 and 2024 reflecting the early withdrawal of fiscal support and the rapid pass-through of tighter financing conditions. Both countries exhibit signs of gradual stabilization in subsequent years, suggesting a partial adjustment of firms to the post-pandemic financial environment. Germany, in contrast, maintained a consistent upward trajectory throughout the period, which may indicate deeper structural fragilities linked to production costs, energy dependency, and weaker external demand.

Spain presents a markedly different dynamic. Following a steep decline in insolvencies in 2023, likely associated with the resilience of domestic demand and the extension of certain fiscal buffers, the country experienced only a moderate rebound that appears to stabilize by 2026. This evolution suggests that Spanish firms have managed to contain financial distress more effectively than their European counterparts, though the persistence of slight year-on-year increases points to remaining vulnerabilities, particularly among small and medium-sized enterprises. However, even this more contained rebound is a source of concern. The *Concursalidad 2024* report published by Informa D&B and the Colegio de Registradores recorded 9,015 insolvency proceedings (*concursos de acreedores*) in 2024, the highest figure of the last decade, representing a 22% year-on-year increase. (Informa D&B, 2025) Despite this sharp rise in absolute terms, the insolvency rate relative to the number of active companies remains comparatively low at 0.27%, below that of major European economies

such as France and Germany, where it exceeds 1%. There are also notable regional disparities within Spain. Catalonia (26%), Madrid (18%), and Valencia (13%) together accounted for the majority of insolvency filings, while Murcia (0.38%) and Asturias (0.35%) registered the highest relative rates. The most affected sectors were industry (0.49%), communications (0.44%), and hospitality (0.35%). (Informa D&B, 2025)

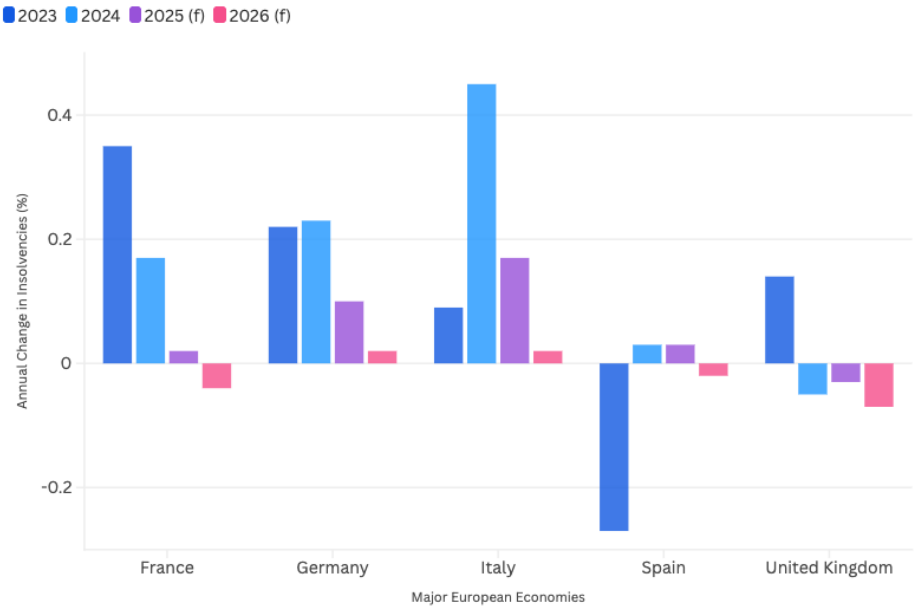

Figure 1.2: Annual Change in Insolvencies by Major European Economies (2023–2025f). Own Elaboration. Data Source: (Allianz Trade, 2025). Note: Years marked with "f" indicate forecasted data.

Figure 1.3 further illustrates the evolution of insolvency proceedings in Spain since 2019, showing a sustained and pronounced upward trend. The number of recorded cases rose from 4,376 in 2020 to 9,015 in 2024 and is expected to reach 9,480 by 2025, marking the highest level in the past decade. This steady escalation underscores the persistence of financial fragility among Spanish firms despite the broader macroeconomic recovery. The structure of the Spanish business landscape provides additional insight into these developments. Small and medium-sized enterprises dominate the national economy and account for the vast majority of insolvency filings. In 2024, micro-enterprises represented 88% of all cases. However, their relative insolvency rate (0.25%) remained below that of small (0.68%) and medium-sized firms (0.61%). These figures indicate that while smaller firms are more numerous, medium-sized companies display higher proportional vulnerability, likely due to their greater exposure to credit constraints and operating leverage. The *Concursalidad 2024* report also highlights that the average insolvent company was between six and ten years old, revealing the fragility of mid-aged businesses with limited financial resilience. (Informa D&B, 2025)

The most recent evidence confirms that this upward trend has persisted into 2025. According to Solunion's Quarterly Insolvency Report (April 2025), Spanish insolvencies in-

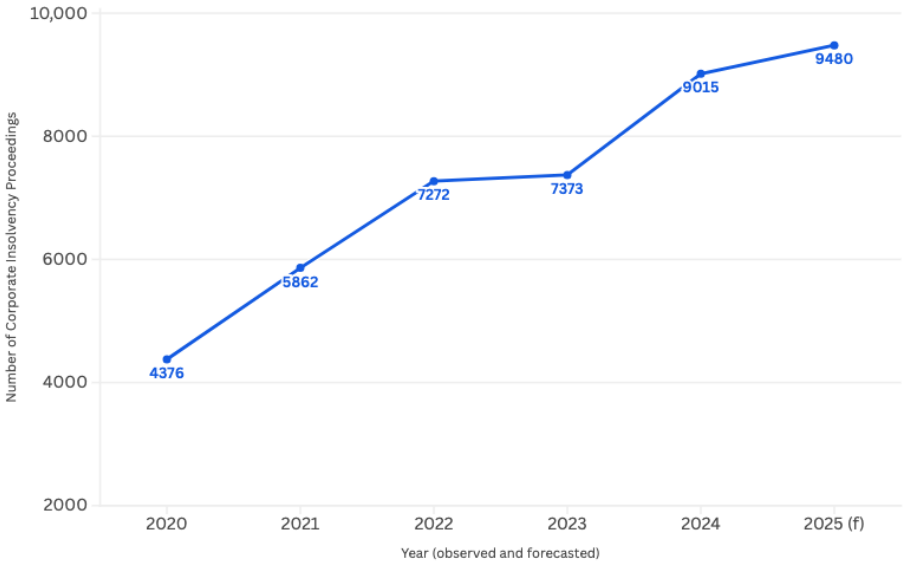

Figure 1.3: Corporate Insolvencies in Spain (2019–2025f). Own Elaboration. Data Sources: (Informa D&B, 2025),(Informa D&B, 2024),(Informa D&B, 2023),(Informa D&B, 2022),(Informa D&B, 2021),(Solunion, 2025). Note: Years marked with "f" indicate forecasted data.

creased by +5.2% in the first quarter of the year. Catalonia once again led the ranking (+20%), followed by Madrid and Valencia, while the services and construction sectors together accounted for more than half of all insolvencies. Particularly sharp increases were observed in retail (+25%), textiles (+43%), and transport (+32%), suggesting mounting difficulties in sectors sensitive to consumption and supply chain pressures. Furthermore, the simplified micro-enterprise insolvency procedure introduced by the 2023 legal reform now represents more than 16% of total cases, underscoring the growing vulnerability of small firms operating in an increasingly restrictive financial environment. (Solunion, 2025)

This increase in corporate distress has implications that extend well beyond individual firms. Bankruptcies disrupt production networks, erode investor confidence, and threaten employment, generating systemic effects that propagate across sectors and regions (Allianz Trade, 2025). For financial institutions and regulatory bodies, anticipating corporate defaults is essential to ensure financial stability, manage credit exposure, and design early intervention mechanisms (Allianz Trade, 2025; Solunion, 2025). Within this context, Business Analytics constitutes a key discipline that transforms vast amounts of financial, macroeconomic, and operational data into actionable insights that enhance decision-making. Predictive analytics enables the identification of patterns and anomalies that precede insolvency, supporting risk management strategies, and improving the allocation of financial resources. For companies, such data-driven approaches strengthen resilience by enabling proactive monitoring of liquidity and solvency indicators, thereby facilitating timely responses to emerging financial pressures.

Given the far-reaching economic and social consequences of corporate bankruptcy, the development of accurate and reliable predictive models has become a relevant challenge in both academic research and financial practice. Early detection of firms at risk of insolvency allows stakeholders to mitigate losses, preserve employment, and maintain market confidence (Mihalovic, 2016). Traditional statistical approaches, such as discriminant analysis or logistic regression, have provided valuable insights but often struggle to capture the complex, nonlinear relationships that characterize financial distress (Mihalovic, 2016). The growing availability of high-dimensional data and advances in computational methods have therefore spurred the adoption of machine learning (ML) techniques, which offer superior predictive accuracy(Qu, Quan, Lei, & Shi, 2019; Shetty, Musa, & Brédart, 2022). Nevertheless, the opacity of many of these models has raised concerns regarding interpretability and accountability, particularly in highly regulated financial contexts. This tension between predictive performance and transparency underscores the need for novel approaches that combine analytical precision with interpretability, a gap that this study seeks to address.

1.2 Objectives

1.2.1 General Objective

This Bachelor's Thesis in Business Analytics aims to examine the predictive capacity of neural networks in the early detection of corporate bankruptcy and incorporate the NN2Poly technique as an explanation tool to interpret relationships derived from financial indicators.

1.2.2 Specific Objectives

The General Objective can be broken down into Specific Objectives that provide a more concrete vision of the study that is going to be carried out, these are presented below:

- Contextualize the relevance of corporate bankruptcy prediction within the field of Business Analytics by reviewing the existing literature and the main approaches used.
- Develop a predictive model based on neural networks capable of accurately estimating the probability of corporate bankruptcy from financial data.
- Incorporate explainability into the predictive model by applying advanced techniques such as NN2Poly to generate interpretable representations of the relationships learned by the neural network.

1.3 Structure of the Document

This Bachelor's Thesis is organized into five chapters. Chapter 1 outlines the motivation, objectives, and general structure of the work. Chapter 2 presents an overview of the state of the art in corporate bankruptcy prediction, reviewing traditional statistical models and recent approaches based on ML techniques; as well as examining the main interpretability methods applied to these models. The chapter aims to clarify the current research landscape, summarize key methodological advances, and demonstrate the relevance of integrating explainability into predictive modeling. Chapter 3 describes the methodology followed in this investigation, including the preprocessing and exploratory analysis of the selected dataset, the design and training of the neural network model, and the incorporation of NN2Poly to improve the interpretability of the model. Chapter 4 outlines the main results obtained, focusing on the model's performance and the insights derived from the use of explainability techniques. Finally, Chapter 5 presents the conclusions, highlighting the most influential variables in corporate bankruptcy prediction and discussing the implications of the findings.

Chapter 2

Review of Predictive Models for Corporate Bankruptcy and Interpretability Approaches

2.1 Machine Learning Techniques for Bankruptcy Prediction

Corporate bankruptcy prediction has constituted a central research stream in financial economics for more than five decades, with implications that extend well beyond academic inquiry. Reliable early warning models support creditors in assessing default risk, assist investors in portfolio allocation, inform supervisory authorities concerned with financial stability, and provide firms with timely signals that can guide preventive actions and restructuring decisions. Episodes of widespread stress, including the 2008 global financial crisis and subsequent economic disruptions, have further emphasized the relevance of this task, as clustered corporate failures can propagate through supply chains and credit markets, amplifying systemic risk and generating substantial macroeconomic costs (Qu et al., 2019). Within this setting, predictive accuracy alone is insufficient. For models to be adopted in risk management and supervisory settings, their outputs must be accompanied by credible explanations that enable stakeholders to understand the drivers of predicted distress, justify decisions, and translate model insights into actionable interventions (Došilović, Brčić, & Hlupić, 2018; Moen, 2020).

The methodological trajectory of the field reflects a shift from parsimonious statistical models toward increasingly complex machine learning (ML) and deep learning approaches. While modern algorithms have improved predictive performance by capturing nonlinearities and high-dimensional interactions, they have also reduced transparency, intensifying concerns about accountability and motivating the development of interpretability and ex-

plainable artificial intelligence techniques (XAI) (Došilović et al., 2018; Moen, 2020). This evolution is particularly evident in the earliest foundations of bankruptcy prediction, which emphasized transparent statistical rules linking financial ratios to corporate solvency and laid the groundwork for subsequent advances.

The foundations of bankruptcy prediction relied on transparent statistical methods that explicitly linked accounting-based financial ratios to corporate solvency. In this line, (Altman, 1968) introduced the Multiple Discriminant Analysis (MDA) model, commonly referred to as the Z-Score, combining five ratios into a linear discriminant function to classify firms as solvent or insolvent. Its appeal stemmed from its parsimony and straightforward interpretability, which made it a cornerstone of early credit-risk assessment. However, the predictive validity of MDA was constrained by strong distributional and structural assumptions, including linear separability and approximate normality of predictors, which are rarely satisfied in real-world financial data. Building on these early contributions, Ohlson (1980) advanced the field by adopting logistic regression to model the probability of bankruptcy directly, thereby relaxing several limitations of discriminant analysis and providing a probabilistic output that supported risk ranking. As a result, logistic regression became a widely used baseline due to its coefficient-based transparency and the ease with which analysts can interpret marginal effects. Nevertheless, both MDA and logistic regression remain limited in their capacity to capture nonlinearities, interaction effects, and complex dependence structures among financial indicators. These limitations motivated a progressive shift toward more flexible, data-driven modeling paradigms, including machine learning approaches designed to improve predictive performance under fewer restrictive assumptions (Mihalovic, 2016; Qu et al., 2019).

These limitations motivated a progressive shift toward more flexible, data-driven modeling paradigms, including ML approaches designed to improve predictive performance under fewer restrictive assumptions (Mihalovic, 2016; Qu et al., 2019). In particular, the rise of ML from the 1990s onward introduced a family of methods capable of capturing nonlinear patterns, threshold effects, and interaction structures that classical statistical models could not represent adequately. Tree-based models became especially influential in this transition. Decision trees partition the feature space into hierarchical decision rules, often expressed as ratio-based thresholds, which can yield intuitive classifications but may suffer from instability and sensitivity to sampling variation. To address these limitations, ensemble methods such as Random Forests and Gradient Boosting combine multiple trees to produce more robust predictors. Random Forests reduce variance by aggregating predictions from many decorrelated trees trained on bootstrap samples, whereas Gradient Boosting constructs an additive model by sequentially refining weak learners to minimize predictive error (Qu et al., 2019).

Evidence from the literature suggests that these ensemble approaches frequently surpass traditional baselines in predictive accuracy. After synthesizing decision-tree-based research,

(Alaka et al., 2018) reported that Random Forest and Gradient Boosting generally outperform classical statistical models in bankruptcy prediction tasks. Nevertheless, these gains typically come at the expense of interpretability. Because ensemble predictions reflect the aggregation of many weak learners, the contribution of individual variables and the logic underlying specific classifications become difficult to trace. This methodological shift therefore accentuated a central tension in the field, namely the transition from transparent but constrained models to highly accurate yet increasingly opaque predictive systems.

As computational capabilities expanded and larger-scale datasets became more accessible, boosting frameworks and neural networks increasingly emerged as dominant approaches in bankruptcy prediction research. In particular, gradient-boosting methods have gained prominence due to their strong performance on tabular financial data, while neural architectures have enabled the modeling of more complex functional forms and, in some settings, temporal dependencies. For instance, (Moen, 2020) proposed a time-aware modeling strategy to capture longitudinal patterns in financial statements by studying Norwegian small and medium-sized enterprises (SMEs) and comparing logistic regression, CatBoost, and recurrent neural networks. Their findings indicate that gradient-boosting models and neural architectures can substantially improve predictive performance relative to classical baselines, especially under the severe class imbalance that characterizes bankruptcy datasets (Moen, 2020; Qu et al., 2019). At the same time, these gains are typically accompanied by increased model complexity, which complicates the interpretation of the mechanisms driving predictions and raises concerns regarding transparency and accountability in practical risk-management and supervisory applications.

Deep learning techniques have further advanced bankruptcy prediction by learning complex nonlinear mappings and, crucially, by exploiting temporal structure in firm-level accounting information. Whereas many classical and machine learning approaches treat financial ratios as static snapshots, recurrent architectures can incorporate the sequential evolution of corporate fundamentals and thereby capture early deterioration patterns that precede insolvency. In this context, (Kaspersen & Lindemark, 2022) trained Long Short-Term Memory (LSTM) networks on longitudinal accounting records from Norwegian firms and reported superior performance relative to conventional machine learning models and shallower neural networks, achieving an AUC of 0.93. These results support the view that modeling temporal dependencies constitutes a substantive methodological improvement in bankruptcy prediction, particularly in settings where risk accumulates over multiple reporting periods (Kaspersen & Lindemark, 2022; Qu et al., 2019). Nevertheless, the enhanced predictive accuracy of deep learning models is often accompanied by limited transparency. Their internal representations are difficult to inspect and their decision logic is not readily expressible in human-interpretable terms, which constrains adoption in high-stakes financial contexts. This tension between predictive power and limited transparency motivates the incorporation of interpretability frameworks capable of translating complex model behavior into explanations

that are meaningful for financial decision-making.

2.2 Interpretability Techniques for Bankruptcy Prediction

The growing reliance on complex ML and deep learning models in bankruptcy prediction has intensified concerns regarding model transparency and interpretability. While these approaches have demonstrated substantial gains in predictive accuracy, their limited explainability poses a major challenge in financial applications where decisions must be justified, audited, and aligned with regulatory requirements. This accuracy–interpretability trade-off has stimulated a parallel stream of research on interpretability and XAI, which aims to clarify the mechanisms through which predictive models infer insolvency risk from financial information (Došilović et al., 2018). In this literature, explanation methods are leveraged to make model behavior intelligible to domain experts, strengthening trust, accountability, and practical usability in bankruptcy prediction settings.

Within this emerging research stream, a foundational contribution is the conceptual framework proposed by (Došilović et al., 2018), which provides a structured taxonomy of interpretability in artificial intelligence. The authors define interpretability as the degree to which a human can understand a model's reasoning and distinguish between intrinsically interpretable and post-hoc explanation methods. Intrinsically interpretable models, such as linear regressions or shallow decision trees, are transparent by design, whereas post-hoc techniques aim to explain complex models after training. Moreover, interpretability can be considered at different levels, including local explanations that describe individual predictions and global explanations that characterize overall model behavior (Došilović et al., 2018). This framework has served as a theoretical foundation for subsequent applications of interpretability in bankruptcy prediction.

Notably, concerns regarding the loss of transparency in predictive models predate the formalization of XAI methodologies. Even before the widespread adoption of interpretability frameworks, several studies identified opacity as a key limitation of ML-based bankruptcy prediction models (Alaka et al., 2018; Qu et al., 2019). Although ensemble methods and deep learning architectures consistently achieved superior predictive performance, their internal decision processes remained difficult to interpret. As emphasized by (Alaka et al., 2018) and (Qu et al., 2019), this lack of transparency hindered practical adoption and reinforced the emerging trade-off between predictive accuracy and interpretability, thereby motivating the development of dedicated explanation techniques tailored to financial risk modeling.

Building on these conceptual distinctions, subsequent research in bankruptcy prediction increasingly shifted from general concerns about opacity toward the practical implementation of post-hoc explanation methods, particularly those that deliver local, case-specific rationales. Among the earliest and most widely adopted techniques is LIME (Local Interpretable

Model-agnostic Explanations), which approximates the behavior of a complex classifier in the neighborhood of a single observation and produces an interpretable representation of the factors driving that prediction. In this context, (Park, Son, Hyun, & Hwang, 2021) applied LIME to tree-based models such as LightGBM and XGBoost, producing instance-level explanations that highlighted the financial variables exerting the strongest influence on each classification. Specifically, liquidity indicators such as the cash ratio and cash and short term investments to current assets emerged as dominant drivers, highlighting the role of short term solvency constraints. Growth indicators, including sales and asset growth rates, were also identified as reflecting the vulnerability of firms experiencing stagnation. In addition, profitability measures such as income before taxes, as well as firm-specific characteristics like age and specific industries (construction, wholesale, and retail), were shown to exert a significant influence on individual predictions. This local perspective provided by LIME supports the justification of individual risk assessments and strengthens decision-making in regulatory and credit-risk settings where transparency and accountability are essential.

As the literature matured, a prominent development was the adoption of Shapley Additive Explanations (SHAP), which provides a unified framework for both local and global interpretability grounded in Shapley values from cooperative game theory (Moen, 2020). Beyond case-level explanations, SHAP enables systematic assessments of variable relevance at the model level. For example, (Moen, 2020) applied SHAP to a logistic regression model, a neural network, and a CatBoost classifier, illustrating that the most influential financial variables vary across the models used but consistently concentrate on firm maturity, size, and liabilities measures. In the logistic regression and CatBoost models, firm maturity was the dominant feature, followed by firm size (log of total assets) and liabilities indicators such as public taxes payable relative to total assets and accounts payable ratios. In contrast, the neural network assigned the highest importance to public taxes payable relative to total assets, with age and size following among the top-ranked variables. Moreover, the CatBoost model also highlighted the industry classification (NACE code) as an important indicator, reflecting the non-linear and sector-specific effects captured by tree-based ensembles. Subsequent work further strengthened the methodological robustness of SHAP-based global explanations. (Liu, Li, Ouyang, Liu, & Wu, 2023) integrated TreeSHAP with Shapley regression, linking SHAP attributions to statistical significance testing and thereby providing a more rigorous inferential basis for global interpretation. Their results show that net asset value per share and net profit after deducting nonrecurring gains and losses consistently rank among the most important features across all horizons, while indicators of short-term operating solvency, such as the ratio of operating profits to current liabilities and the ratio of inventory to current liabilities, increase in relevance for longer-term predictions.

In addition, (Kaspersen & Lindemark, 2022) extended SHAP to deep learning settings through DeepSHAP, using it to explain both individual time-dependent predictions and overall model behavior in LSTM architectures. Their work found that dividends relative to net

income constitute the most influential predictor of bankruptcy across model specifications. Other high-impact variables included measures of leverage and liabilities structure, such as total and long-term liabilities relative to assets, the effective tax rate, and indicators of expense burden and liquidity. Moreover, their temporal analysis shows that SHAP contributions decline as the distance from the prediction year increases, with variables from the most recent period exerting the strongest influence on model outputs. More recently, the literature has shown a growing interest in approaches that embed interpretability within the model itself rather than relying exclusively on post-hoc explanation tools. In this direction, Beade, Rodríguez, and Santos (2024) proposed a Single-Model Multiperiod Bankruptcy Prediction framework based on Genetic Programming (GP), in which bankruptcy risk over multiple horizons is modeled through a mathematical expression composed of financial ratios. The interpretability analysis reveals that the GP models consistently rely on economically relevant variables related to leverage and liabilities structure, such as total and current liabilities relative to assets, profitability and operating performance, including operating income–based ratios, and firm size proxies, specifically the log of total assets and the number of employees. These variables appear explicitly within the mathematical expressions of the model, and thus, the model remains comprehensible to humans while achieving competitive predictive accuracy (Beade et al., 2024).

Following the review of the studies discussed, Table 2.1 provides a structured summary of the existing literature on corporate bankruptcy prediction and interpretability approaches. It synthesizes the reviewed studies by highlighting the data context, prediction task, model families employed, interpretability strategies, and the main application-level insights obtained regarding bankruptcy risk. The table illustrates a clear methodological evolution from traditional statistical models, such as Multiple Discriminant Analysis and logistic regression, which provide high transparency but limited predictive power, towards more complex machine learning and deep learning models applied to diverse data contexts, including cross-sectional, panel, and sequential financial data. Ensemble methods, particularly Random Forests, Gradient Boosting, and CatBoost, as well as neural network architectures, including recurrent models such as LSTMs, are shown to achieve strong predictive performance in challenging settings such as imbalanced bankruptcy datasets, but at the expense of reduced transparency. As a result, post-hoc interpretability techniques are used in most empirical studies, with LIME and SHAP emerging as the most widely applied tools to explain model outputs. Across these studies, interpretability analyses consistently identify a common set of financially significant predictors, including firm size or maturity proxies, liquidity measures, profitability ratios, and leverage and liabilities-related indicators (e.g., total and current liabilities relative to assets).

Despite these advances, the literature reveals a persistent gap regarding the form in which interpretability is delivered. Most empirical studies rely on post-hoc attribution methods—primarily LIME and SHAP—to explain opaque predictors such as ensembles and deep neural

networks. Although these tools provide informative local and global attributions, they typically leave the original model unchanged and maintain a separation between the predictor and its explanation. As a result, interpretability remains dependent on auxiliary explanation procedures, which can be sensitive to modeling choices and may not provide an explicit, model-level representation that can be inspected directly.

In this context, the present study adopts a complementary post-hoc strategy based on model transformation rather than attribution. A neural-network model is trained to predict bankruptcy risk, and NN2Poly is subsequently used to approximate the network with an explicit polynomial representation (Morala, Cifuentes, Lillo, & Ucar, 2025). This transformation yields a compact set of analytical expressions that can be examined to identify nonlinear effects and interactions learned by the network, thereby improving transparency at the model-structure level while retaining much of the predictive capacity of neural architectures. This approach contributes to the bankruptcy prediction literature by linking high-performance modeling with an explicit symbolic surrogate that facilitates communication, validation, and decision support in Business Analytics.

Table 2.1: Summary of selected studies on corporate bankruptcy prediction, highlighting data context, interpretability strategies, and application-level insights.

Reference	Data Context	Prediction Task	Model Family	Interpretability Strategy	Key Interpretable Insights for Bankruptcy Prediction
(Altman, 1968)	U.S. manufacturing firms; cross-sectional financial statements (1960s).	Binary classification: bankrupt vs. solvent firms.	Linear statistical models.	Intrinsic (linear discriminant function).	Showed that bankruptcy risk can be explained through a small set of accounting ratios capturing profitability, liquidity, leverage, and operating efficiency. Established that financial distress is driven by structural balance-sheet weaknesses rather than isolated indicators.
(Ohlson, 1980)	U.S. public firms; accounting and market-based data.	Probability estimation of bankruptcy.	Generalized linear models.	Intrinsic (logistic regression).	Demonstrated that firm size, leverage, liquidity, and profitability jointly determine bankruptcy probability, introducing a probabilistic interpretation that allows continuous risk assessment rather than threshold-based classification.
(Alaka et al., 2018)	Multiple countries and firm types; review of empirical studies.	Binary bankruptcy prediction.	Tree-based and ensemble models.	Post-hoc (feature importance in trees).	Found consistent evidence that leverage, liquidity, and profitability ratios dominate bankruptcy prediction across tree-based models, but noted that ensemble aggregation obscures global functional relationships among variables.
(Moen, 2020)	Norwegian SMEs; multi-year firm-level panel data.	Temporal bankruptcy prediction.	Logistic regression, neural networks, boosting.	Post-hoc (SHAP-based local and global analysis).	Revealed that firm age and size act as stabilizing factors, while liabilities-related indicators (e.g. taxes payable and payables ratios) and industry affiliation systematically increase bankruptcy risk, with importance patterns varying across model classes.
(Kaspersen & Lindemark, 2022)	Norwegian firms; sequential annual accounting records.	Sequence-based bankruptcy prediction.	Recurrent neural networks (LSTM).	Post-hoc (DeepSHAP).	Showed that dividends relative to net income, leverage structure, effective tax rate, and recent operating performance dominate predictions, with interpretability analysis highlighting that the most recent accounting periods exert the strongest influence on bankruptcy risk.
(Došilović et al., 2018)	Conceptual analysis across ML applications.	Not application-specific.	Explainable AI frameworks.	Conceptual taxonomy.	Established the distinction between intrinsic vs. post-hoc and global vs. local interpretability, providing a conceptual basis for evaluating transparency in financial risk models.
(Qu et al., 2019)	International literature on bankruptcy prediction.	Binary and multi-horizon bankruptcy prediction.	ML and deep learning models.	Review of interpretability limitations.	Identified lack of global interpretability as a major barrier to deploying high-performing bankruptcy models in regulated financial environments, despite strong predictive accuracy.
(Park et al., 2021)	Korean firms; structured financial datasets.	Firm-level bankruptcy classification.	Tree-based ensembles.	Post-hoc (LIME, local explanations).	Local explanations highlighted short-term liquidity constraints, profitability measures, firm age, and sector effects as key drivers of individual bankruptcy decisions, supporting case-level justification but not full model transparency.

Reference	Data Context	Prediction Task	Model Family	Interpretability Strategy	Key Interpretable Insights for Bankruptcy Prediction
(Liu et al., 2023)	Chinese listed firms; annual accounting statements.	Multi-horizon bankruptcy prediction.	Gradient boosting models.	Post-hoc (TreeSHAP + Shapley regression).	Identified net asset value per share, net profit after nonrecurring items, and operating solvency ratios as statistically significant and stable predictors of financial distress across different forecast horizons.
(Beade et al., 2024)	Spanish SMEs; multi-period financial records.	Single-model multiperiod bankruptcy prediction.	Symbolic learning models.	Intrinsic (Genetic Programming).	Showed that leverage, profitability, and firm size variables can be explicitly combined within a single symbolic expression, enabling direct inspection of global decision logic without reliance on post-hoc explanations.

Chapter 3

Methodology for Predictive Modeling and Interpretability in Bankruptcy Prediction Environments

This chapter describes the methodology employed to develop a corporate bankruptcy prediction model and to enhance its interpretability through NN2Poly, with the aim of identifying the financial variables that exert the strongest influence on insolvency risk. Following methodological frameworks adopted in related studies, such as (Morala, Cifuentes, Lillo, & Ucar, 2021) and (Morala, Cifuentes, Lillo, & Ucar, 2023), the procedure is organized into three main phases, summarized in Figure 3.1. The first phase covers data collection and preprocessing, including dataset selection, data cleaning, feature preparation, and the application of class-imbalance handling techniques to address the skewed distribution typically observed in bankruptcy datasets. The second phase focuses on training an artificial neural network for bankruptcy prediction, including model specification, hyperparameter selection, and performance evaluation using standard classification metrics. The third phase applies NN2Poly to distill the trained neural model into an explicit polynomial representation, enabling transparent inspection of the learned relationships and supporting model-level interpretation. Together, these stages provide a unified framework that balances predictive performance with interpretability, facilitating the generation of actionable insights for financial decision-making.

3.1 Data Collection and Preprocessing

The first stage of the methodology focuses on dataset selection and preprocessing. This step begins with the identification and evaluation of publicly available datasets suitable for corporate bankruptcy prediction. To ensure methodological validity, the selected dataset must provide a sufficiently large number of firm-level observations labeled as bankrupt or non-

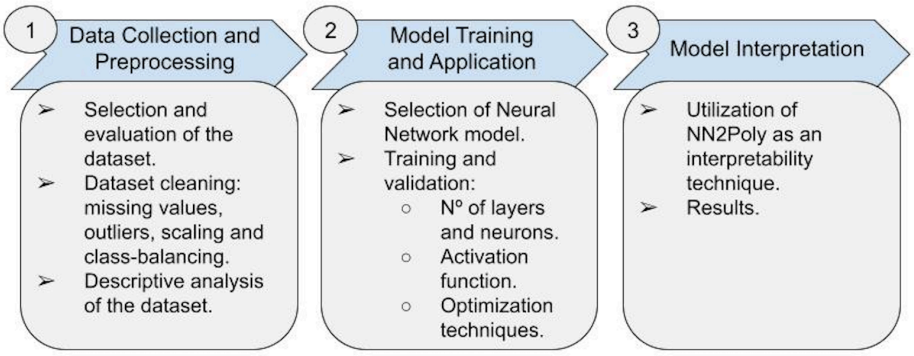

Figure 3.1: Data Analysis Methodology. Own Elaboration.

bankrupt, as well as a consistent set of accounting-based financial indicators measured prior to the bankruptcy event. In addition, the dataset should include variables that capture multiple dimensions of corporate financial health, enabling the model to learn patterns associated with financial distress rather than relying on a narrow set of signals. From a methodological perspective, the feature space is designed to reflect commonly adopted categories in the bankruptcy literature, including liquidity, profitability, leverage, activity or efficiency, and firm size proxies. Rather than fixing a definitive set of ratios at this stage, these categories serve as a guiding structure for feature selection and engineering, ensuring that the final variable set is grounded in financial theory and aligned with empirical practice in distress prediction. The definitive set of predictors is subsequently determined after data quality checks and exploratory analysis, taking into account availability, missingness, redundancy, and multicollinearity patterns.

Once the dataset is selected, a comprehensive preprocessing pipeline is applied to improve data quality and to ensure that the resulting inputs are appropriate for neural-network training. Missing values are handled through imputation strategies selected according to the missingness structure and variable type, while variables exhibiting excessive missingness are evaluated for removal when reliability cannot be ensured. Outliers are identified through statistical criteria and exploratory visual diagnostics (e.g., boxplots and scatter plots) and are managed through transformations, winsorization, or exclusion, depending on their nature and potential influence on model training. Finally, all numerical variables are scaled to a common range, standardizing them into the interval $[-1, 1]$. This normalization step is important in neural networks because gradient-based optimization is sensitive to differences in feature magnitudes. Unscaled variables can dominate the loss landscape, slow convergence, and bias weight updates. In addition, according to (Morala et al., 2023), this scaling is essential for the theoretical validity of NN2Poly, as the Taylor-based approximation relies on activation inputs remaining in a region where the Taylor expansion is sufficiently accurate. Lastly, categorical variables will be appropriately encoded to ensure compatibility with the modeling framework.

In addition to these standard preprocessing steps, a class-balancing procedure is incorporated prior to neural network training. Bankruptcy prediction datasets are typically characterized by severe class imbalance, where the bankrupt class constitutes only a small fraction of the observations. Let the binary target be $y \in \{0, 1\}$, where $y = 1$ denotes bankruptcy and $y = 0$ denotes non-bankruptcy. If the empirical class prior satisfies $\pi_1 = \mathbb{P}(y = 1) \ll \pi_0 = \mathbb{P}(y = 0)$, a classifier trained by minimizing an empirical risk such as the cross-entropy loss can become biased toward the majority class. In extreme cases, a trivial classifier that predicts $\hat{y} = 0$ for every observation attains an apparent accuracy of approximately π_0, yet exhibits very low recall for the bankrupt class, which is the class of greatest practical relevance. This motivates the use of resampling strategies that modify the class composition in the training data to improve the learner's sensitivity to minority-class patterns.

To avoid information leakage and preserve an unbiased evaluation, the balancing step is applied exclusively to the training split, while the test set is retained in its original distribution. This ensures that performance metrics reflect the real-world prevalence of bankruptcy and remain comparable across studies. In this thesis, data balancing is performed using the ROSE (Random Over-Sampling Examples) framework, which generates a more balanced training sample by combining undersampling of the majority class with the creation of synthetic minority observations through a smoothed bootstrap procedure. Formally, ROSE draws artificial samples from a conditional kernel density estimate defined around the observed points of each class, generating synthetic observations x^* in the neighborhood of real observations. A common representation of this mechanism is

$$x^* = x_i + \epsilon, \qquad \epsilon \sim K_h(\cdot), \tag{3.1}$$

where x_i denotes an observed training instance (typically from the minority class), $K_h(\cdot)$ is a kernel distribution with bandwidth h, and ϵ introduces controlled perturbations that increase diversity while preserving local structure. By generating synthetic minority samples and simultaneously reducing the dominance of the majority class, ROSE increases the effective representation of bankruptcy cases during training, encouraging the neural network to learn decision boundaries that better discriminate financially distressed firms.

Following the preprocessing pipeline, an exploratory descriptive analysis is conducted to characterize the empirical properties of the dataset and to provide a statistically grounded overview of the candidate predictors. Summary statistics are computed for each financial indicator, including measures of central tendency (e.g., mean and median) and dispersion (e.g., standard deviation and interquartile range), with the objective of assessing distributional shape, variability, and potential asymmetries. In parallel, correlation analysis is performed to examine linear associations among variables and to identify potential multicollinearity patterns that may affect model estimation and subsequent interpretation. A comparative analysis between bankrupt and non-bankrupt firms is then carried out to detect systematic differences

in financial profiles across classes. Group-wise summaries and distributional comparisons offer preliminary evidence on which dimensions of financial health tend to diverge prior to bankruptcy, thereby motivating hypotheses regarding the most informative predictors. Visual diagnostics, including distribution plots and correlation heatmaps, complement the numerical analysis by providing a clearer representation of data structure, outlier behavior, and relationships across variables. Overall, this exploratory stage contextualizes the financial characteristics present in the sample and informs the interpretability phase by establishing an empirical baseline against which the neural network's learned patterns can be assessed. Building on the resulting preprocessed dataset and exploratory findings, the subsequent section details the neural network specification, training procedure, and performance evaluation strategy.

3.2 Model Training and Application

3.2.1 Neural Network Model Architecture

The second stage of the methodology focuses on the development of a predictive model based on artificial neural networks. In line with prior evidence highlighting the suitability of neural models for representing complex functional relationships in business forecasting, this study adopts a Multilayer Perceptron (MLP) as the primary predictive architecture (Zhang, 2004). The choice of an MLP is methodologically aligned with the nature of the available data, which consist of cross-sectional firm-level observations described by structured tabular variables and a binary target indicating bankruptcy status. In this setting, the objective is to learn a nonlinear mapping from a vector of financial and operational characteristics to an estimated probability of insolvency, without imposing restrictive assumptions such as linear separability or additive effects.

An MLP is a feedforward artificial neural network composed of an input layer, one or more hidden layers, and an output layer. Each layer contains computational units (neurons) that transform the incoming information through a weighted aggregation followed by a nonlinear activation function. By stacking these transformations across layers, the network can approximate complex decision boundaries and interaction effects that are difficult to represent using traditional linear models (Barboza, Kimura, & Altman, 2017; Dubey, Singh, & Chaudhuri, 2022). This modeling flexibility is particularly relevant in bankruptcy prediction, where financial distress typically emerges from joint patterns across liquidity, leverage, profitability, and operational indicators rather than from any single ratio in isolation.

Figure 3.2 illustrates the structure of the MLP adopted in this study. The input layer contains the predictors for each firm, denoted by $x_1, \ldots, x_p$, where p is the number of preprocessed features. These inputs are propagated through one or more hidden layers and finally mapped to a single output neuron that produces the estimated bankruptcy risk. Connections

between neurons are weighted, as indicated by the labels $w_{i,j}$ in the figure. Here, $w_{i,j}$ denotes the weight associated with the connection from neuron i in a given layer to neuron j in the subsequent layer. Each hidden-layer neuron also includes a bias term (e.g., b_1, b_2, b_3), which shifts the activation threshold and increases the flexibility of the representation. For a neuron j in a hidden layer, the pre-activation value is computed as

$$z_j = \sum_{i=1}^{m} w_{i,j}\, a_i + b_j, \tag{3.2}$$

where a_i denotes the output of neuron i in the previous layer (with $a_i = x_i$ at the input layer), m is the number of neurons in the previous layer, and b_j is the bias of neuron j. The neuron output is then obtained by applying a nonlinear activation function $g(\cdot)$:

$$a_j = g(z_j). \tag{3.3}$$

This fully connected structure implies that each hidden neuron aggregates information from all inputs (directly or indirectly through previous hidden layers), enabling the MLP to capture interaction effects among financial indicators. Through successive nonlinear transformations, the network constructs internal representations that progressively separate financially stable firms from those exhibiting a higher probability of distress. The output layer combines the final hidden representation into a single score z_{out}, which is mapped to the interval $(0, 1)$ using a sigmoid activation function to yield a probability-like estimate of bankruptcy:

$$\hat{p}(y = 1 \mid x) = \sigma(z_{\text{out}}), \qquad \sigma(t) = \frac{1}{1 + e^{-t}}. \tag{3.4}$$

In Equation (3.4), $\hat{p}(y = 1 \mid x)$ denotes the predicted probability that a firm is bankrupt ($y = 1$), conditional on its feature vector x. The sigmoid mapping is particularly suitable in binary classification settings because it constrains outputs to the probability range and supports threshold-based decision rules.

A central component enabling the MLP to move beyond linear decision rules is the activation function applied at each neuron. Activation functions introduce nonlinearity, allowing the network to represent threshold effects, saturation behaviors, and interaction patterns that cannot be captured by purely linear transformations (Dubey et al., 2022). Different activation functions exhibit distinct mathematical properties, and their choice affects both the expressive capacity of the model and the stability of gradient-based optimization. In this context, the sigmoid function, defined in Equation (3.4), is commonly used in the output layer of binary classification models because it produces values in $(0, 1)$ that can be interpreted probabilistically. However, sigmoid activations may lead to vanishing-gradient behavior when the input magnitude $|t|$ is large, since their derivatives become very small in saturation regions. This phenomenon can slow learning, especially as network depth increases, and motivates the use

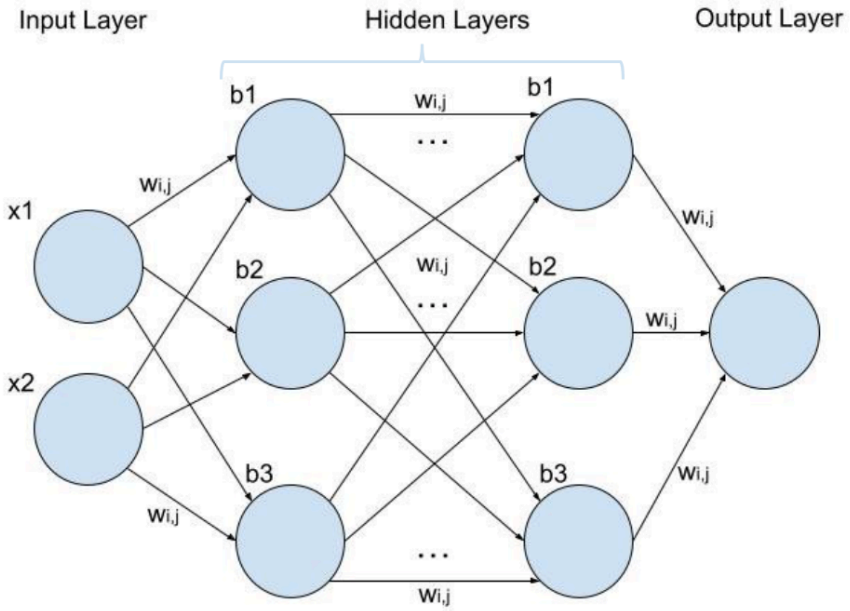

Figure 3.2: Structure of an MLP. Own elaboration.

of alternative activation functions in hidden layers to maintain efficient optimization (Dubey et al., 2022).

Another commonly used activation function is the hyperbolic tangent (*tanh*), defined as

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}. \tag{3.5}$$

The mapping in Equation (3.5) transforms inputs to the bounded interval $[-1, 1]$ and is zero-centered, which can be advantageous for optimization because positive and negative activations propagate more symmetrically through the network. In practice, this centering may improve numerical stability and convergence relative to sigmoid activations, particularly in hidden layers where balanced gradients are desirable. Nevertheless, as with other saturating functions, $\tanh(\cdot)$ can still exhibit vanishing gradients for large $|x|$ due to its flat tails.

Beyond these general optimization considerations, the choice of $\tanh(\cdot)$ is also directly motivated by the interpretability objective of this thesis. NN2Poly relies on approximating the nonlinear activations of a trained neural network through polynomial expansions derived from Taylor series. For such approximations to remain accurate, the activation function must be sufficiently smooth and differentiable, and the inputs to the activation (i.e., synaptic potentials) must remain within a range where the Taylor expansion provides a good approximation. As emphasized by Morala et al. (2023), smooth activations such as $\tanh(\cdot)$ are particularly compatible with NN2Poly when features are scaled to $[-1, 1]$, since this scaling helps constrain the magnitude of synaptic potentials and supports accurate polynomial representations over a wider domain.

A further activation considered in this study is the *softplus* function, defined as

$$\text{softplus}(x) = \log\left(1 + e^{x}\right). \tag{3.6}$$

Softplus can be interpreted as a smooth approximation to the Rectified Linear Unit (ReLU), behaving approximately linearly for large positive inputs (i.e., $\text{softplus}(x) \approx x$ when $x \gg 0$) and approaching zero for large negative inputs (i.e., $\text{softplus}(x) \approx 0$ when $x \ll 0$). Its derivative is the sigmoid function, $\frac{d}{dx}\text{softplus}(x) = \sigma(x)$, which implies that softplus is continuously differentiable and avoids the non-differentiability at zero that characterizes ReLU. This smoothness makes softplus suitable in contexts where differentiability is important for subsequent analytical transformations. In the context of NN2Poly, softplus is also attractive because it is analytic and smooth, facilitating Taylor-based polynomial approximations of the activation response. Compared to piecewise-linear activations, the absence of kinks supports a more stable polynomial distillation step, since higher-order derivatives exist everywhere. However, similar to sigmoid-based derivatives, softplus may yield small gradients for strongly negative inputs due to saturation. For this reason, constraining input magnitudes through feature scaling and careful regularization remains relevant to maintain stable learning dynamics and, simultaneously, to support accurate polynomial approximations in the NN2Poly stage (Morala et al., 2023).

Taken together, these considerations motivate the evaluation of smooth activation functions that are compatible with both neural-network optimization and the subsequent NN2Poly transformation. Accordingly, this study considers $\tanh(\cdot)$, $\text{softplus}(\cdot)$, and $\sigma(\cdot)$ as candidate activation functions for the hidden layers. The final selection is guided by predictive performance and training stability. For the output layer, a sigmoid activation function will be employed to produce bankruptcy probabilities in the range $[0, 1]$, enabling direct interpretation as the probability of financial distress (Dubey et al., 2022).

3.2.2 Training Process and Optimization

The estimation of the MLP parameters is formulated as a supervised optimization problem in which the objective is to minimize a loss function that measures the discrepancy between the predicted bankruptcy probabilities and the observed class labels. Let $y_i \in \{0, 1\}$ denote the bankruptcy status of firm i, where $y_i = 1$ indicates insolvency and $y_i = 0$ otherwise, and let $\hat{p}_i = \hat{p}(y = 1 \mid x_i)$ be the probability predicted by the network for observation i. For binary classification problems, the binary cross-entropy (logistic) loss is adopted due to its probabilistic interpretation and strong penalization of confident misclassifications. For a sample of n firms, the loss is defined as

$$L = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)\right]. \tag{3.7}$$

This objective function corresponds to the negative log-likelihood of a Bernoulli model and is consistent with the sigmoid output layer, thereby ensuring that training directly optimizes probabilistic predictions. Parameter estimation is performed using backpropagation, an algorithm that efficiently computes the gradient of the loss function with respect to all network parameters by applying the chain rule of calculus (Maclaurin, Duvenaud, & Adams, 2015). Starting from the output layer, where the prediction error is first evaluated, the error signal is propagated backward through the hidden layers. At each layer, partial derivatives of the loss with respect to weights and biases are computed, allowing each connection in the network to be updated according to its contribution to the final prediction error. This mechanism ensures that learning is distributed across all layers, enabling the network to refine both low-level and high-level representations of financial distress patterns.

The optimization step itself is carried out through gradient-based methods. In its basic form, stochastic gradient descent (SGD) updates each weight parameter w iteratively according to

$$w^{(t+1)} = w^{(t)} - \eta \frac{\partial L}{\partial w}, \tag{3.8}$$

where $w^{(t)}$ denotes the value of the parameter at iteration t, $\eta > 0$ is the learning rate controlling the step size, and $\frac{\partial L}{\partial w}$ is the gradient of the loss with respect to that parameter. In practice, mini-batch training is employed, so gradients are estimated using subsets of the training data, which stabilizes convergence and improves computational efficiency. In addition to SGD, adaptive optimization algorithms such as Adam are considered. These methods adjust the effective learning rate for each parameter using first and second moment estimates of past gradients, often leading to faster convergence and improved robustness in high-dimensional parameter spaces.

Training proceeds over multiple epochs, where one epoch corresponds to a complete pass through the training dataset. Across epochs, the iterative updates progressively reduce the loss in Equation (3.7), enabling the network to learn a mapping from firm-level financial features to bankruptcy probabilities. This optimization process ultimately yields a set of weights and biases that encode the nonlinear relationships underlying corporate financial distress, which are subsequently transformed into polynomial form through the NN2Poly procedure for interpretability.

3.2.3 Hyperparameter Tuning

Hyperparameters constitute key design and training choices in neural-network modeling. Unlike weights and biases, which are estimated from data through backpropagation, hyperparameters are specified prior to training and determine the functional capacity of the network, the stability of the optimization process, and the model's generalization performance (Zhang,

2004). In bankruptcy prediction, where datasets are often noisy and imbalanced and the cost of misclassification is asymmetric, an appropriate hyperparameter configuration is required to obtain a model that is both accurate and reliable. A first set of hyperparameters concerns the network architecture, namely the number of hidden layers and the number of neurons per layer. These elements control model depth and representational capacity. Architectures with insufficient capacity may underfit by failing to capture nonlinearities and interaction effects across financial indicators, whereas overly complex architectures may overfit, learning idiosyncratic patterns in the training data that do not transfer to unseen firms (Zhang, 2004). For this reason, alternative architectural configurations are evaluated in order to balance flexibility and generalization.

The activation function employed in the hidden layers is also treated as a hyperparameter. As discussed in the previous subsection, this choice affects both optimization dynamics and the type of nonlinearities that the network can represent. In addition, the selection of smooth activation functions (e.g., $\tanh(\cdot)$ or $\text{softplus}(\cdot)$) is particularly relevant in this thesis because NN2Poly approximates the trained network through Taylor-based polynomial expansions, for which differentiability and smoothness facilitate more accurate distillation (Morala et al., 2023). Accordingly, $\tanh(\cdot)$, $\text{softplus}(\cdot)$, and $\sigma(\cdot)$ are considered as candidate activations for hidden layers, while a sigmoid activation is retained in the output layer to produce probability estimates in $(0, 1)$. Additional training-related hyperparameters include the batch size and the number of training epochs. The batch size determines the number of observations used to compute each gradient update and therefore influences the variance of the gradient estimate and the stability of convergence. Smaller batches introduce higher stochasticity, which can improve generalization but may slow convergence, whereas larger batches yield smoother updates at the potential cost of poorer generalization. The number of epochs controls the total number of passes over the training set and, together with early-stopping criteria, determines the point at which training is terminated to avoid overfitting (Zhang, 2004).

Regularization is incorporated as an additional hyperparameter dimension to control model complexity and, importantly, to support the subsequent NN2Poly transformation. In this thesis, L1 regularization is considered to penalize large synaptic weights and thereby constrain the magnitude of the pre-activation values (synaptic potentials) within the hidden layers. This is particularly relevant because NN2Poly relies on Taylor-based polynomial approximations of the activation functions, which are accurate only within a neighborhood of the expansion point. If weight magnitudes become large, the resulting pre-activations z can move outside the region where a truncated Taylor series provides a reliable approximation, even when inputs are scaled to $[-1, 1]$. By discouraging large weights, L1 regularization helps keep the network operating in a regime that is more compatible with stable polynomial distillation. As a secondary effect, the penalty can also promote sparser parameterizations, which may lead to more compact polynomial expressions after NN2Poly.

To identify an effective configuration, a grid-search strategy is adopted over a selected

set of hyperparameters, including (i) the hidden-layer activation function, (ii) the number of neurons (and, when applicable, the number of hidden layers), and (iii) the batch size. For each candidate configuration, the network is trained on the balanced training set, and its performance is evaluated using validation-based criteria aligned with the bankruptcy prediction objective. Following (Barboza et al., 2017), Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) will be the main evaluation metric due to its robustness in assessing bankruptcy classifiers with imbalanced class distributions. The best-performing configuration is then retained for final training and subsequent NN2Poly transformation.

3.2.4 Model Evaluation

The final stage of this phase consists of training and evaluating the neural network in order to select the most appropriate model configuration. The dataset is partitioned into training and testing subsets, with 80% of the observations allocated to training and 20% reserved for final testing. The training subset is used for hyperparameter tuning through a grid search, following standard practices in predictive modeling to obtain a reliable estimate of generalization performance (Zhang, 2004). Because bankruptcy datasets are typically characterized by pronounced class imbalance, model performance cannot be adequately summarized by accuracy alone. In such settings, a classifier may attain high accuracy by predominantly predicting the majority (non-bankrupt) class while failing to detect bankrupt firms. For this reason, evaluation is conducted using complementary metrics that quantify both overall discrimination and minority-class detection, as recommended in the financial distress literature (Barboza et al., 2017).

Let $y_i \in \{0, 1\}$ denote the observed class label for firm i, where $y_i = 1$ indicates bankruptcy and $y_i = 0$ indicates a non-bankrupt firm. Let $\hat{p}_i = \hat{p}(y = 1 \mid x_i) \in (0, 1)$ be the predicted probability output by the neural network given the feature vector x_i. To convert these continuous risk scores into a binary classification, a decision threshold $\tau \in (0, 1)$ is introduced. The predicted label $\hat{y}_i \in \{0, 1\}$ is then defined as

$$\hat{y}_i = \mathbb{I}(\hat{p}_i \geq \tau), \tag{3.9}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Intuitively, τ determines the operating point of the classifier: lower values of τ produce a more sensitive screening rule (more firms flagged as bankrupt), whereas higher values yield a more conservative rule (fewer bankruptcy flags, but potentially more missed bankruptcies).

Given $\hat{y}_i$, classification outcomes can be summarized through the confusion matrix counts: true positives (TP, bankrupt firms correctly flagged), false positives (FP, non-bankrupt firms incorrectly flagged), true negatives (TN, non-bankrupt firms correctly not flagged), and false negatives (FN, bankrupt firms missed by the model). Standard performance measures follow

directly. Accuracy is defined as the overall fraction of correct classifications:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{3.10}$$

However, in bankruptcy prediction, accuracy can be misleading due to class imbalance, since predicting the majority class may already yield high accuracy without detecting distressed firms. Two threshold-dependent measures that better reflect the screening trade-off are sensitivity and specificity. Sensitivity (also referred to as the true positive rate, TPR, or recall) measures the proportion of bankrupt firms correctly identified:

$$\text{Sensitivity} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{3.11}$$

Specificity (also referred to as the true negative rate, TNR) measures the proportion of non-bankrupt firms correctly classified:

$$\text{Specificity} = \text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{3.12}$$

In bankruptcy applications, sensitivity is often prioritized because false negatives (missed bankruptcies) may entail substantial economic and systemic costs, whereas false positives typically imply additional monitoring or screening effort. Because sensitivity and specificity depend on the chosen threshold τ, it is common to evaluate discrimination in a threshold-independent manner using the Receiver Operating Characteristic (ROC) curve. The ROC curve traces the trade-off between the TPR and the false positive rate (FPR) across all thresholds $\tau \in (0, 1)$. The false positive rate is defined as

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = 1 - \text{Specificity}. \tag{3.13}$$

Thus, each value of τ yields a point $(\text{FPR}(\tau), \text{TPR}(\tau))$ on the ROC space, and varying τ produces the full curve. The Area Under the ROC Curve (AUC) summarizes this curve into a single scalar measure of discriminative ability. Formally, AUC is the integral of TPR as a function of FPR and admits a probabilistic interpretation: it equals the probability that the model assigns a higher risk score $\hat{p}$ to a randomly selected bankrupt firm than to a randomly selected non-bankrupt firm. Higher AUC values therefore indicate stronger ranking performance, independent of any particular operating threshold.

While AUC characterizes global discrimination, operational decisions still require choosing a specific threshold τ. In this study, the operating threshold is selected using Youden's index, which identifies the value of τ that maximizes the sum of sensitivity and specificity:

$$\tau^\star = \arg\max_{\tau \in (0,1)} \left\{ \text{Sensitivity}(\tau) + \text{Specificity}(\tau) - 1 \right\}. \tag{3.14}$$

This criterion selects the point on the ROC curve that is farthest from the diagonal (random classifier), providing a balanced compromise between detecting bankrupt firms and limiting false alarms.

The final selected model is evaluated on the independent test subset using these metrics to obtain an objective estimate of predictive performance on previously unseen firms. Demonstrating robust results at this stage provides evidence of adequate generalization and supports the use of the trained network as the foundation for the subsequent interpretability phase. Once predictive reliability is established, NN2Poly is applied to transform the trained neural network into an explicit polynomial representation, enabling a transparent assessment of the financial variables that most strongly drive bankruptcy risk predictions.

3.3 Model Interpretation via NN2Poly

Following the estimation and validation of the predictive model, the NN2Poly interpretability framework is employed to obtain an analytical representation of the learned decision function. Despite their strong predictive performance, deep neural networks are widely regarded as opaque models due to their multilayered structure and the composition of nonlinear transformations, which obscure the relationship between input variables and final predictions (Zhang, 2004). This limited transparency poses a substantive challenge in financial applications, where model outputs often inform high-stakes decisions related to credit allocation, regulatory oversight, and corporate risk management. In such contexts, predictive accuracy alone is insufficient; stakeholders require interpretable models that allow the economic drivers of bankruptcy risk to be examined and justified.

To address this limitation, the present study adopts NN2Poly, a methodology that converts a trained neural network into an explicit polynomial representation of the learned mapping (Morala et al., 2021). Rather than approximating model behavior through external surrogate explanations, NN2Poly analytically reconstructs the functional form encoded in the network by expressing the sequence of linear transformations and nonlinear activations as a polynomial expansion (Morala et al., 2023). This transformation yields a global representation of the model that is mathematically explicit and directly interpretable, enabling a transparent examination of how financial variables and their interactions contribute to predicted bankruptcy risk.

3.3.1 NN2Poly Transformation

NN2Poly provides an analytical framework to express the function learned by a trained neural network as an explicit polynomial representation. The method builds on the Taylor series expansion of smooth activation functions and leverages the fact that a feedforward neural network is a composition of linear transformations and nonlinear activations. By approximating

the nonlinear components with polynomials, the overall network mapping can be rewritten as a polynomial function of the original input variables (Morala et al., 2021, 2023). Consider a neuron j in a hidden layer, whose pre-activation value (synaptic potential) is given by

$$z_j = \sum_{i=0}^{p} w_{i,j} x_i, \tag{3.15}$$

where $x_0 = 1$ represents the bias input, $x_1, \ldots, x_p$ are the model features, and $w_{i,j}$ are the trained weights. The neuron output is $a_j = g(z_j)$, where $g(\cdot)$ denotes a smooth activation function such as $\tanh(\cdot)$ or $\text{softplus}(\cdot)$, previously introduced for their compatibility with Taylor-based approximations.

NN2Poly approximates each activation function $g(u)$ by a Taylor series expansion centered at zero,

$$g(u) \approx \sum_{n=0}^{q} \frac{g^{(n)}(0)}{n!} u^n, \tag{3.16}$$

where q denotes the truncation order and $g^{(n)}(0)$ is the n-th derivative evaluated at zero. This step replaces the nonlinear activation by a polynomial in the synaptic potential u. Substituting Equation (3.15) into the expansion and applying the multinomial theorem yields

$$z_j^n = \left(\sum_{i=0}^{p} w_{i,j} x_i \right)^n = \sum_{m_0 + \cdots + m_p = n} \binom{n}{m_0, \ldots, m_p} \prod_{i=0}^{p} (w_{i,j} x_i)^{m_i}, \tag{3.17}$$

where $\binom{n}{m_0,\ldots,m_p} = \frac{n!}{m_0! \cdots m_p!}$. Consequently, the output of each hidden neuron can be expressed as a polynomial in the original inputs $x_1, \ldots, x_p$. The procedure is applied iteratively across layers. Once a hidden layer is expressed as polynomial functions of the inputs, these polynomials become the inputs to the next layer, where the same Taylor expansion and substitution steps are performed. Through this recursive composition, the final network output z_{out} is rewritten as a polynomial function of the original feature vector $\mathbf{x}$. The resulting expression has the general structure

$$P(\mathbf{x}) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \sum_{i \le j} \beta_{ij} x_i x_j + \sum_{i,j,k} \beta_{ijk} x_i x_j x_k + \cdots, \tag{3.18}$$

where the coefficients β are functions of the trained network weights and the derivatives of the activation functions. Linear terms correspond to main effects of individual variables, while higher-order terms capture interactions and nonlinear effects learned by the network. Through this construction, NN2Poly converts the trained neural network into an analytically tractable polynomial whose structure mirrors the learned relationships among financial indicators. This representation enables a direct examination of variable contributions and interaction effects within a unified mathematical framework, providing global interpretability of

the predictive model (Morala et al., 2021, 2023).

Beyond the general polynomial structure in Equation (3.18), NN2Poly provides explicit expressions that link the polynomial coefficients directly to the trained network parameters (Morala et al., 2021, 2023). These formulas make it possible to compute the coefficients analytically from the weights and biases, thereby preserving a direct correspondence between the neural representation and its polynomial counterpart. The intercept term of the resulting polynomial model is given by

$$\beta_0 = v_0 + \sum_{j=1}^{h_1} v_j \left(\sum_{n=0}^{q} \frac{g^{(n)}(0)}{n!} (w_{0,j})^n \right), \tag{3.19}$$

where β_0 represents the baseline output of the model in the absence of feature contributions. The parameter v_0 corresponds to the output-layer bias, while $w_{0,j}$ denotes the bias weight associated with hidden neuron j. The inner summation reflects the Taylor expansion of the activation function evaluated at the bias-driven synaptic potential, thus incorporating the contribution of hidden-layer biases into the global polynomial representation. For interaction terms involving t input variables, NN2Poly derives the general coefficient

$$\beta_{l_1,l_2,\ldots,l_t} = \sum_{j=1}^{h_1} v_j \left(\sum_{n=t}^{q} \frac{g^{(n)}(0)}{(n-t)!} \cdot \frac{1}{m_1! \cdots m_p!} (w_{0,j})^{n-t} (w_{1,j})^{m_1} \cdots (w_{p,j})^{m_p} \right), \tag{3.20}$$

where the indices $l_1, \ldots, l_t$ identify the variables involved in the interaction term and the exponents m_i satisfy $m_1 + \cdots + m_p = t$. The multinomial structure arises from the expansion of $(\sum_i w_{i,j} x_i)^n$ and systematically accounts for all combinations of inputs that contribute to a term of order t. Each coefficient therefore aggregates the contributions of all hidden neurons, weighted by the output-layer parameters v_j and scaled by the derivatives of the activation function. These expressions imply that every polynomial coefficient β is a deterministic function of the trained weights and biases and the derivatives of the activation function, under the selected Taylor truncation order. As a result, NN2Poly does not rely on surrogate explanations or local approximations, instead, it delivers a global analytical representation of the learned mapping that supports the direct examination of main effects and higher-order interactions within a unified mathematical framework (Morala et al., 2021, 2023).

NN2Poly entails practical conditions to ensure that the polynomial approximation remains faithful to the trained network. Since the Taylor expansion is centered at zero, the synaptic potentials should remain close to this point. Therefore, all financial variables are scaled to $[-1, 1]$ to constrain activation inputs within a stable range. In the same way, ℓ_1-norm regularization is incorporated to discourage excessively large weight magnitudes that could amplify synaptic potentials and degrade the approximation quality (Morala et al., 2023). Moreover, the Taylor truncation order q must be selected to balance approximation fidelity

with interpretability, as polynomial complexity increases rapidly when higher-order terms are propagated across layers (Morala et al., 2021, 2023). Prior evidence suggests that lower-order interactions often capture most of the predictive signal, motivating the use of a maximum order $Q_{\max}$ that preserves analytical tractability. Finally, smooth activation functions are required to support stable Taylor-based distillation. Accordingly, $\tanh(\cdot)$, $\text{softplus}(\cdot)$, and $\sigma(\cdot)$ are considered as suitable candidates for the hidden layers (Morala et al., 2023).

Chapter 4

Empirical Analysis and Results

This chapter presents the results obtained from the application of a neural network model for corporate bankruptcy prediction based on financial ratios, together with an analysis of its interpretability using the NN2Poly methodology. The chapter follows the methodology pipeline described in Chapter 3, covering dataset selection and preprocessing, model building and evaluation, and the interpretation of the extracted polynomial structure, with particular emphasis on the economic meaning of the results. All codes and scripts developed for the analysis, including data preprocessing, model training, and the implementation of NN2Poly, are publicly available in the following repository: (Valverde Hueso, 2025) to ensure reproducibility.

4.1 Dataset Selection

The empirical analysis begins with the identification of a dataset that is appropriate for corporate bankruptcy prediction and consistent with the dual objective of this study, namely achieving competitive predictive performance while enabling economically meaningful interpretation through NN2Poly. To this end, three publicly available datasets frequently employed in the bankruptcy prediction literature were considered: the Polish Companies Bankruptcy dataset (Zięba, Tomczak, & Tomczak, 2016), the Taiwanese Bankruptcy dataset (Yeh, 2010), and an American firm-level panel dataset compiled from public sources (sowide, 2023). The candidate datasets were assessed according to criteria that directly affect both model development and the feasibility of polynomial distillation. These criteria include sample size, unit of analysis, degree of class imbalance, prevalence of missing values, dimensionality, and the resulting implications for interpretability and computational tractability under NN2Poly. Table 4.1 reports a comparative summary of their main characteristics.

Although the Taiwanese dataset benefits from the absence of missing values, its high dimensionality (95 financial ratios) introduces two practical limitations in the present setting. First, the interpretability of the learned mapping becomes less tractable as the number

of predictors increases, particularly when explanatory conclusions must remain grounded in well-established financial constructs. Second, high-dimensional inputs substantially increase the complexity of the polynomial representation produced by NN2Poly, since the number of candidate interaction terms grows combinatorially with both the number of variables and the maximum polynomial order. The American dataset, while offering a considerably larger number of observations, is organized as a firm-year panel. This structure introduces temporal dependence and repeated observations per firm, implying that appropriate validation would require panel-specific strategies (e.g., time-aware splitting and leakage control) and potentially alternative modeling choices. Such extensions, although relevant, fall outside the methodological scope of this study, which focuses on cross-sectional firm-level prediction and subsequent polynomial distillation.

Table 4.1: Comparative characteristics of candidate bankruptcy datasets

Dimension	**Polish Companies (1-year)**	**Taiwanese Companies**	**American Companies**
Data source	UCI ML Repository	UCI ML Repository	GitHub (U.S.)
Unit of analysis	Firm (financial snapshot)	Firm (financial snapshot)	Firm–year
Unique firms	7,027	6,819	8,971
Total observations	7,027	6,819	78,682
Number of variables	64	95	20
Time structure	Cross-sectional	Cross-sectional	Panel (1999–2018)
Missing values	5,835	0	0
Bankruptcy rate (%)	3.86	3.23	6.63

Note: Own elaboration.

By contrast, the Polish Companies Bankruptcy dataset (1-year horizon) provides a favorable balance between statistical adequacy and interpretability requirements. It contains a sufficiently large number of firms, a moderate number of variables (64 ratios), and a cross-sectional structure that avoids temporal dependence while remaining aligned with the classical bankruptcy prediction tradition. Although the dataset includes missing values, these can be addressed through the preprocessing procedures described in the methodology chapter, thereby preserving data quality without compromising the analytical objective. Importantly, the set of financial ratios is consistent with standard categories used in the literature (liquidity, profitability, leverage, and efficiency), which facilitates economic interpretation of the resulting polynomial terms. For these reasons, the Polish Companies Bankruptcy dataset (1-year) is selected as the primary data source for the empirical analysis reported in the subsequent sections.

4.2 Data Preprocessing and Descriptive Analysis

Prior to model construction, a structured preprocessing and exploratory descriptive analysis was conducted to ensure data quality, numerical stability, and the preservation of economic meaning in the explanatory variables. The selected dataset comprises 7,027 firm-level observations and 64 explanatory financial variables, in addition to a binary target variable labeled *class*, which indicates whether a firm becomes bankrupt in the subsequent years (1) or remains solvent (0). In the raw dataset, explanatory variables were generically labeled (Attr1–Attr64). To ensure economic interpretability, each attribute was renamed according to the financial ratio definition provided in the original documentation. This relabeling is essential for two reasons. First, it allows the predictive model outputs to be interpreted in terms of well-defined financial constructs (e.g., liquidity, leverage, profitability). Second, it ensures that the coefficients obtained from the NN2Poly polynomial representation can be directly associated with economically meaningful variables rather than abstract indices. A detailed description of the variables, including their financial interpretation and formulaic definition, is presented in Table 5.1 in the Appendix.

An initial inspection of data types confirms that all explanatory variables are numerical (integer or floating-point), which is appropriate for neural network modeling and avoids the need for categorical encoding or additional transformation of qualitative variables. The target variable reflects a one-year-ahead bankruptcy outcome, aligning the prediction task with a short-term early warning framework. As is typical in corporate bankruptcy datasets, the class distribution is highly imbalanced. As illustrated in Figure 4.1, only 3.86% of firms are labeled as bankrupt, while the vast majority correspond to non-bankrupt companies. This imbalance has important implications for model training and evaluation, as standard accuracy measures can be misleading in such settings. For this reason, the imbalance is explicitly addressed in the modeling phase through resampling strategies applied to the training data, as described in the methodology section.

A further relevant aspect of the data quality assessment concerns the presence and distribution of missing values across financial variables. Although the dataset provides a relatively large number of firm observations, several variables exhibit incomplete records, which may affect both model estimation and the stability of the subsequent polynomial approximation. Figure 4.2 presents the ten variables with the highest proportion of missing values. The results reveal a markedly uneven pattern. The ratio *Quick assets / Long-term liabilities* shows the highest level of incompleteness, with nearly 40% of observations missing. This is followed by the *Sales growth rate*, with approximately one quarter of its entries absent. A second group of variables, including *Operating profit / Financial expenses*, *Sales / Inventory*, and *Net profit / Inventory*, displays moderate levels of missingness, generally below 10%. The remaining variables in the figure, such as *Gross profit (3y avg) / Total assets*, *Debt repayment period*, *Gross profit – extraordinary items – financial expenses / Total assets*, *Cur-*

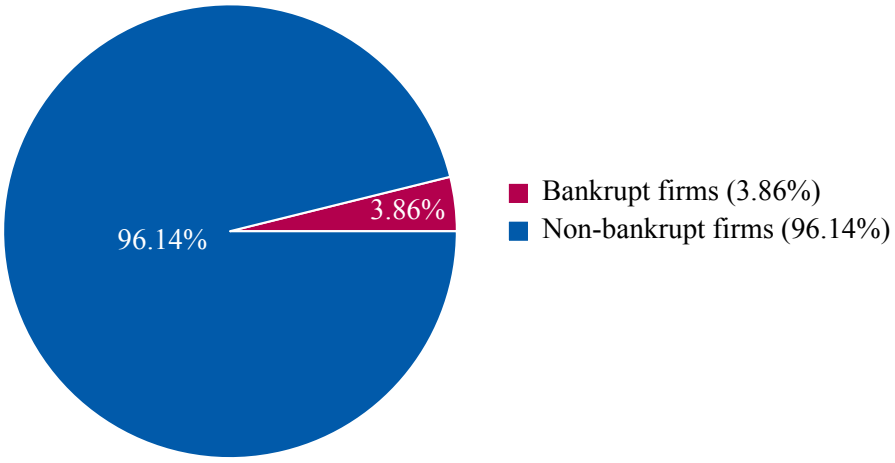

Figure 4.1: Distribution of bankrupt vs non-bankrupt firms. Source: Own Elaboration.

rent liabilities turnover days, and *Constant capital / Fixed assets*, present comparatively low percentages of missing values.

This heterogeneity in missingness suggests that data gaps are not random across financial dimensions, but rather concentrated in specific ratios, particularly those related to liquidity structure and growth dynamics. From a modeling perspective, high levels of missing data in certain variables can introduce noise, reduce effective sample size, and distort learned relationships if not properly addressed. To ensure robustness and to avoid introducing excessive noise through imputation, a threshold-based criterion was adopted. Variables with more than 20% missing observations were excluded from the analysis, as their effective information content is limited and their inclusion could distort both the learning process and the subsequent NN2Poly polynomial distillation. For the remaining variables, missing values were imputed using the median computed on the training data. Median imputation is preferred over mean imputation in this context because financial ratios often exhibit skewness and heavy tails, and the median provides a more robust estimate of central tendency under outliers. This approach preserves the overall distributional structure more effectively and supports numerical stability during neural network training

In addition, the descriptive statistics revealed substantial dispersion and the presence of extreme values in several financial ratios, particularly in turnover measures and indicators related to liabilities and profitability. Although extreme realizations are plausible in bankruptcy prediction settings, often reflecting firms under acute financial stress, they may hinder model estimation by inducing numerical instability and disproportionate gradient updates during neural network training. A targeted outlier inspection was conducted using boxplots for representative variables, as reported in Figure 4.3. The figure shows highly right-skewed distributions for ratios such as such as *Inventory turnover (days)*, *Receivables turnover (days)*, and their combined measure *Receivables and inventory turnover (days)*, where a small number

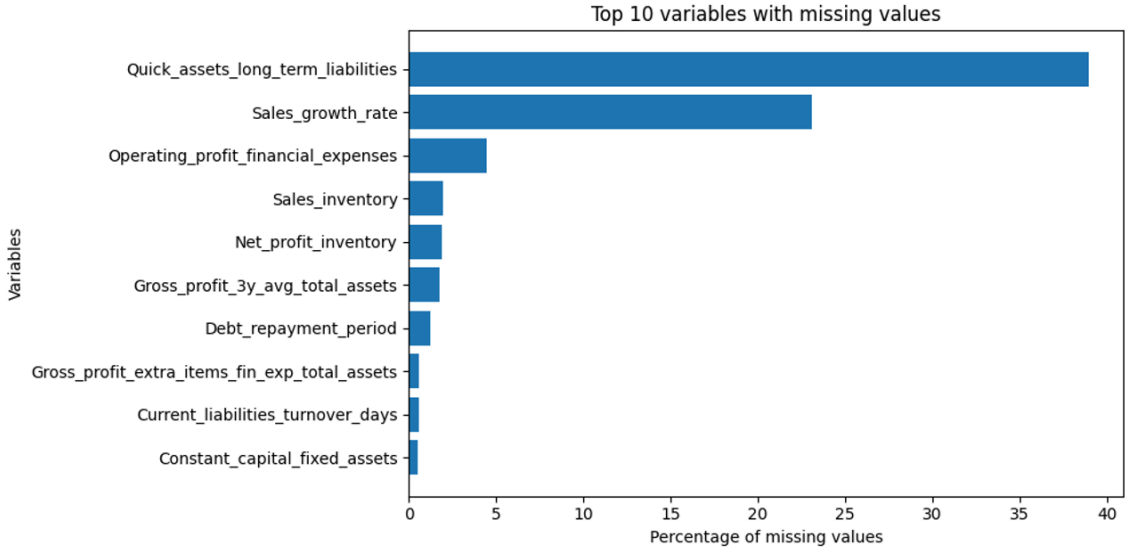

Figure 4.2: Top 10 variables with missing values. Source: Own Elaboration.

of observations take extremely large values relative to the central mass of the data. Similarly, ratios involving sales components in the denominator, such as *Liabilities (excluding cash) / Sales*, exhibit pronounced outliers that are consistent with near-zero denominators and volatile accounting components. The plots also highlight extreme values in profitability-related ratios, such as *Retained earnings / Total assets* and *Operating profit / Financial expenses*, which may arise from sharp operating losses or highly leveraged cost structures. To mitigate the impact of such numerical extremes while preserving economically meaningful information associated with financial distress, a conservative winsorization strategy was adopted. Specifically, only the ratios exhibiting clear numerical noise and heavy-tailed behavior were winsorized at the 1st and 99th percentiles. This approach limits the influence of extreme realizations.

The distributional analysis further confirms that several core financial ratios exhibit pronounced departures from symmetry, with strong skewness and heavy tails. Figure 4.4 reports representative histograms for key profitability, leverage, liquidity, and efficiency indicators. The return-on-assets ratio, measured as *ROA (net profit / total assets)*, displays a highly asymmetric distribution with a long left tail, reflecting the presence of firms reporting substantial operating losses. This behavior is economically plausible in a bankruptcy setting, as distressed firms may experience severe negative profitability well before formal insolvency. A similarly skewed pattern emerges for leverage and liquidity ratios. The histogram of *Total liabilities / total assets* concentrates most observations at relatively low to moderate leverage levels, while a small fraction of firms displays extremely high values, consistent with unusually large liabilities relative to asset bases. In parallel, the *Current assets / short-term liabilities* ratio shows a dense concentration near the lower range and a long right tail, indicating that while many firms operate with limited short-term coverage, a minority exhibits

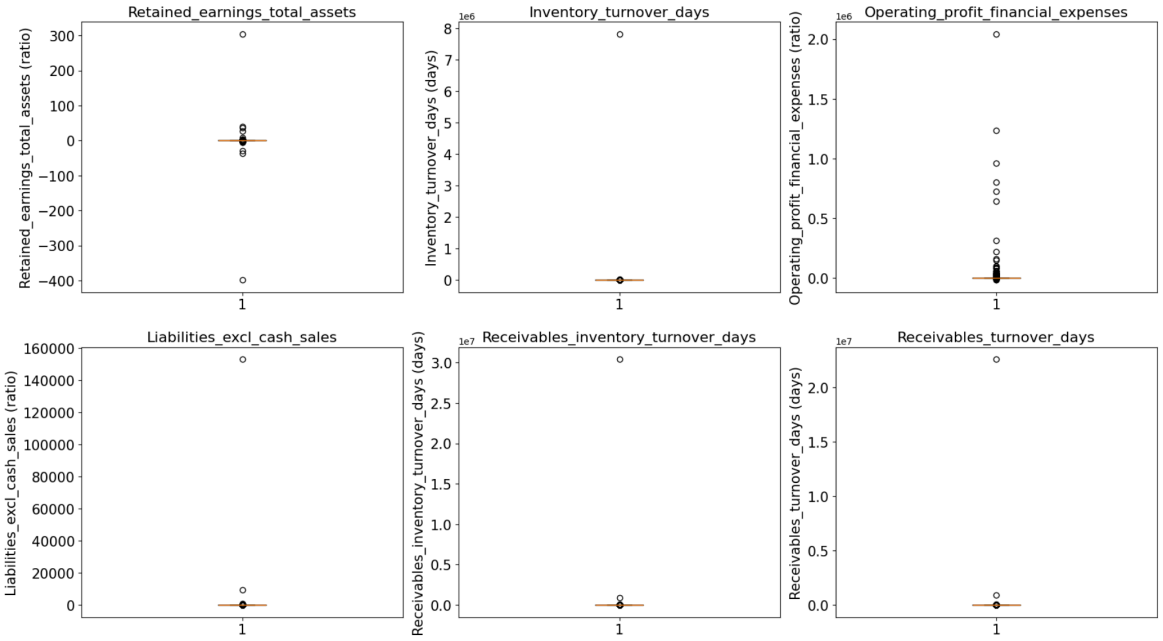

Figure 4.3: Boxplots of selected ratios before transformation. Source: Own Elaboration.

exceptionally high liquidity buffers. Efficiency measures display comparable heavy-tailed behavior. As illustrated by the distribution of *Inventory turnover (days)*, the majority of firms cluster at lower turnover durations, whereas the right tail extends toward substantially larger values. This right-skewness is consistent with heterogeneous operating cycles across industries and with potential inventory accumulation among firms facing weakening demand. Firm size, proxied by the logarithm of total assets, follows a more regular and approximately symmetric distribution, consistent with the stabilizing effect of the logarithmic transformation.

Complementing the univariate distributional analysis, Figure 4.5 compares the behavior of selected financial ratios across bankrupt and non-bankrupt firms, providing preliminary evidence of systematic differences between the two groups. Liquidity-related measures show clear contrasts. Working capital is, on average, markedly lower among bankrupt firms, with the distribution for distressed firms concentrated closer to zero and extending further into negative values. This pattern is consistent with tighter short-term financial positions and limited buffers to absorb operational shocks. In parallel, the cash flow adequacy ratio is substantially lower for bankrupt firms, with the median shifting toward more negative values, indicating weaker internal cash-generation capacity relative to obligations.

Leverage and payment dynamics also differ between the groups. Liabilities turnover days and short-term liabilities to sales days both exhibit higher central values among bankrupt firms, suggesting slower repayment cycles and greater pressure from short-term commitments. Similarly, current liabilities turnover days are higher for distressed firms, reflecting delays in settling short-term obligations. These patterns align with the expected deterioration

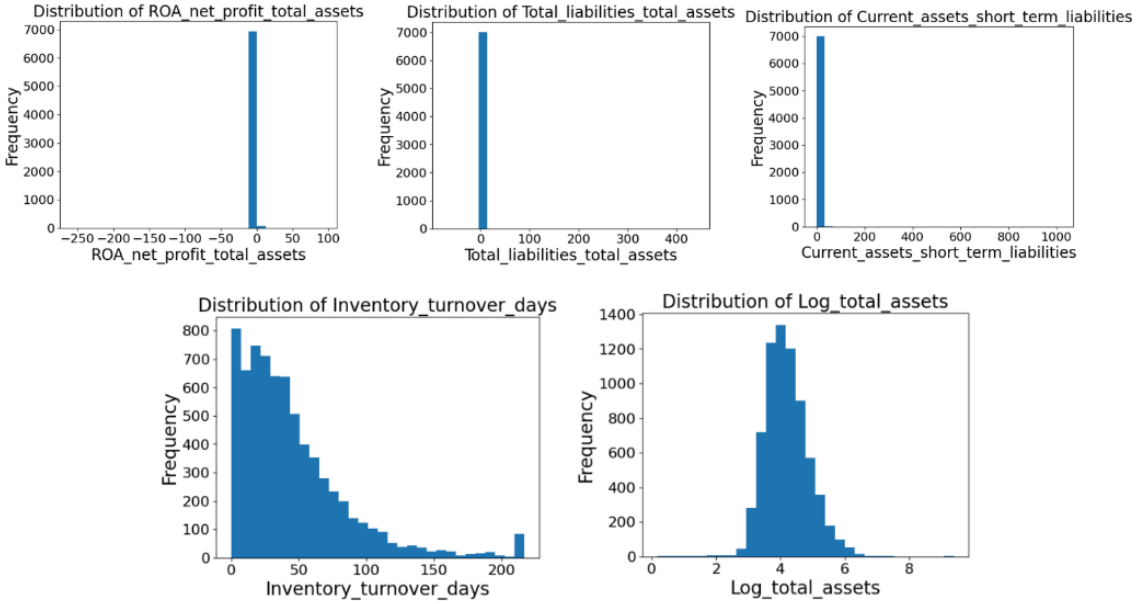

Figure 4.4: Histograms of selected financial ratios. Source: Own Elaboration.

in payment discipline and creditor pressure as firms approach insolvency. Operational efficiency indicators display more moderate but still visible differences. Inventory turnover days (COGS-based) tend to be higher for bankrupt firms, indicating slower inventory rotation and potential demand weakness or inventory accumulation prior to failure. Although dispersion is high in both groups, the upward shift in the distribution for bankrupt firms is consistent with operational inefficiencies frequently associated with financial distress. Overall, the boxplots show that bankrupt firms are characterized by weaker liquidity positions, slower liability turnover, and signs of declining operational efficiency.

A more detailed inspection of Figure 4.6 reveals several economically meaningful pairs and clusters of variables with strong linear associations (yellow tones) and marked inverse relationships (dark blue tones). Profitability ratios form a particularly dense cluster in the upper-left region, with measures such as EBIT to total assets, gross profit to total assets, and gross profit scaled by interest or depreciation presenting perfect positive correlations that reflect their shared focus on operating performance. Similarly, liquidity indicators, such as the current ratio and the quick ratio, are also highly correlated, as they capture very similar aspects of short-term solvency based on overlapping balance-sheet components. High correlations are also present among turnover indicators expressed in days, such as current and short-term liabilities turnover, indicating consistent patterns in payment behavior. Moreover, strong negative correlations are observed between leverage and liquidity measures. For instance, working capital relative to total assets is almost perfectly inversely correlated with short-term liabilities relative to total assets, highlighting the direct trade-off between reliance on short-term financing and available liquidity buffers. Overall, the correlation matrix confirms substantial multicollinearity across related profitability, liquidity, and leverage ratios.

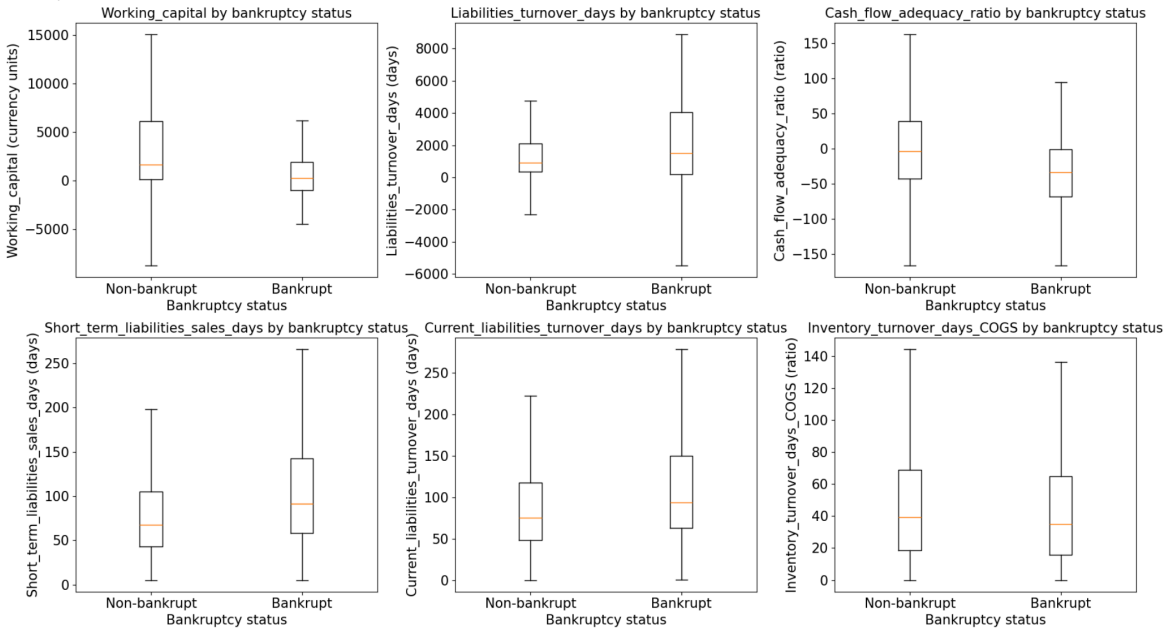

Figure 4.5: Boxplots of key ratios by bankruptcy status. Source: Own Elaboration.

While this is not a direct limitation for neural networks, it supports the use of regularization and motivates NN2Poly to obtain a structured and interpretable representation of the joint effects of correlated predictors on bankruptcy risk.

Lastly, all explanatory variables were scaled to the interval $[-1, 1]$. This transformation supports numerical stability during neural network training by placing predictors on a comparable scale, preventing variables with large magnitudes from disproportionately influencing gradient updates and accelerating convergence in gradient-based optimization. In addition, the chosen scaling is methodologically aligned with the subsequent NN2Poly stage, since Taylor-based polynomial representation assumes that synaptic potentials remain within a bounded neighborhood of the expansion point. Constraining inputs to $[-1, 1]$ contributes to keeping pre-activation values in a stable range, thereby improving the reliability of the polynomial approximation (Morala et al., 2023). At the conclusion of the preprocessing pipeline, the resulting dataset contains no missing values, as all remaining gaps were addressed through median imputation after excluding variables exceeding the missingness threshold. These transformations yield a numerically well-conditioned input matrix that is suitable for predictive modeling and provides an appropriate foundation for the NN2Poly-based interpretability analysis developed in subsequent sections.

4.3 Implementation of the Neural Network Predictive Model

The implementation process begins with partitioning the dataset into training and test subsets using an 80/20 stratified *split*, so that the original proportion of bankrupt and non-bankrupt

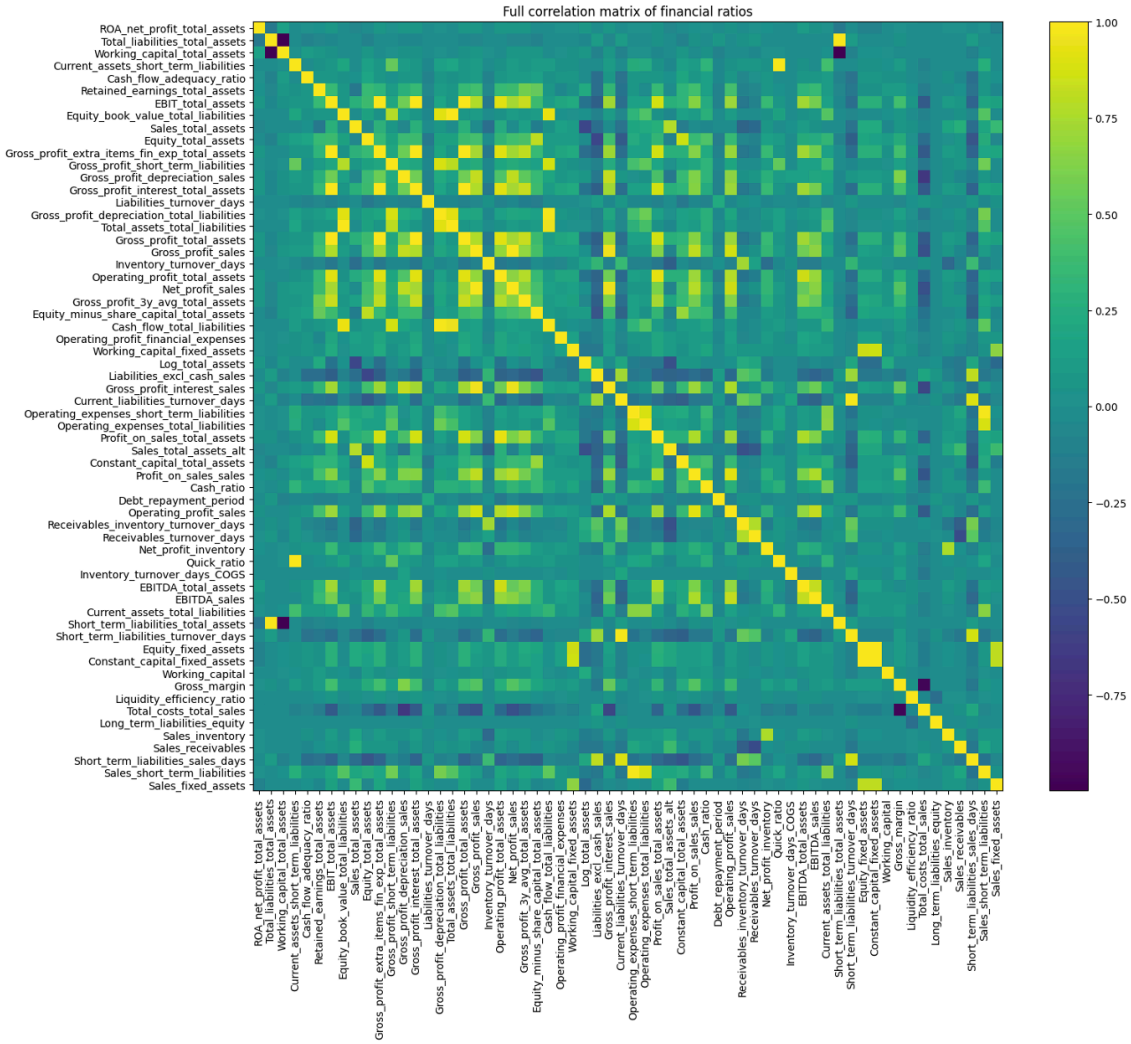

Figure 4.6: Correlation matrix of financial ratios. Source: Own Elaboration.

firms is preserved in both subsets. Given the pronounced class imbalance characteristic of bankruptcy prediction, the training set was balanced prior to model estimation using the ROSE strategy, described in Section 3.1, with the objective of obtaining an approximately balanced class distribution (around 50/50) that facilitates the learning of patterns associated with the minority class. In contrast, the test set was deliberately left unbalanced to ensure a realistic evaluation under conditions similar to those observed in practice. As an initial reference, a majority-class baseline was computed on the original (unbalanced) training set. In this dataset, a trivial rule that systematically predicts the majority class would achieve an *accuracy* close to 0.96, which illustrates why accuracy alone is of limited informativeness in highly imbalanced settings, as high values can be obtained without correctly identifying any cases of insolvency.

After the initial train-test partition, model development proceeded exclusively within the training subset. To support hyperparameter selection without compromising the integrity

of the final test evaluation, the training data were further split into an internal training set (80%) and a validation set (20%). The validation subset was used to compute the AUC, which served as the primary criterion for model comparison and selection given the strong class imbalance. A grid search over 60 configurations of key hyperparameters was then conducted. The search space included the number of hidden neurons (10, 20, 30, 40, and 50), the activation function in the hidden layer (sigmoid, softplus, and tanh), and the batch size (8, 16, 32, and 64). All candidate networks were trained under identical optimization settings to ensure comparability. Specifically, models were compiled with the Adam optimizer, a learning rate of 0.001, and the binary cross-entropy loss function. The top five configurations ranked by validation AUC are reported in Table 4.2. The results indicate that architectures using the tanh activation function consistently achieve superior discriminative performance relative to sigmoid and softplus alternatives.

Units	**Activation**	**Batch Size**	**AUC**
50	tanh	8	0.9751
40	tanh	8	0.9634
30	tanh	8	0.9633
20	tanh	8	0.9508
50	tanh	16	0.9466

Table 4.2: Neural network hyperparameter configurations and corresponding AUC scores. Source: Own elaboration.

Based on the validation-based grid search, the best-performing configuration corresponded to a network with 50 hidden units, a tanh activation function, and a batch size of 8, achieving a validation AUC of 0.9751 (Table 4.2). This configuration was therefore selected as the final architecture and subsequently retrained under the same optimization settings before being assessed on the held-out test set. It is important to note that the validation AUC is computed on data drawn from the balanced training procedure (via ROSE), whereas the final evaluation is conducted on the original, highly imbalanced test set. As a consequence, these AUC values are not directly comparable in magnitude. When the class distribution is artificially balanced, the classifier is exposed to a higher prevalence of bankrupt firms during training and validation, which typically facilitates discrimination and can improve performance estimates relative to evaluations under the natural base rate. In contrast, the unbalanced test set reflects the true prevalence of insolvency, where the minority class is rare and the score distributions of bankrupt and non-bankrupt firms overlap more strongly, yielding a more conservative but practically relevant estimate of generalization performance. The discriminative capacity of the selected neural network was then evaluated using the ROC curve, which provides a threshold-independent assessment of classification performance by characterizing the trade-off between sensitivity (TP rate) and specificity (TN rate) across all

possible decision thresholds. This representation is particularly informative in highly imbalanced settings, where single-threshold metrics may provide a misleading view of model performance. Figure 4.7 reports the ROC curve obtained on the test set, yielding an AUC of 0.7578. This value indicates a satisfactory level of global discrimination between bankrupt and non-bankrupt firms and suggests that the model captures meaningful separation between the two classes independently of the operating threshold selected for classification.

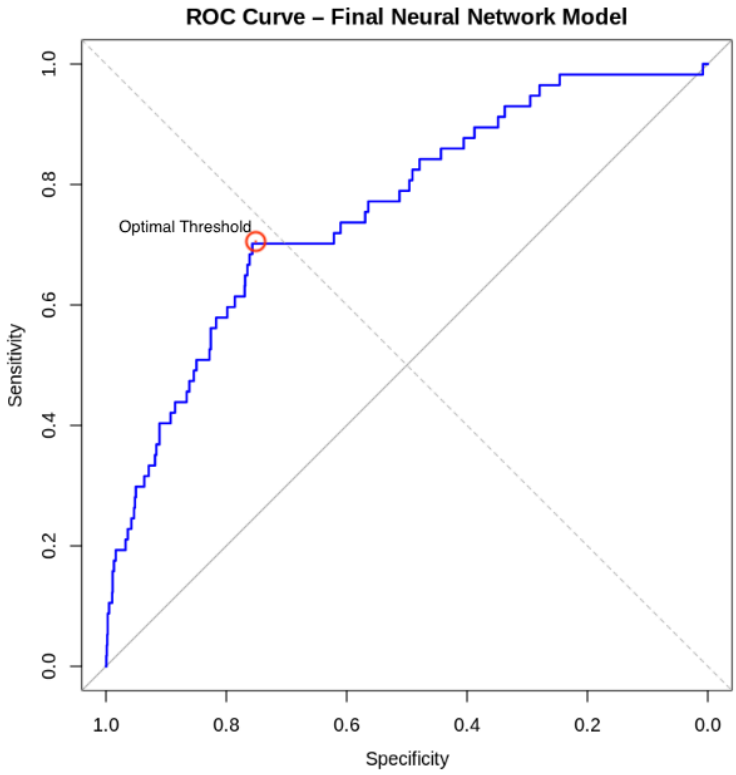

Figure 4.7: ROC Curve for the Final Neural Network Model on the Test Set. Source: Own elaboration.

While the ROC curve evaluates discrimination irrespective of a specific threshold, practical deployment requires selecting an operating point that reflects the desired balance between detecting bankrupt firms and limiting false alarms. Accordingly, an optimal classification threshold of 0.2418 was identified on the test set using Youden's criterion, which maximizes the sum of sensitivity and specificity. To contextualize the implications of this threshold choice, performance was compared against the standard cutoff of 0.5. Under the default threshold, the model attains high specificity (0.8873) but comparatively low sensitivity (0.4211), resulting in an accuracy of 0.8684. This profile indicates that the classifier is conservative, correctly identifying most solvent firms but missing a substantial share of bankrupt cases. In contrast, applying the Youden-based threshold substantially improves sensitivity to 0.7018, while specificity decreases to 0.7569, producing an accuracy of 0.7546. This shift reflects a more balanced operating regime that prioritizes the detection of distressed firms at

the cost of additional false positives, which is often preferable in early-warning applications where the cost of failing to flag insolvency risk may be high.

These results are further detailed in Table 4.3, which reports the confusion matrices obtained under the standard threshold (0.5) and the Youden-based threshold (0.2418). The comparison makes explicit the operational trade-off between false positives and false negatives in bankruptcy prediction. Under the default threshold, the model produces relatively few false alarms but misses a substantial number of bankrupt firms (FN). In contrast, lowering the threshold increases the number of detected bankruptcies (TP) while also increasing false positives, reflecting a more sensitive screening regime. In many early-warning and credit-risk applications, prioritizing sensitivity is often justified because failing to flag a truly distressed firm can entail higher economic costs than conducting additional screenings triggered by false positives. From this perspective, the Youden-based operating point provides a more suitable balance for bankruptcy detection, even if it reduces specificity relative to the conservative standard threshold.

Threshold = 0.5 (Standard)

	Class 0	**Class 1**
Predicted 0	1197	33
Predicted 1	152	24

Optimal Threshold (Youden = 0.2418)

	Class 0	**Class 1**
Predicted 0	1021	17
Predicted 1	328	40

Table 4.3: Confusion matrices for the final model using the standard threshold (0.5) and the optimal Youden threshold. Source: Own elaboration.

4.4 Application of the Interpretability Technique: NN2Poly

This section applies the NN2Poly methodology to the final neural network model in order to obtain a transparent, analytically tractable representation of the learned decision function. The aim of this stage is not to enhance predictive performance, but rather to assess the extent to which the black-box behavior of the trained model can be distilled into a low-order polynomial and interpreted in economically meaningful terms, namely through financial ratios and their interaction structure. The analysis proceeds in two steps. First, the selected neural network is taken as the reference predictive model and its role as a black-box benchmark is established. Second, NN2Poly is used to extract the corresponding polynomial approximation, which is then examined with respect to fidelity to the original network, structural complexity, and interpretability of the resulting main and interaction effects.

4.4.1 Final Neural Network Model

Once the optimal hyperparameter configuration was selected, the neural network was retrained under the constraints required for the NN2Poly transformation. The final architecture consists of a single hidden layer with 50 neurons and tanh activation, followed by a linear output that produces logits. Bankruptcy probabilities are obtained by applying a sigmoid mapping after training. This specification is consistent with NN2Poly, which operates on smooth activations and leverages Taylor-based expansions of the network function.

In addition, an ℓ_1 regularization term was incorporated to encourage moderate synaptic weights, a condition that supports the stability and interpretability of the subsequent polynomial representation. The constrained model was trained on the balanced training set using the Adam optimizer (learning rate 0.001), a batch size of 8, and 100 epochs. Training was conducted using the binary cross-entropy loss function applied to the network logits, which measures the discrepancy between predicted probabilities and observed class labels, penalizing confident misclassifications more heavily. Figure 4.8 summarizes the training history. Both training and validation curves stabilize after the initial epochs and remain at comparable levels, with only sporadic fluctuations in validation loss. A similar behavior is observed for accuracy, which remains stable and closely aligned between training and validation. Overall, these dynamics suggest stable convergence without evident overfitting, supporting the use of the final network as a reliable black-box reference for the NN2Poly approximation.

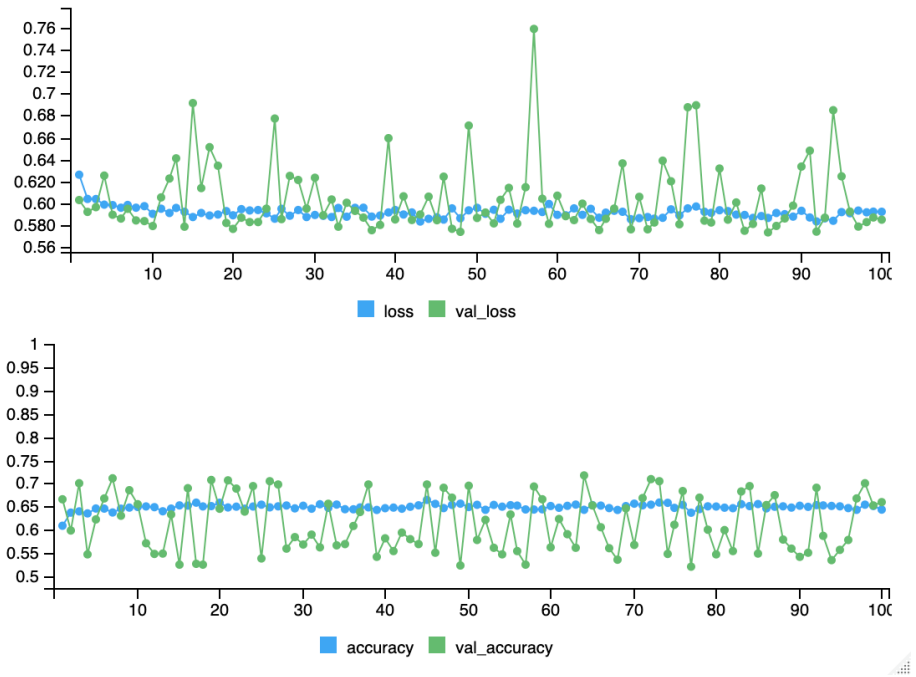

Figure 4.8: Training history of neural network. Source: Own Elaboration.

The final constrained model was subsequently evaluated on the test set, which preserves the original class imbalance in order to provide a realistic assessment of performance. The re-

sulting confusion matrix is reported in Table 4.4. The model correctly identifies 39 bankrupt firms (TP) and 954 non-bankrupt firms (TN), while producing 395 false positives and 18 false negatives. These results correspond to a sensitivity of 0.707 and a specificity of 0.684, indicating a relatively balanced detection profile between distressed and non-distressed firms. Compared with the unconstrained neural network presented in Section 4.3, the constrained specification maintains a similar level of sensitivity while exhibiting a moderate reduction in specificity. This slight loss in discriminative sharpness can be interpreted as the cost of imposing structural constraints, namely ℓ_1 regularization and smooth activation compatibility, to facilitate the NN2Poly transformation. Importantly, the constrained model continues to capture a substantial proportion of bankrupt firms, preserving its usefulness as an early-warning instrument. Overall, although the constrained architecture does not aim to maximize predictive performance, it achieves a satisfactory trade-off between sensitivity and specificity while satisfying the mathematical conditions required for polynomial extraction.

	Class 0	**Class 1**
Predicted 0	954	18
Predicted 1	395	39

Table 4.4: Confusion matrix for the evaluated model. Source: Own elaboration.

4.4.2 Polynomial Representation

After training the constrained neural network, the NN2Poly algorithm is applied to derive a global polynomial representation of the learned decision function. In this study, the Taylor expansion is truncated at a maximum order of three. This choice reflects a deliberate trade-off, third-order polynomials are sufficiently expressive to capture nonlinear main effects and low-order interactions among financial ratios, while avoiding the exponential growth in terms that would arise with higher expansion orders. Consequently, the resulting polynomial remains structurally interpretable, as it is composed of linear terms, pairwise interactions, and a limited number of third-order interactions.

The polynomial obtained through NN2Poly constitutes a functional approximation of the trained neural network. It does not involve retraining or re-estimation. Rather, its coefficients are deterministically computed from the weights and biases of the constrained model, as described in the previous section. The objective is to reconstruct the mapping learned by the neural network in an explicit algebraic form, enabling direct inspection of the contribution of individual ratios and their joint effects. To assess whether the polynomial faithfully reproduces the behavior of the neural network, a direct comparison of their classification outputs is conducted. Specifically, predictions generated by the neural network and by its polynomial counterpart are evaluated using the same decision threshold, and their agreement is summa-

rized in Table 4.5. Importantly, this confusion matrix does not measure predictive performance against the true bankruptcy labels. Instead, it quantifies the concordance between the two decision functions. The results indicate perfect agreement. All 972 instances classified as non-bankrupt by the neural network are also classified as non-bankrupt by the polynomial, and all 434 instances classified as bankrupt coincide exactly. The resulting agreement accuracy of 1 confirms that, under the selected polynomial order, the NN2Poly representation preserves the classification decisions of the original constrained neural network on the evaluated dataset.

	Pred. NN Class 0	**Pred. NN Class 1**
Pred. Poly Class 0	972	0
Pred. Poly Class 1	0	434

Table 4.5: Confusion matrix NN vs Poly. Source: Own elaboration.

A complementary fidelity assessment can be conducted at the continuous level by comparing the raw decision scores produced by the neural network and by its NN2Poly approximation. Figure 4.9 reports a diagonal scatter plot of neural network logits against polynomial logits. The points lie almost entirely on the 45° reference line, indicating an almost one-to-one correspondence between the two scoring functions across the full range of predictions. This result corroborates the perfect agreement observed in Table 4.5 and provides stronger evidence that fidelity is not limited to thresholded classifications but extends to the continuous decision function itself. Minor departures from the diagonal appear only at the most extreme logit values, suggesting that approximation errors are negligible in the region where most observations lie and remain limited even in the tails.

Figures 4.10 and 4.11 provide a layer-wise diagnostic of the Taylor approximation underlying the NN2Poly transformation. In the hidden layer (Figure 4.10), the polynomial approximation closely overlaps the true activation function in the central region of the synaptic potential domain. Importantly, this region coincides with the highest density of activation potentials, as illustrated by the green density curve. Consequently, the approximation error (blue curve) remains close to zero for the vast majority of observations effectively processed by the network. Noticeable deviations between the true activation and its Taylor expansion arise only at the tails of the activation range, where synaptic potentials are less frequently observed due to prior feature scaling and ℓ_1-based regularization. In the output layer (Figure 4.11), the Taylor approximation is virtually indistinguishable from the true function across the entire range of activation values. This behavior is consistent with the linear structure of the logit layer prior to the final sigmoid transformation, which facilitates an almost exact polynomial representation. Taken together, these diagnostics confirm that the NN2Poly approximation maintains high fidelity precisely in the regions of the input space that carry the greatest empirical relevance. Approximation discrepancies are confined to sparsely popu-

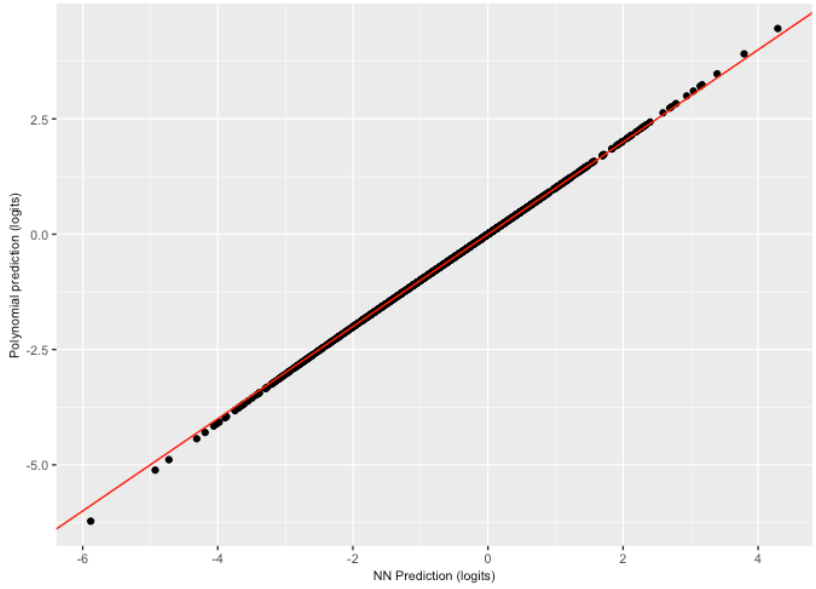

Figure 4.9: Diagonal plot NN logits versus Poly logits. Source: Own Elaboration.

lated activation extremes and therefore exert negligible influence on the model's effective classification behavior.

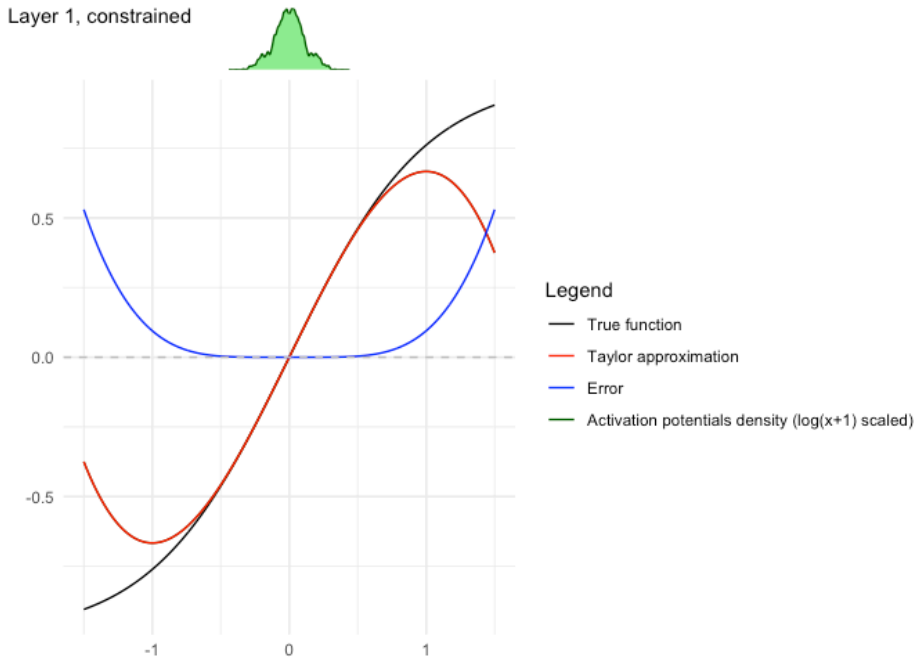

Figure 4.10: Taylor approximation for the hidden layer. Source: Own Elaboration.

To further evaluate the structural interpretability of the extracted polynomial, Figure 4.12 presents the coefficients with the largest absolute magnitude. The bars are ordered according to their absolute value, which reflects the relative strength of each term in the global decision function. The height of each bar represents the magnitude of the coefficient, while the color

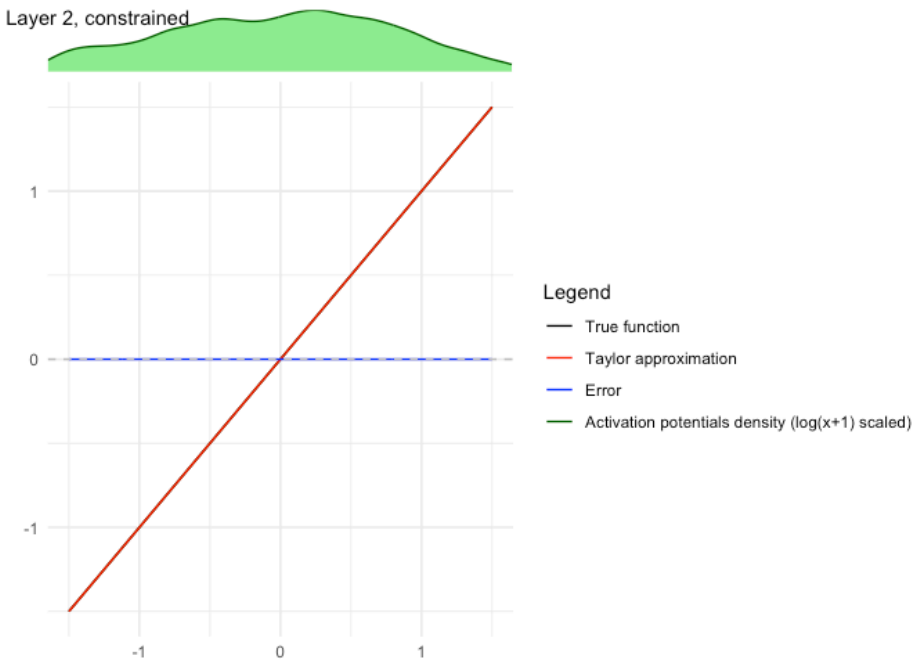

Figure 4.11: Taylor approximation for the output layer. Source: Own Elaboration.

indicates its sign, distinguishing between positive and negative contributions to the predicted bankruptcy logit. The ranking reveals a clear concentration of explanatory power in a limited subset of terms. Most of the dominant contributions correspond to main effects rather than higher-order interaction terms. Although the polynomial was computed up to third order, interaction coefficients do not appear among the largest magnitudes, suggesting that the predictive structure of the constrained neural network is primarily driven by individual financial ratios rather than by complex multi-way interactions. This pattern is consistent with the earlier regularization strategy and the bounded weight configuration imposed to support NN2Poly stability. From an interpretability standpoint, the predominance of main effects facilitates economic interpretation, as the model's behavior can be largely understood in terms of the marginal influence of specific financial indicators.

Once the structure of the polynomial approximation has been established, Figure 4.13 identifies the financial ratios associated with the ten most influential terms in the model. These dominant terms are primarily linked to core dimensions of firm performance, particularly profitability, repayment capacity, balance sheet strength, and firm size, highlighting economically meaningful drivers of bankruptcy risk.

The most influential term corresponds to *gross profit over a three-year average relative to total assets*. Its large negative coefficient indicates that sustained operating profitability substantially reduces the predicted bankruptcy logit. This result is economically consistent, as firms capable of generating stable operating returns over time are better positioned to absorb shocks and maintain solvency. Similarly, *EBITDA to total assets* appears among the

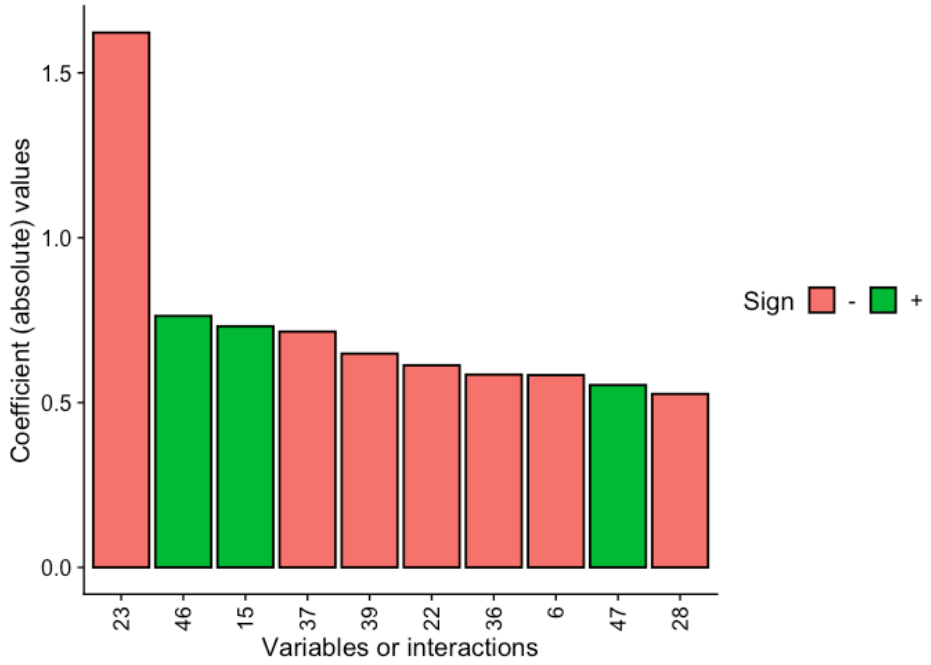

Figure 4.12: Top polynomial coefficient magnitudes and signs. Source: Own Elaboration.

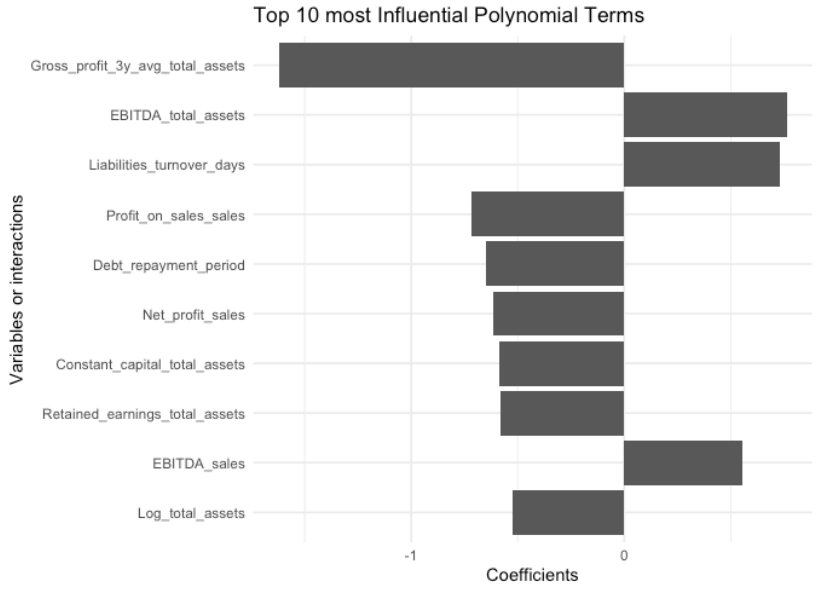

Figure 4.13: Top 10 most influential polynomial terms. Source: Own Elaboration.

strongest contributors with a positive magnitude in absolute terms (sign interpreted relative to the logit). This ratio captures operating cash-generating capacity before financial structure effects. A higher EBITDA-to-assets ratio reflects stronger operating performance and therefore contributes to lowering bankruptcy risk when the coefficient is negative, reinforcing the central role of operating profitability.

The presence of *liabilities turnover in days* highlights the importance of payment behav-

ior and short-term financial management. Longer turnover periods typically signal delays in meeting obligations, which may reflect liquidity constraints or operational stress. The sign of its coefficient suggests that deterioration in payment efficiency increases predicted distress. *Profit on sales (sales margin)* and *net profit to sales* both represent profitability at different levels of the income statement. Their influence confirms that both operating and bottom-line margins are important in distinguishing solvent from distressed firms. Lower margins reduce the firm's internal capacity to generate buffers against adverse conditions. The *debt repayment period* directly captures repayment capacity. A longer repayment horizon generally implies greater leverage pressure or weaker cash flow generation relative to debt, increasing insolvency risk. Its prominence in the polynomial underscores the importance of debt sustainability in the learned decision function.

Balance sheet strength is further reflected in variables such as *retained earnings to total assets*. This ratio proxies cumulative profitability and internal capital accumulation. Firms with higher retained earnings relative to assets typically exhibit stronger capital structures and greater resilience, thereby reducing bankruptcy likelihood. Size effects are captured by *log of total assets*. Firm size often proxies diversification capacity, access to credit markets, and bargaining power. Larger firms may benefit from scale advantages and greater resilience, which is reflected in the direction of its coefficient. Finally, *EBITDA to sales* reinforces the role of operating efficiency, complementing asset-based profitability measures. Together, these profitability and efficiency indicators dominate the top-ranked polynomial terms, indicating that the neural network—and its polynomial representation—relies primarily on economically intuitive performance metrics rather than obscure interactions. Overall, the dominance of profitability, repayment capacity, leverage-related, and size variables confirms that the polynomial approximation captures financially coherent drivers of insolvency. The absence of higher-order interaction terms among the most influential coefficients further suggests that bankruptcy risk in this model is largely explained by strong marginal effects of key financial ratios rather than by complex multi-way interactions. This enhances the transparency and practical interpretability of the extracted polynomial representation.

Chapter 5

Conclusions

Corporate bankruptcy prediction is a highly relevant and well-researched problem in financial analysis due to its significant economic and social consequences. Firm failure affects not only shareholders and creditors, but also employees, suppliers, and the stability of financial systems as a whole. The ability to identify companies at risk of bankruptcy at an early stage is therefore relevant for supporting preventive actions, improving credit-risk assessment, and mitigating systemic risk. As financial environments become increasingly complex and data intensive, predictive models have gained importance as tools to support decision making in this context. However, for such models to be effectively adopted in practice, they must not only deliver strong predictive performance but also provide transparent and economically meaningful explanations of their predictions.

As discussed in Chapter 2, the literature on bankruptcy prediction spans both traditional statistical approaches and more recent machine learning techniques. Early models based on discriminant analysis and logistic regression established the relevance of financial ratios related to liquidity, leverage, profitability, and firm size, while subsequent work has shown that machine learning models—particularly ensemble methods and artificial neural networks—often deliver superior predictive performance. However, the limited transparency of these approaches has constrained their practical applicability, especially in financial and regulatory environments where accountability and justification are required. Most empirical studies address this limitation through post-hoc interpretability tools such as SHAP or LIME, which provide useful case-level attributions but remain indirect, as they do not yield an explicit representation of the learned decision function across the input space. This thesis contributes to closing this gap by moving beyond post-hoc attributions and providing a global, algebraic representation of a neural bankruptcy classifier through NN2Poly. By transforming the trained network into a low-order polynomial with demonstrated fidelity, the approach makes it possible to interpret bankruptcy risk in terms of identifiable financial ratios and a parsimonious set of interaction structures.

Building on this motivation, the study developed and implemented a methodological

framework that integrates the predictive capacity of artificial neural networks with the global transparency enabled by NN2Poly. The empirical analysis was conducted using the Polish Companies Bankruptcy dataset, which provides firm-level financial ratios measured one year prior to the bankruptcy event and includes both bankrupt and non-bankrupt firms. This dataset, widely adopted in the bankruptcy prediction literature, contains indicators spanning key dimensions of financial condition, including liquidity, leverage, profitability, efficiency, and firm size. The central aim was to evaluate whether a neural network trained on these variables could achieve competitive discrimination between distressed and non-distressed firms, while also allowing its learned decision function to be translated into an explicit, economically interpretable form. The methodology followed a structured sequence of steps. First, a comprehensive preprocessing stage was conducted to improve data quality and numerical stability, including the treatment of missing values and extreme observations, the scaling of predictors to a bounded interval to support stable optimization and subsequent polynomial approximation, and the application of resampling strategies to address severe class imbalance in the training data. The dataset was then partitioned into training, validation, and test subsets, preserving the original base rate of bankruptcy in the hold-out evaluation set. Next, an MLP with a single hidden layer was estimated and tuned through a grid search over core hyperparameters, including the number of hidden units, the choice of activation function, and the mini-batch size. Model selection relied primarily on validation AUC to ensure threshold-independent comparability across configurations. The best-performing architecture consisted of 50 hidden neurons with a tanh activation. Finally, to ensure methodological compatibility with NN2Poly, the selected network was retrained under additional constraints, including a linear logit output, smooth hidden-layer activations, and ℓ_1-based regularization to promote bounded and sparse weight structures that facilitate faithful and parsimonious polynomial extraction.

The selected architecture was subsequently evaluated on the test data to obtain an unbiased estimate of out-of-sample performance. Using the optimal classification threshold identified via Youden's index, the model obtained a sensitivity of 0.707 and a specificity of 0.684, resulting in an overall accuracy of 0.706. These results reflect a more balanced operating regime that prioritizes the detection of financially distressed firms while maintaining reasonable control over false positives.

Following the construction of the predictive model, NN2Poly was applied to obtain a global polynomial approximation of the trained neural network. The quality of the polynomial approximation was assessed by comparing neural network and polynomial classifications, as well as through logit alignment and Taylor expansion diagnostics. These analyses confirmed that the polynomial reproduced the neural network's decision function with very high fidelity. The resulting polynomial structure revealed that the most influential terms correspond to financially intuitive variables related primarily to profitability, repayment capacity, balance sheet strength, and firm size. Several profitability-based ratios, including

gross profit averaged over three years relative to total assets, EBITDA scaled by assets, and EBITDA margin, emerged among the dominant components, highlighting the central role of sustained operating performance and cash-generating ability in determining bankruptcy risk. In addition, variables capturing debt servicing and payment behavior, such as liabilities turnover days and the debt repayment period, exhibited strong influence, indicating that delays in meeting financial obligations are closely associated with financial distress. Measures of capital structure and internal financing strength, including retained earnings and constant capital relative to total assets, entered the polynomial with stabilizing effects, reflecting the protective role of accumulated resources and balance sheet solidity. Firm size, proxied by the logarithm of total assets, also appeared as a key stabilizing factor, consistent with the greater resilience of larger firms. Overall, these results suggest that the neural network bases its predictions on a coherent set of economic mechanisms grounded in profitability, financial structure, and repayment capacity, rather than on opaque or purely statistical patterns.

Despite the robustness of the results, several limitations should be acknowledged. First, the empirical analysis relies on a single dataset of Polish firms and adopts a static one-year-ahead prediction setting. This design constrains the external validity of the findings across institutional contexts, sectoral structures, and alternative forecasting horizons. Second, the NN2Poly framework imposes methodological constraints on the predictive model, particularly regarding network architecture and the use of smooth activation functions. These requirements narrow the set of admissible neural configurations and exclude certain alternatives that may yield stronger predictive performance but are not suitable for Taylor-based polynomial extraction. Future research can extend this framework along several lines. First, applying the methodology to multi-period or panel datasets would enable the study of dynamic bankruptcy risk and strengthen early-warning applications by capturing the temporal accumulation of distress. Second, the approach should be validated on larger and more heterogeneous datasets to assess robustness, scalability, and cross-context generalization. Third, integrating NN2Poly with complementary explanation methods such as SHAP could enrich interpretability by combining global, algebraic representations with firm-level attributions that support case-specific assessment. Finally, methodological advances that allow NN2Poly to scale to deeper neural architectures would broaden its applicability to more complex financial prediction settings while preserving the objective of transparent, model-consistent explanations.

Appendix

Table 5.1: Description of financial variables used in the analysis

Variable	Financial interpretation	Formula
ROA_net_profit_total_assets	Return on assets, measuring profitability relative to total assets	Net Profit / Total Assets
Total_liabilities_total_assets	Leverage ratio indicating financial risk	Total Liabilities / Total Assets
Working_capital_total_assets	Liquidity position relative to firm size	Working Capital / Total Assets
Current_assets_short_term_liabilities	Short-term liquidity (current ratio)	Current Assets / Short-term Liabilities
Cash_flow_adequacy_ratio	Ability of operating cash flow to cover obligations	Operating Cash Flow / Total Liabilities
Retained_earnings_total_assets	Cumulative profitability retained in the firm	Retained Earnings / Total Assets
EBIT_total_assets	Operating profitability before financing costs	EBIT / Total Assets
Equity_book_value_total_liabilities	Solvency measure comparing equity buffer to debt	Equity / Total Liabilities
Sales_total_assets	Asset turnover, measuring efficiency	Sales / Total Assets
Equity_total_assets	Capital structure indicator	Equity / Total Assets
Gross_profit_extra_items_fin_exp_total_assets	Profitability adjusted for extraordinary and financial items	Gross Profit / Total Assets
Gross_profit_short_term_liabilities	Short-term debt coverage by gross profit	Gross Profit / Short-term Liabilities
Gross_profit_depreciation_sales	Margin after depreciation	(Gross Profit – Depreciation) / Sales
Gross_profit_interest_total_assets	Profitability net of interest burden	(Gross Profit – Interest) / Total Assets

Variable	Financial interpretation	Formula
Liabilities_turnover_days	Average payment period to creditors	365 × Liabilities / Sales
Gross_profit_depreciation_-total_liabilities	Debt coverage after depreciation	(Gross Profit – Depreciation) / Total Liabilities
Total_assets_total_liabilities	Inverse leverage ratio	Total Assets / Total Liabilities
Gross_profit_total_assets	Overall profitability ratio	Gross Profit / Total Assets
Gross_profit_sales	Gross margin	Gross Profit / Sales
Inventory_turnover_days	Average inventory holding period	365 × Inventory / Sales
Sales_growth_rate	Year-over-year sales growth	$(\text{Sales}_t - \text{Sales}_{t-1}) / \text{Sales}_{t-1}$
Operating_profit_total_assets	Core operating profitability	Operating Profit / Total Assets
Net_profit_sales	Net margin	Net Profit / Sales
Gross_profit_3y_avg_total_-assets	Smoothed profitability over three years	$\text{Avg}(\text{Gross Profit}_{t-2:t})$ / Total Assets
Equity_minus_share_capital_total_assets	Retained equity intensity	(Equity – Share Capital) / Total Assets
Cash_flow_total_liabilities	Debt repayment capacity via cash flows	Cash Flow / Total Liabilities
Operating_profit_financial_-expenses	Interest coverage ratio	Operating Profit / Financial Expenses
Working_capital_fixed_assets	Long-term financial equilibrium	Working Capital / Fixed Assets
Log_total_assets	Firm size proxy	log(Total Assets)
Liabilities_excl_cash_sales	Non-cash leverage intensity	(Liabilities – Cash) / Sales
Gross_profit_interest_sales	Margin after interest costs	(Gross Profit – Interest) / Sales
Current_liabilities_-turnover_days	Short-term payment cycle	365 × Current Liabilities / Sales
Operating_expenses_short_-term_liabilities	Cost pressure on short-term debt	Operating Expenses / Short-term Liabilities
Operating_expenses_total_liabilities	Cost burden relative to total debt	Operating Expenses / Total Liabilities
Profit_on_sales_total_assets	Profitability scaled by asset base	Profit on Sales / Total Assets
Sales_fixed_assets	Fixed asset productivity	Sales / Fixed Assets
Quick_assets_long_term_liabilities	Long-term liquidity buffer	Quick Assets / Long-term Liabilities

Variable	Financial interpretation	Formula
Constant_capital_total_assets	Capital intensity indicator	Constant Capital / Total Assets
Profit_on_sales_sales	Operating margin	Profit on Sales / Sales
Cash_ratio	Immediate liquidity measure	Cash / Current Liabilities
Debt_repayment_period	Years needed to repay debt	Total Liabilities / Cash Flow
Operating_profit_sales	Operating margin	Operating Profit / Sales
Receivables_inventory_-turnover_days	Cash conversion cycle component	365 × (Receivables + Inventory) / Sales
Receivables_turnover_days	Average collection period	365 × Receivables / Sales
Net_profit_inventory	Inventory profitability	Net Profit / Inventory
Quick_ratio	Acid-test liquidity ratio	(Current Assets – Inventory) / Current Liabilities
EBITDA_total_assets	Cash-based profitability	EBITDA / Total Assets
EBITDA_sales	EBITDA margin	EBITDA / Sales
Short_term_liabilities_total_-assets	Short-term leverage intensity	Short-term Liabilities / Total Assets
Short_term_liabilities_-turnover_days	Short-term payment cycle	365 × Short-term Liabilities / Sales
Equity_fixed_assets	Capital adequacy for fixed investments	Equity / Fixed Assets
Working_capital	Net short-term financial position	Current Assets – Current Liabilities
Gross_margin	Core operating margin	Gross Profit / Sales
Liquidity_efficiency_ratio	Short-term solvency efficiency	Current Assets / Current Liabilities
Total_costs_total_sales	Cost intensity	Total Costs / Sales
Long_term_liabilities_equity	Long-term solvency indicator	Long-term Liabilities / Equity
Sales_inventory	Inventory efficiency	Sales / Inventory
Sales_receivables	Collection efficiency	Sales / Receivables
Sales_short_term_liabilities	Operating leverage vs short-term debt	Sales / Short-term Liabilities
Sales_fixed_assets	Fixed asset utilization	Sales / Fixed Assets

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que *ChatGPT* u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Nombre del Estudiante, estudiante de Programa de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Nombre del TFG", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa *ChatGPT* u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. Referencias: Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
2. Interpretador de código: Para realizar análisis de datos preliminares.
3. Estudios multidisciplinares: Para comprender perspectivas de otras comunidades sobre temas de naturaleza multidisciplinar.
4. Constructor de plantillas: Para diseñar formatos específicos para secciones del trabajo.
5. Corrector de estilo literario y de lenguaje: Para mejorar la calidad lingüística y estilística del texto.
6. Generador previo de diagramas de flujo y contenido: Para esbozar diagramas iniciales.
7. Sintetizador y divulgador de libros complicados: Para resumir y comprender literatura compleja.
8. Revisor: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
9. Traductor: Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado *ChatGPT* u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: Marzo 2026

Firma: Claudia Valverde Hueso

References

Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O., & Bilal, M. (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert systems with applications*, *94*, 164–184.

Allianz Trade. (2025). *Global insolvencies report 2025: The corporate battlefield.* Retrieved from `https://www.allianz-trade.com/` (Includes forecasts for 2025–2026.)

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, *23*(4), 589–609.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert systems with applications*, *83*, 405–417.

Beade, Á., Rodríguez, M., & Santos, J. (2024). Multiperiod bankruptcy prediction models with interpretable single models. *Computational Economics*, *64*(3), 1357–1390.

Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st international convention on information and communication technology, electronics and microelectronics (mipro)* (pp. 0210–0215).

Dubey, S. R., Singh, S. K., & Chaudhuri, B. B. (2022). Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, *503*, 92–108.

Informa D&B. (2021). *Demografía empresarial 2020: Concursos y disoluciones.* Madrid. (Informe anual de concursos empresariales en España.)

Informa D&B. (2022). *Concursos y disoluciones 2021.* Madrid. (Informe anual de concursos empresariales en España.)

Informa D&B. (2023). *Concursabilidad 2022.* Madrid. (Informe anual sobre concursos de acreedores en España.)

Informa D&B. (2024). *Concursabilidad 2023.* Madrid. (Informe anual sobre concursos de acreedores en España.)

Informa D&B. (2025). *Concursabilidad 2024.* Madrid. (Informe anual sobre concursos de acreedores en España.)

Kaspersen, A., & Lindemark, O. (2022). *Interpretable deep learning for bankruptcy prediction* (Unpublished master's thesis). NTNU.

Liu, J., Li, C., Ouyang, P., Liu, J., & Wu, C. (2023). Interpreting the prediction results of the tree-based gradient boosting models for financial distress prediction with an

explainable machine learning approach. *Journal of Forecasting*, *42*(5), 1112–1137.

Maclaurin, D., Duvenaud, D., & Adams, R. (2015). Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning* (pp. 2113–2122).

Mihalovic, M. (2016). Performance comparison of multiple discriminant analysis and logit models in bankruptcy prediction. *Economics & Sociology*, *9*(4), 101.

Moen, P. A. (2020). *Bankruptcy prediction for norwegian enterprises using interpretable machine learning models with a novel timeseries problem formulation* (Unpublished master's thesis). NTNU.

Morala, P., Cifuentes, J. A., Lillo, R. E., & Ucar, I. (2021). Towards a mathematical framework to inform neural network modelling via polynomial regression. *Neural Networks*, *142*, 57–72.

Morala, P., Cifuentes, J. A., Lillo, R. E., & Ucar, I. (2023). Nn2poly: A polynomial representation for deep feed-forward artificial neural networks. *IEEE Transactions on Neural Networks and Learning Systems*.

Morala, P., Cifuentes, J. A., Lillo, R. E., & Ucar, I. (2025). Nn2poly: A polynomial representation for deep feed-forward artificial neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(1), 781-795. doi: 10.1109/TNNLS.2023.3330328

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, 109–131.

Park, M. S., Son, H., Hyun, C., & Hwang, H. J. (2021). Explainability of machine learning models for bankruptcy prediction. *Ieee Access*, *9*, 124887–124899.

Qu, Y., Quan, P., Lei, M., & Shi, Y. (2019). Review of bankruptcy prediction using machine learning and deep learning techniques. *Procedia Computer Science*, *162*, 895–899.

Shetty, S., Musa, M., & Brédart, X. (2022). Bankruptcy prediction using machine learning techniques. *Journal of Risk and Financial Management*, *15*(1), 35.

Solunion. (2025). *Insolvencias empresariales en españa: Informe trimestral 1t 2025.* Madrid. Retrieved from `https://www.solunion.es/` (Informe de abril de 2025 sobre la evolución de los concursos empresariales en España.)

sowide. (2023). *U.s. corporate bankruptcy dataset.* `https://github.com/sowide/bankruptcy_dataset`. (GitHub repository)

Valverde Hueso, C. (2025). *Nn2poly bankruptcy prediction.* `https://github.com/claudiavalverdehueso-tfg/NN2Poly-Bankruptcy-Prediction`. (GitHub repository)

Yeh, I.-C. (2010). *Taiwanese bankruptcy prediction dataset.* `https://archive.ics.uci.edu/dataset/572/taiwanese+bankruptcy+prediction`. (UCI Machine Learning Repository)

Zhang, G. P. (2004). *Neural networks in business forecasting*. IGI global.

Zięba, M., Tomczak, S. K., & Tomczak, J. M. (2016). *Polish companies bankruptcy dataset.* `https://archive.ics.uci.edu/dataset/365/polish+companies+bankruptcy+data`. (UCI Machine Learning Repository)