University of Reading

Innovations in Credit Risk Assessment of Small and Medium Enterprises

Bogdan Pleshkevich

Henley Business School

This thesis is submitted for the degree of Doctor of Philosophy

Initial Submission: September 2025

Final Submission: May 2026

Declaration

According to the Rules for Submission of Theses for Research Degrees, Section 8(a), the candidate is required to include in each copy of the thesis (including the electronic copy), a signed declaration of original authorship. I hereby declare my contribution to a co-authored publication that will be included in my PhD dissertation. Regarding the following publication:

For the publication

Pleshkevich, B., & Han, L. (2025). Application of Econometric Techniques and Machine Learning Algorithms for Credit Risk Assessment of Small and Medium-sized Enterprises (SMEs). In *HANDBOOK OF FINANCIAL ECONOMETRICS, STATISTICS, TECHNOLOGY, AND RISK MANAGEMENT: (In 4 Volumes)* (pp. 1963-1990).

which is included into the following section:

Chapter 1. Introduction and

Chapter 2. Review of Econometric Techniques and Machine Learning Algorithms for Small & Medium-sized Enterprises (SME) Credit Risk Assessment

Contribution of Bogdan Pleshkevich is 75%

Acknowledged by: Bogdan Pleshkevich. Date: 30/09/2025

Acknowledgements

It’s been a long path since the time I decided to work on my research ideas up to this moment. I am very grateful for opportunities I was granted and supported I received along the way. First and foremost, I want to share gratitude to my parents, whose unwavering support, encouragement, and sacrifices have been the foundation of all my achievements. Their belief in me has been a constant source of strength and motivation throughout this journey.

Further, I would like to extend my heartfelt appreciation to my supervisor, Dr. Liang Han, whose invaluable guidance, constructive feedback, and constant encouragement have been instrumental in shaping this work. His expertise has not only contributed greatly to the quality of this thesis but have also inspired me to strive for excellence in my future endeavours.

I am also sincerely thankful to my friends, who stood by me with kindness, patience, and encouragement during all those years.

Abstract

The credit risk assessment of small and medium enterprises (SMEs), as a separate class, has been specifically segmented due to their high dependence on external finance as well as critical importance for global economies. This thesis deepens understanding of tools and methods that can be utilized for credit risk modelling to reveal various new patterns in lending to SMEs with three essays.

First essay summarizes the key developments in the SME credit risk modelling considering the modern estimation techniques. Through a literature review and an empirical study, it outlines the historical evolution of the SME credit risk evaluation and presents current trends to discuss recent contributions from utilization of advanced techniques for SME scoring such as higher performance and robustness.

Second essay proceeds by investigating obstacles for quantitative assessment. By reviewing various definitions of credit risk events in application to multinational SME loans, this part demonstrates how utilization of comprehensive robustness specifications promotes capturing similarities across European countries through empirical study on Spanish and Italian SMEs.

Finally, third expands the discussion by looking into Generative AI (GenAI) applications and their great potential to improve credit risk assessments for SMEs in Europe. Assessment of new capabilities of GenAI tools with regulatory and ethical considerations are used to propose a quantitative credit risk framework to go beyond standard financial and

transactional data for their implementation by policy makers and financial institutions.

The important contributions of this thesis include theoretical frameworks to evaluate the hardly observable, small firms with higher precision through different data utilization (granular credit event definitions) and data augmentation (collecting and transforming unstructured data). Empirical findings provide actual benefits for multi-region, cross-events evaluation, demonstrating how the focus on robustness can translate into higher confidence outcomes, such as importance of early warning signals, non-financial information, and regional effects.

Table of Contents

Table of Figures

Table of Tables

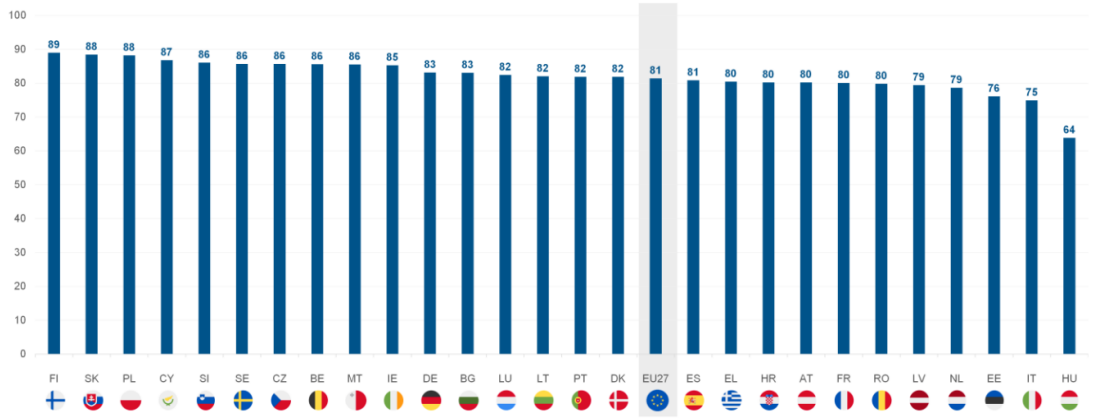
1 Introduction

1.1 Importance of the SME sector and key challenges for quantitative risk assessment

Small and medium-sized enterprise (SME) sector is a key contributor of various economies worldwide. Covering various economic and social needs, SMEs are involved in job creation, innovation transmitting, and other activities. Statistically, in European Union for example, SMEs generate more than half of value added, represent more than 99% of non-financial organisations and secure jobs for 85 million people equivalent to 64.4 % of total employment (Schulze Brock et al., 2025).

Unsurprisingly, from financial perspective most SMEs are dependent on external funding, as evidenced by Figure 1. About 82% of the firms rely on external funding including overdrafts (48%), leasing (47%) and banking loans (43%) as top relevant forms of financing (European Central Bank, 2026). In the light of recent events, such as the COVID pandemic, the current state of the world is a natural example of how hazards can significantly affect the usual business routine and decrease the performance. The risk that corporate and individual liabilities will not be met due to various reasons or factors is non-negligible and relevant assessment is essential for lenders, creditors, and financial regulators.

Figure 1. Relevance of debt financing per Survey on the access to finance of enterprises (SAFE), Source: European Central Bank, 2026



Several factors catalysed a renewed focus on SME credit risk modelling in the 2000s. The implementation of the Basel II Capital Accord (2004) required banks to develop internal rating systems for SMEs, pressuring lenders to improve their models. Moreover, economic crises highlighted weaknesses in existing models – during the 2007-2009 Global Financial Crisis, rising SME default rates and tightened credit underscored the need for better predictive tools. Scholars noted that traditional models based on financial statements were too static, failing to reflect rapid changes in SME finances or macroeconomic shocks. Research showed that applying large-firm models to SMEs yielded low prediction accuracy due to SMEs' unique financial characteristics (Ciampi, 2021).

One of key differentiation from larger firms is the informational characteristic of smaller firms. SMEs are usually operating under less reporting scrutiny, thus less focused on generating audited and detailed financial statements or publicly available information. This results into information asymmetry during the lending process where banks and other financial

institutions cannot verify and confidently assess creditworthiness of SMEs, causing additional uncertainty for decision making (Beck, Demirgüç-Kunt, & Maksimovic, 2008). As a result, incomplete or unreliable data creates challenges for financial institutions to develop and maintain quantitative models.The quantitative analysis of SME credit risk is essential due to its direct involvement is acquiring additional capital for small private firms. Without robust models, the access to finance can be restricted or imprecise, creating additional constraints for SME sector and, potentially, whole economy (as described in Calabrese et al., 2021). At the same time, the application of credit scoring models requires tedious efforts to account for all potential risk patterns realization (Chiampi, 2021), as if models are not robust and accurate, they can cause higher credit losses for lender, which will translate in higher costs of lending for borrowers and imposing additional caveats like collateral and tighter requirements.

These constraints are more pronounced due to biases in financial institution policies (the underwriting procedures can be more scrutinized, or acceptance criteria could be stricter when compared to larger counterparties, Garcia-Martinez et al., 2023) which may result in even higher capital costs for SMEs or limited access to financing opportunities (De Blick et al., 2024). Additionally, in the context of the guarantee or support programs, credit risk evaluation, creditworthiness and its connection with financial constraints should be considered by such program providers during the design as well as outcome evaluation (Kersten et al., 2017). Therefore, expanding understanding of SME credit risk evaluation has vast implications for financing and policy making.

Correct and precise evaluation of credit risk of SMEs through building and utilising relevant models is a critical activity for banks, credit bureaus and other market participants. While the external financing is crucial for SMEs at all stages of their lifecycle, government policies might ease access to funds or to help with loans securing[1]. In the meantime, the SME credit risks might transmit to a wider range of agents through loan guarantees, SME re-financing, factoring, and supply chain. To sum up contributions, if SMEs are playing the role of the “backbone” in an economy, the credit risk assessment of SME defines the quality and healthiness of the financial system.

The empirical research on SME credit risk is considerably less developed relatively to corporate credit risk of larger corporations and public companies. Thus, improving robustness and reliability of SME models represents important objective both for researchers and practitioners. These improvements can be constituted through adopting better statistical methods, exploring better credit risk measurements, and improving data collection techniques to reduce uncertainty. In particular, the definition of financial distress may significantly influence empirical results, raising concerns about the robustness of model predictions.

1.2 Problem statement

Modelling SME credit risk requires multiple challenges to be considered. Firstly, a substantial piece of corporate-oriented literature utilized bankruptcy status as the main default indicator. Indeed, bankruptcy-identified

[1] Analysis of a finance gap reduction has been presented, for instance, in Tucker & Lean (2003).

(or other reasons) flags are easier to be tracked as (i) they might be available from financial reports; and (ii) they have clear connection with liquidity and solvency principles which are key indicators for repayment risk formulation. However, the legal status of bankruptcy is not necessarily observed, even if firm experienced financial distress (Balcaen & Ooghe, 2006). With Basel Committee on Banking Supervision release of capital framework (proposed in 2001, published II accord in 2006), it has boosted research oriented towards default definition and credit risk realization (Hayden, 2003; Lin, Ansell, and Andreeva, 2012). The choice of distress definition in the SME context remains important for empirical modelling and need to be resolved, especially when applied to risk management and lending decisions (Campbell, Hilscher, & Szilagyi, 2008).

Secondly, empirical literature on measuring SME credit risk has been historically limited by the quality and amount of eligible datasets to work with. The key reason behind worse (when compared to public or larger counterparties) coverage is reluctance of SME obligations to submit financial reporting which leads to data scarcity and general opaqueness. This has affected the adoption of International Financial Reporting Standards (IFRS) for SMEs when multiple countries experienced issues with transition of their policies due to complexity of implementation leading to inconsistencies with reporting frameworks (Fearnley & Hines, 2007; Perera & Chand, 2015). Additionally, the data is usually collected on annual or biannual frequency which might be insufficient for capturing behavioural patterns. As a result, the majority policy-oriented research is mostly restricted to very specific regions or data-dependent designs.

The issue of getting micro-level data on SME has a direct impact on amount and versatility of research on the topic. For example, only few studies perform assessment of multiple countries like Dietsch and Petey (2004) and Filipe et al. (2016). To address this, there are few initiatives that aim to improve the data availability and SME sector transparency. The data collection is definitive for model construction and decision maker, thus higher information disclosure would reduce information asymmetry and reduce opaqueness penalty on SME, as described, for instance, in Song et al. (2016). Example of such initiative is AnaCredit that was initiated in 2011 by ECB.

The project is oriented on collecting detailed information on individual bank loans in euro area. Prior to the project, the ECB has only partial information for SME sector which has been obtained from different surveys. The data structure of new data will allow providing extensive credit risk analysis. The obvious limitation of AnaCredit is its availability in a short-term perspective as data collection had only started in 2018. Another constrain is data access: due to confidentiality reasons, the information is only available to certain participants and ECB with some users will only have access to part of the dataset (ECB, 2016).

Other sources of company-level data are vendors focusing on publishing private company data. They collect data pieces from multiple providers and aggregate information in one database. Example of such company and database can be Bureau van Dijk with Orbis database. The latter has been widely used (Filipe et al.,2016; Calabrese et al., 2013) as data for SME research. While such solutions offer highly dimensional and granular data, the quality and data structure require extra efforts to clean (Kalemli-Ozcan,

Sebnem, et al., 2015). In addition, access to the database might require a substantial fee which makes it more difficult to use for independent parties. As a result, while data aspect of SME landscape is improving, its perception remains to be limited and obscures potential research opportunities due to those constraints.

Another challenging aspect of SME portfolios is their ambiguous position between corporate and consumer credit exposures bringing additional uncertainty in respect to risk definition and modelling approach (Lin, Ansell, and Andreeva, 2012). Like larger corporates, SMEs ability to repay is dependent on financial performance of business activities which imply that revenue, profit and equity value will determine the success of liability maintenance. At the same time, smaller firm credit risk could experience idiosyncratic disturbances associated with non-financial factors (Altman, and Sabato, 2010), being driven by behavioural and individual characteristics and nuances of the particular entity. These factors are more typical for retail portfolios risk assessment. The comparison of the SME credit risk nature with other classes as well as intra-SME segregation into micro, small and medium subgroups is quintessential for most studies, in one way or another.

1.3 Synthesis of key literature

The academic literature relevant to SME credit risk modelling can broadly be grouped into several research streams. The first stream focuses on the development of statistical models for predicting bankruptcy and financial distress. Early contributions include Altman's (1968) Z-score model and Ohlson's (1980) logistic regression framework, which demonstrated the predictive power of financial ratios derived from accounting statements.

Subsequent research has expanded these approaches by incorporating market-based variables, survival analysis techniques, and hazard models (Shumway, 2001; Duan et al., 2018).

A second body of literature examines the specific characteristics of SME credit risk with studies adapting traditional bankruptcy prediction models to the SME context. As notable example, Altman and Sabato (2007) developed a model specifically tailored to SMEs using financial ratios that better capture the financial structures of smaller firms, with wider expansion of accounting ratios and larger dataset in Altman and Sabato (2010). Similarly, Dietsch and Petey (2004) an investigated determinants of SME default across European countries, highlighting the importance of firm size, leverage, profitability, and liquidity indicators. Generally, differentiation of SMEs and underlying risk factors, especially qualitative and soft information, is being brought as key discussion points within this stream.

A third stream of research focuses on the role of information asymmetry in credit markets. Theoretical work by Stiglitz and Weiss (1981) demonstrated how imperfect information between lenders and borrowers can lead to credit rationing, even when borrowers are willing to pay higher interest rates. In the SME context, informational opacity increases the difficulty of borrower screening and risk assessment. Berger and Udell (1998) emphasized that relationship lending and soft information may help mitigate these challenges, although quantitative risk models remain an important component of credit evaluation processes.

Finally, recent literature has explored the application of more advanced modelling techniques, including machine learning methods and

hybrid modelling frameworks, to improve credit risk prediction. While these approaches often demonstrate strong predictive performance, relatively limited attention has been devoted to evaluating their robustness under alternative definitions of financial distress. This gap is particularly relevant for SME datasets where measurement uncertainty may be more pronounced.

1.4 Theoretical frameworks

The theoretical background for the present study relies on two key frameworks – stakeholder theory and agency theory – to formulate research objectives and contribution. From perspective of agency theory, the information asymmetry creates multiple challenges arise since lenders cannot confidently assess SME borrowers, which hold significantly more information about their operation than lenders. Such imbalance can lead to adverse selection and moral hazard problems, ultimately affecting credit allocation and loan pricing (Akerlof, 1970; Stiglitz & Weiss, 1981). One of such consequences is credit rationing when instead of charging higher rates (for potential high-risk borrowers), lenders prefer to cut supply of lending and reject SME applicants to avoid high-risk borrowers and maintain proven part of the portfolio. Such pattern is particularly present in turbulent times or deteriorating phase of credit cycle.

In the presence of data and information gaps, financial institutions often rely on predicted or proxied outputs. While predictive accuracy is important, the theory implicitly highlights the importance of model robustness in credit decision-making. Maintaining stable performance across different time periods and economic conditions reduces uncertainty surrounding confidence in risk estimates. This is particularly relevant in SME lending, where data is

often subject to structural change. Empirical evidence shows that instability in risk models can lead to procyclical lending behaviour, with banks tightening credit excessively during downturns due to lack of confidence in model outputs (Bernanke, Gertler, & Gilchrist, 1999; Jiménez & Saurina, 2006). Therefore, robustness enhances the reliability of risk differentiation, which is essential for mitigating precautionary credit rationing. This aligns with broader financial stability literature emphasizing the importance of model reliability and resilience, particularly under stress scenarios (BCBS, 2017; Danielsson et al., 2018).

At the same time, stakeholder theory extends SME credit risk analysis beyond the lender-borrower relationship to a broader network of agents whose interactions shape firm outcomes. Building on R. Edward Freeman (1984), the theory argues that financial distress reflects not only contractual default but also disruptions across stakeholders such as suppliers, employees, and customers. In the context of SMEs, these relationships are even more as risk signals often emerge outside traditional financial performance indicators. Empirical evidence shows that governance quality, ESG exposure, and stakeholder relationships contain additional information for the financial risk (Hörisch et al., 2014; Boubaker et al., 2020). On the other hand, various stakeholders might be focusing on different dimensions of the credit risk (e.g., overall solvency, maintaining mid- and short-term loans repayment), implying that credit risk assessment is required to be strong universally across those aspects.

From this perspective, robust modelling becomes essential since different definitions of “distress” lead to variability in observed credit risk outcomes. Models calibrated on a single default definition may therefore be

unstable or incomplete. Robustness across alternative definitions and information sets should be better aligned with stakeholder theory, as they capture risk consistently across heterogeneous perspectives. Thus, robustness is not only a technical consideration to test model validity but a theoretically grounded requirement for reflecting the multi-faceted nature of the SME credit risk.

1.5 Identification of research gaps

Despite the substantial progress made in the literature on corporate and SME default prediction, much of the existing research focuses on improving predictive accuracy through the introduction of new variables or modelling techniques. While these contributions are valuable, relatively fewer studies systematically examine the robustness of credit risk models as an independent characteristic, which is important as highlighted per theoretical concept above. This point is amplified when different definitions of financial distress are considered. As a result, it remains unclear to what extent empirical findings depend on the specific measurement choices used in model estimation. From a practical perspective, financial institutions require models that provide stable and reliable predictions across different data environments. However, if model performance is highly sensitive to the definition of credit risk, the practical applicability of such models may be reduced. Addressing this gap requires a systematic examination of how alternative distress definitions affect the predictive performance and stability of SME credit risk models.

Additionally, geographical versatility remains limited as empirical evidence on SME credit risk is provided unevenly across countries. While several studies have analysed SME datasets in specific national contexts,

comparative evidence remains limited for certain European economies characterized by high SME prevalence and strong reliance on bank financing. Additionally, due to data constraints and specifics of the SME portfolios, only few studies provided detailed multi-country analysis.

Another empirical gap is situated around inclusion of qualitative information to the assessment. While soft information can provide better understanding of driving forces behind risk patterns, it might be challenging to provide a systematic and scalable mechanism of processing and accounting of such information. While artificial intelligence is being widely used for operational efficiency and recent development of Large Language Models (LLMs) suggests a technical path to work with textual data, there is an absence of a well-established methodology to apply language models into credit risk frameworks, that would provide sufficient clarity on the outcome structure.

1.6 Research Aims

To address these gaps, this thesis formulates several research questions to be answered. Those questions can be split in two dimensions – methodological and empirical. From methodological standpoint, it is essential to understand how SME credit risk assessment can be improved from methodological view in the light of increasing data availability, and specifically:

M1. What methodological limitations exist in traditional SME credit risk models, and how can machine-learning and Artificial Neural Networks address these limitations?

M2. How effective are advanced machine-learning techniques for credit risk assessment of SMEs compared to conventional econometric approaches?

M3. Which factors and methods contribute the most to the robustness property of credit risk events identification?

M4. Which aspects need to be accounted for in a systematic LLM framework to incorporate unstructured SME-level information into credit risk models?

M5. When the use of a LLM framework is justified and how different are the outcomes when compared to more conventional modelling?

Micro, small and medium firms could demonstrate different levels of credit risk due to variation in repayment nature. Indeed, individual entrepreneurs are usually attached to retail profiles while larger organisations could be more associated with corporate processes and performance measures. As a result, we would observe distinctive SME-size effects. At the same time, while those effects are strongly (or rather directly) connected with financial performance, the smaller firms are not just their balance book. The financials are not being updated frequent enough to reflect idiosyncratic shock that cause repayment disturbance. Such idiosyncrasy can be driven by (or at least correlated with) those non-financials factors. To account for those, the empirical side is focused on robustness and its role in multi-country SME credit risk assessment, by addressing the following list of questions:

E1. How well can alternative definitions of SME credit deterioration be empirically predicted using granular risk factors in across-country setup?

E2. Do non-financial risk factors contribute differently under various credit event definitions in SME credit models?

E3. Does pooling SME data across countries improve statistical robustness and inference when modelling low-frequency credit events?

E4. Does increased granularity of explanatory information enhance portfolio-level risk understanding and robustness for SMEs?

E5. How does credit risk behaviour differ between micro, small, and medium enterprises?

1.7 Thesis contribution

Small and medium enterprises credit assessment can benefit from better mechanisms that can deepen understanding of specific patterns and phenomena. There are 2 important ways how the contribution of this thesis can be summarized.

Methodologically, the study provides a broad review of existing methodologies, highlighting various innovative approaches, testing the novel and advanced techniques, and commenting on applicability and benefits of machine learning techniques application for SME obligors. The key takeaway from applying alternative techniques is connected to robustness which can be positioned as additional dimension of risk assessment rather than just a validity check – while in-sample performance can be high, model stability and applicability over different samples contribute to the picture separately.

While robustness plays huge role throughout the thesis, enabling qualitative information in quantitative assessment is another methodological improvement. In that respect, the Chapter 4 proposes a novel way of credit risk assessment with GenAI approach of data collection and transformation that can substantially increase number of potential explanatory factors in credit risk of

less observable firms. Data augmentation and translation of unstructured data into metrics that can be later observed and back-tested is key element of the proposed framework which complemented with exact specification setups such as a classic and Bayesian regime-switching regression analysis. The proposed approach provides a conceptual way of mitigating data collection and processing challenge when assessing the SME credit risk. While empirical study is yet to be commenced, the outlined principles provide a fundamental basis for analytical work.

From an empirical perspective, this study extends the literature on credit risk modelling for SMEs by explicitly incorporating granular definitions of default and risk increase, including early warning signal events. By applying alternative credit event specifications to a consistent set of explanatory variables, the analysis evaluates how the level of granularity in default definitions affects predictive power and model stability. The results indicate that more granular, EWS-oriented definition has its benefits when compared to classical default definitions. Furthermore, incorporating such granular event definitions enables a more detailed decomposition of portfolio risk structure, supporting enhanced monitoring and more pronounced effects. Consequently, this study argues that systematic testing of multiple, granular credit risk definitions should become standard practice, rather than reliance on a single default configuration. As a result, it leads to materially stronger reliability, stability, and interpretability of empirical findings in SME credit risk modelling.

Empirical contribution also lays clues for multi-region studies, providing joint empirical findings for Spain and Italy. The study demonstrates that pooled data can be beneficial for **robustness** of findings, when the low

frequency events (such as defaults) are used as target variable. Study confirmed financial patterns and provided extra observations for SME pool of both countries. It also suggests that cross-regional studies (e.g., pan-European) can be of better confidence than single country studies that are common in the literature. While expanding the research with more focused comparison between large corporates and SMEs would provide better clarity on specifics of the latter, the empirical part supports the importance of qualitative information for smaller firms to support their distinctive position.

Such development has positive consequences for the business environment as well as the research state of the field. The pattern analysis under stress benefits from higher accuracy of negative effects and demand for financial support. The latter should be specifically important for SME sectors of the EU and the USA given latest global pandemic crisis (Kalemli-Ozcan et al., 2020). Subsequently, application of described methodologies and revealing its usefulness is also in the interest of policy makers and regional regulatory organizations.

1.8 Analytical framework and thesis structure

This research is structured around a unified analytical framework that decomposes the credit risk problem into three interrelated components:

(i) the definition of the outcome of interest,
(ii) the identification and structuring of its driving factors, and
(iii) the mechanism through which these factors are connected to the outcome.

This structure can be formalized using a standard functional representation:

$$y = f(x)$$

Within this formulation, y represents the target variable capturing the event of credit risk, x denotes the set of explanatory variables reflecting firm-specific, sectoral, and macroeconomic conditions. The function $f(.)$ defines the mapping between inputs and outcomes, reflecting both the modeling technique and the underlying assumptions about the data-generating process.

First, the connection between factors and outcomes (f) captures the methodological contribution of the research. Multiple modeling approaches are evaluated, including traditional statistical techniques and advanced machine learning methods, with a focus on robustness, stability, and generalizability across time and economic conditions. The functional form is not treated as fixed; instead, it is tested under different specifications to assess sensitivity to modeling assumptions and input definitions. This aspect is addressed through Chapter 2 and Chapter 4.

Second, the definition of the problem (y) is explicitly addressed by exploring alternative representations of credit risk events. Rather than relying on a single default definition, the research analyses various setups to robustness of risk evaluation. This is particularly relevant in the SME context, where heterogeneity and data limitations can lead to significant variation in observed outcomes. This aspect is addressed within Chapter 3.

Third, the factors of the problem (x) are systematically identified and categorized across multiple dimensions, including firm-level financial indicators, qualitative characteristics, industry dynamics, and macroeconomic conditions. Special emphasis is placed on expanding the information set

through the transformation of unstructured data into quantifiable features, thereby addressing information asymmetries and enhancing predictive power. This aspect is present in all subsequent Chapters.

The thesis continues in the following way. Chapter 2 provides existing literature review, discusses theoretical aspects that has been considered, introduces the data for the empirical part and tests of the most relevant approach to address questions M1 and M2 from section 1.6 above. Chapter 3 digs into definition of credit risk events and provides cross-definition, multi-regional with ultimate focus on findings robustness over various specifications and setups to reveal answers for questions M3 and E1-5. Chapter 4 discusses limitations as per question M4 and deepens the framework by looking into GenAI applications, regulatory considerations of associated data treatment and proposing combined framework of data augmentation and subsequent credit risk modelling. Finally, Chapter 5 concludes.

2 Review of Econometric Techniques and Machine Learning Algorithms for Small & Medium-sized Enterprises (SME) Credit Risk Assessment

2.1 Aspects of SME credit risk evaluation

The purpose of this chapter is to discuss phenomena of SME credit risk and provide critical overview of existing practices and research. Specifically, it focusses on identifying long-term trends and how machine learning and other modern modelling techniques contribute to developing modelling standards. Comparing robustness of machine learning model outcomes to more conventional analysis should help explain whether performance gain can sometimes be worth the loss in explainability.

2.1.1 Defining SME credit risk

The SME credit risk refers to the likelihood that a small or medium-sized enterprise will fail to meet its contractual obligations such as missing loan payment or breaching covenants, which altogether is widely known as default. However, interpretation of “failure” may vary, and exact definition can be expressed in a numerous ways depending on the assessment: purely qualitative (e.g., deterioration in quality, business operations disruptions), legal (e.g., classified status or bankruptcy), quantitative regulatory driven (like IFRS9 or CECL – 90 days past due, IASB 2014; BCBS, 2017), quantitative derived

(potential financial loss, financial indicators depreciation or any negative dynamics), and many others. It is indeed influenced by a wide range of factors, including the SME's financial health, business performance, industry conditions, and the general economic environment.

Who is particularly interested in credit risk assessment? Lenders must assess SME credit risk when making lending decisions to ensure that they are lending to creditworthy borrowers and to manage the risk of loan defaults. At the same time, SME credit risk is a key consideration not just for lenders but also for borrowers: poor risk management leads to lower availability and higher cost of credit for SMEs and can have significant implications for their ability to grow and succeed.

2.1.2 Quantitative models to estimate credit risk

2.1.2.1 Classical corporate credit risk methodologies

Based on the fundamental data that is being used for model construction, corporate credit risk modelling methodology could be divided into two group: structural approach and reduced-form methods. The structural models take their name from focusing on a relationship between default and capital structure of a firm. Concisely, the idea is to compare actual assets values against certain “default thresholds” (which are usually linked to value of debt). If assets fall below liabilities, a firm might experience default. However, direct application of such an idea leads only to contemporaneous recognition of default. The necessity to predict upcoming defaults brings complexity to the model. To evaluate equity value, the studies usually drive a la Merton (1974) model which links default events with security market information. Merton

assumes that the value of firm's assets could be presented as Brownian motion process with growth rate of μ and volatility σ to define distance-to-default T periods ahead as:

$$DD_T = \frac{\ln\left(\frac{V}{D}\right) + \left(\mu + \frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}} \quad (1)$$

The distance-to-default could be interpreted as the relative difference between the asset value of the firm to the default barrier given correction and normalization for the asset's volatility.

The classical study of Merton (1964) has been subsequently expanded with different alternatives, like setting the different default boundary exogenously (Longstaff and Schwartz, 1995; Huang and Huang, 2012) or endogenously (started with Black and Cox, 1976; extended through Leland and Toft, 1996). Huang and Huang (2012) encountered the phenomena that firms could continue to operate with negative net value which is not in line with the original model and suggested evaluating the default boundary relative to the debt value. At the same time, setting endogenous default barrier assumes that the threshold can be identified as a result of maximization of the equity value problem. Another widely known extension of Merton's model is the KMV model which is well described in Kealhofer (2003). The development connected to the fact that the KMV model builds the relationship between the company's equity and asset characteristics rather than linking debt to asset value and volatility.

As an alternative, the reduced-form group of studies commences with historical accounting information to evaluate which financial indicators drive the default pattern. The foundation for this direction was set by Altman (1968) with introduction of Z-score which is a weighted average of 5 key financial ratios:

working capital over total assets, retained earnings over total assets, earnings before interest and taxes over total assets, market value of equity over total liabilities, and sales over total assets. Almost a decade later, the Z-score model has been expanded into ZETA framework (Altman et al.,1977); however, the original setup serves as benchmark for large number of studies including Fedorova et al. (2013), Castagnolo and Ferro (2014), Altman and Sabato (2007).

On the downsides of the accounting-based reduced-form models is their reliance on the historical performance and underestimation of the future performance projection as well as volatility. The ignorance of those is driven by the difficulty of collecting more granular and more frequent observations of firm information. As a result, structural models that rely on market information usually demonstrate better performance (Hillegeist et al., 2004)

In contrast to extensive coverage of public and large private companies, the empirical literature on modelling credit risk for SMEs has been in a relative scarcity. At the same time, Dietsch and Petey (2004) provide sufficient evidence that modelling of SMEs default risk is different from modelling for larger companies. That leads to the fact that modelling of SME credit risk can and should be considered as a distinct stratum of the research field.

2.1.2.2 Classical SME Credit risk modelling

The very first effort in credit risk modelling of SME sector has been presented by Edmister (1972). The author applies multivariate discriminant analysis (MDA) which is based on financial ratios similar to Altman (1968). However, MDA approach has several limitations including strong assumptions

on predictors which, for instance, restricted the use of dummies or missing the relative importance of covariates (Ohlson, 1980). Fedorova et al. (2013) has demonstrated the better accuracy rates using classical regression analysis in a form of logit and probit models.

As one of the key studies for the field, Dietsch and Petey (2002) provide analysis of various drivers of credit risk for portfolios consisting of small firms' loans. Beyond the credit risk evaluation, the paper covers construction of loss distribution, capital requirements and pricing. Authors introduced the general concept of internal risk modelling for SME and potential implications of its usage in comparison to Internal Rating-Based (IRB) approach. Further expansion, Dietsch and Petey (2004) continue with direct comparison of retail and corporate exposures with SME. With estimating default probabilities and asset correlation researchers distinguish riskiness of different entities within 2 large European countries: Germany and France. Important finding of the paper is that SME are indeed separate strata and significantly different from large firms; moreover, there is strong evidence of size risk segregation *within* SME sector itself. At the same time, default probability distribution is assumed to be individually independent. From 2007-2008 we have learnt that modelling should account for inter- and cross-portfolio correlations. Second, there is limited robustness check in respect to default definition. As authors fairly mention, the different definition of default might impose deviating results.

One of the most practical ways to model conditional defaults is a mixture model. In such model default probabilities are expected to be dependent not only on a set of individual characteristics, but also on a list of common economic factors known as macroeconomic variables. Given this fact,

interdependence between defaults is modelled as sensitivity to those macroeconomic factors (Frey & McNeil, 2003). The importance of macroeconomic variables inclusion was also highlighted by Filipe et al. (2016) as they identify “significant regional variations”.

Another classical analysis of logistic regression prediction model development for SME is presented by Altman and Sabato (2007). They find that SMEs are more sensitive to idiosyncratic risk than to systemic risk as compared to large corporates. That fact confirms the findings of Dietsch and Petey (2004) regarding separation of SMEs and corporate model since SME-specific model demonstrates better results on SME data. Unfortunately, both studies do not employ qualitative information (including sector, firm age and other), although its importance is widely confirmed (e.g., Lehmann, 2003; Berger and Frame, 2007). Omitting fundamental qualitative information might be especially important in cases of low predictive power of financial ratios.

Going further, Altman and Sabbato (2010) evolve research ideas to account for larger set of various explanatory variables including non-financial information and “event” data for SME credit risk analysis. They use an enriched dataset of the UK firms for the period 2000-2007 which is supposed to be more current and covers a larger span of firms (5.8 million individual observations). The additional non-financial information reasonably increases model prediction accuracy. Altogether, the paper enhances confidence in modelling SME loan portfolio credit risk as separate cluster using both accounting and qualitative information.

Altman, Esentato, Sabato (2020) contribute to the SME credit risk assessment by analyzing specific case of Italy which introduced opportunity of

simplified bond issue for the unlisted firms in 2012. So-called “mini bonds” has created a separate market which required default probabilities estimation. The distinctive feature of the study is extended application analysis which covers both classical credit scoring and bond rating exercise and evaluates mini bonds which were issued through period 2012-2014. While the paper comprises a sequential set of on-purpose modelling steps with concrete business and scientific aims, the analysis is limited to Z-score credit risk model and omits robustness check which confines discussion and implication.

The contribution of Altman and Sabato studies (2007, 2010, 2020) to the field is extensive. First, they refine Altman’s Z-score approach to SME nature which grants a way to perform credit scoring for private firms analogous to the classical corporate credit risk models. Second, studies utilize different geographical and regulatory contexts to gauge credit risk of non-listed firms and underline the importance of building specialized models for those. At the same time, the focus is aligned to Z-score adaptation rather than addressing specificity of SMES and proposition of a radically novel approach (which should not be perceived as disadvantage). However, it highlights the necessity of more granular financial data collection and high dependence of modelling on high quality data. As a result, applicability is limited to well-reported cases only. In addition, while authors defend the point of SME credit risk modelling isolation from corporate/retail, the segmentation within SME is generally limited.

Fidrmuc and Hainz (2010) provide similar analysis on SME loans in Slovakia. They utilize the probit model for default risk estimation and find that financial ratios and qualitative regressors (in particular, set of sector dummies)

are significant determinants of default probability. Similar results for Ireland's SME sector are obtained by McCann and McIndoe-Calder (2012).

Duan et al. (2018) bring novelty to the quantitative aspect of credit risk modelling of private firms through implementing dynamic structure of the default risk for future horizons. It becomes possible via application of forward-intensity model on Korean dataset with monthly frequency. The study results in multiple findings which suggest importance of financial information disclosure for firms and superiority of forward-intensity model for term structure predictions when compared to classical model like logit, probit and Altman Z-score. Also, this approach reimagines the distance-to-default measure which is being proxied for the private firms and provides extra implication for practitioners.

The majority of described studies on SME credit risk assessment use either MDA or logit/probit models. Studies which employ different methods are usually based on hazard models (for instance, Belotti & Crook, 2013). The benefit of the hazard-type models is the existence of time component and subsequent ability of time-varying covariates inclusion. Shumway (2001) demonstrates the ease of model interpretability and proves that hazard model could be equated to a multi-period logit model.

2.1.2.3 Different angle: internationalization and innovation

In addition to the classical way of approaching SME credit risk, there are certain directions of research that extend analysis in considering more specific aspects of certain SME sectors. Angilella and Mazzu (2015) create a judgmental rating model using qualitative information to assess innovative SMEs credit ratings. They segment 4 types of risks – development, technological, market and production – through a set of hierarchically structured

criteria. While the methodology relies on soft information and requires accurate calibration, it presents a view on practical mechanisms which reduce uncertainty in the absence of formally reported data.

Gupta et al. (2014) questions the necessity to account for international trade in case of SME credit risk assessment. Following initial findings of Arslan and Karan (2009), the study builds hypothesis that the risk measure of domestic firms will have different sensitivity to factors than that of international SMEs with amount of trade export affecting default probability. While authors received mixed evidence from multiple logistic regression estimation results, they found new feature of the share of intangible assets to be a strong predictor of the credit risk.

2.2 Current trends in SME credit risk modelling

Meanwhile, there is a substantial development in credit scoring techniques as even a small improvement in performance of scoring grants banks higher profits (Hand & Henley, 1997). The standard approach – that is logistic regression – is widely studied by researchers (Crook, Edelman, & Thomas, 2007) and is usually referred as a benchmark for new models. However, this method has a strict limitation in the form of multiple assumptions to be satisfied for an efficient evaluation (Malley et al., 2012). In addition, modern automatic scoring systems that are based on machine learning or AI have demonstrated several advantages as compared to traditional credit scoring. Cost and time savings (Marques, Garcia, & Sanchez, 2013) and higher accuracy (Kruppa et al., 2013; Abdou et al., 2016) are among the most important improvements reached due to these approaches. The section below

summarizes key contributions, applications, and development in modern SME credit scoring literature.

2.2.1 Scoring: classification models and their variations

Development of credit scoring systems has become immensely popular in recent years. Since even small improvements in those models grant banks with high profits (Hand & Henley, 1997), credit scoring has attracted researchers from both academia and public sector. The accent in the literature from this field has moved from classical methods towards advanced machine learning algorithms.

One of the main points of current critique is related to the models to be designed without consideration of selection bias or data outliers (Calabrese et al., 2019). Thus, as a part of data preparation, loan-level characteristics could undergo classification or binning procedure. The process of coarse and fine classing to create bins is a quite standard technique for scorecard creation (Lin et al., 2012) as it allows modeler to focus on key patterns within one explanatory variable.

Such pre-modelling brings several improvements to the data structure prior to multivariate analysis. First, it eliminates impact of outliers as those are being grouped with the nearest bin and do not bias the overall model outcome without the need of separate cleaning. Second, observations with missing information could be treated as separate group which helps account for omitted values without ignoring them. Third, it controls for expected direction of pattern (ascending/descending for monotonic relationship, u-shape/hump-shape for quadratic one, and so on) and smooth unexpected/unexplained variation which

could be critical for continuous variables with large scale of potential values. If pattern is counterintuitive or binning solution doesn't exist, then it also serves as economic intuition check. To conclude, binning is expected to provide better shaped and tractable outcomes while accounting for classical data preparation problems.

To assess quality of binning for a particular variable, the predictive power can be evaluated through different metrics, for example, the most common indicators are Information value (IV) and Gini coefficient. As IVs rely on the calculation of individual bin weight-of-evidence, those are calculated as well. Weight-of-evidence (WoE) is a measure of relative separation of default from non-defaults expressed as log odds:

$$WoE_i = \ln\left(\frac{\left(\frac{Number\ of\ non-defaults\ in\ group\ i}{Number\ of\ non-defaults}\right)}{\left(\frac{Number\ of\ defaults\ in\ group\ i}{Number\ of\ defaults}\right)}\right) \quad (2)$$

Weight-of-evidence calculation is widely used in the scoring literature and used, for instance, in Calabrese et al. (2019). Overall, the binning process algorithm is extensively described in Siddiqi (2017) which iteratively defines cut-points by merging quantiles into non-overlapping groups with a target of IV maximization.

2.2.2 How machine learning and Artificial Neural Network (ANN) techniques fit into the scoring

As machine learning and AI become increasingly popular in the credit risk applications, the SME risk assessment has benefited from multiple contributions on the model techniques cross-comparison. The section below focuses on the most common and widely used methods.

At the same time, the ANN models are not always outperforming the classical methods. Bekhet and Eletter (2014) identified the logistic regression model to demonstrate better performance than the radial basis function model in terms of the overall accuracy rate (even the latter was superior in screening rejected applications).

2.2.2.1 Classification and Regression Trees (CART) and ANN

Methodology-focused studies within credit scoring field tend to compare several methods on a widely accessed, “golden” dataset. It might be a proposition of several credit systems within one experimental study when authors estimate multiple models on a specifically processed data piece. Abdou et al. (2016) compare the conventional model in the form of logistic regression to machine learning technique in a form of classification and regression tree (CART) and cascade correlation neural network (CCNN). Using area under receiver operating characteristic curve (AUC ROC), they conclude that CCNN demonstrates high accuracy and could be considered “to be of critical interest to bankers” (p.102).

However, obtaining a separate, high-quality dataset for the estimation is a challenge. Thus, a considerable number of studies have utilized UCI (University of California Irvine) Machine Learning repository datasets. For instance, Ala'raj and Abbod (2016) apply 5 ANN algorithms on Australian, German, Iranian, Polish, and Japanese information sets available through the UCI Machine Learning website. They propose a hybrid ensemble model which incorporates outcomes from all 5 scoring systems using a weighting approach. The weights are chosen according to uncertainty that each model creates. The method of creating a hybrid approach might be used for the purposes of this

research proposal. As a result, the hybrid model outperforms other hybrid approaches which are based on rankings and single classifiers.

From a general perspective, the use of the same datasets is controversial. On the one hand, it is an advantage since the outcome of a model could be easily compared to the results of predecessors. For instance, Zhao et al. (2015) link their results to 13 other studies performed on the same dataset from 2011 to 2013. In addition, those datasets are publicly available which makes them attainable at minimum cost and efforts. On the other hand, it limits the application diversity of the models. While the datasets used have been confirmed as reliable enough to be scored by ANN systems, the latest methods may be tailored for specifics of commonly used datasets and ignore or lower utilization of more recent or simply different datasets.

2.2.3 The benefit of ensembles

While each individual estimation technique has its pros and cons, the model ensemble combines multiple individual models to create an aggregated output. The advantage of using ensemble (especially for practitioners) is its ability to include a diverse set of alternative models (“challengers”) to create a joint prediction. This increases the output robustness. The most popular versions of ensemble models are boosting, bagging, and random forest.

Zhu et al. (2017) compare multiple machine learning techniques (decision tree, bagging, boosting, random subspace (RS), RS-boosting and multi-boosting) applied to a Chinese SME data from Shenzhen and Shanghai Stock Exchange. They include 18 financial and operational indicators to assess credit risk for reverse factoring. They observe a superior accuracy of RS- and

multi-boosting algorithms based on AUC and precision, recall and F-measure rates. The RS-boosting is highlighted as superior to multi-boosting; however, the observed difference is not substantial when compared to performance of techniques without ensemble. At the same time, the study is constructed only on 377 points, limiting any validation analysis.

Abellan and Castellano (2017) compare different classification techniques, including both statistical and AI techniques, under different ensemble approaches. Their main candidate – Credal Decision Tree (CDT) model – is subject to comparison with logistic regression, Support Vector Machine (SVM), C4.5 decision tree and Multilayer Perception (MLP) neural network. The main distinctive characteristic of CDT is the use of imprecise probabilities of default. The outcome of the experimental study by Abellan and Castellano (2017) characterizes CDT to deliver better performance in terms of accuracy and AUC measures. One of the strongest parts of their study is the use of different ensemble methods. Comparison has shown that under advanced ensemble schemes even logistic regression could produce high-accuracy outcomes. However, the focus of the paper is biased towards the diversity of used models, while analysis of CDT outcomes under different values of uncertainty parameter might be more illustrative and make a valuable contribution.

Fedorova et al. (2013) analyses the application of various techniques (MDA, logit-regression, CART, ANN and AdaBoost) on the dataset of Russian manufacturing companies. On first step, the selection is performed through MDA, logit and CART methods to identify sets of variables to predict bankruptcy with sufficient level of accuracy. Second step uses these sets for back-

propagation multi-layer and the radial basis function neural networks. The study achieves almost 89% of overall accuracy for the neural network models with 84.7% accuracy for the individual learner. While these indicators are sufficiently high and comparable with other studies, the classical specifications like Altman (1968) and Fulmer (1984) reached 77.5% and 82% respectively. Also, the study would benefit from out-of-sample validation to enrich findings.

2.2.4 Final notes on AI methods

While literature on credit scoring qualifies AI systems as an upgrade of traditional methods (Abellan and Castellano, 2017), such models and algorithms by construction are oriented on a classification problem and typically ignore the time component. The key advantage of this method is that it does not impose additional restrictions on data (e.g., continuous collection, controlling for time component). Second, those scoring models do not incorporate conditional distribution of classes. Finally, all applications are considered separately which has its criticism in the literature (Shumway, 2001).

2.3 Data description

2.3.1 Key details

The dataset for the empirical part of research has been extracted from European Data Warehouse (EDW) covering period from December 2012 to March 2019. Additional information has been later extracted to account for post-COVID period, expanding time coverage up to June 2023. EDW provides information on small and medium enterprises from 7 European countries which in our case comprise the following list of countries: Spain, Italy, Portugal,

Belgium, Netherlands, Germany, and France. To ensure data representativeness, the data has been checked for the following properties: (1) data coverage for financial information and performance (delinquency) information; (2) data continuity for a individual obligors (continuous performance history); (3) financial information reliability (values check). Based on data quality assessment, only Spain and Italy subsets have been chosen for further analysis due to lack of representativeness of other country's data[2].

Focusing on Italy and Spain is defensible as both economies represent large, bank-dependent SME systems characterized by reliance on relationship lending (Financial Stability Board, 2019; European Investment Bank, 2022). Evidence from the ECB SAFE survey further shows that SMEs in these countries depend more on bank financing than core Eurozone peers, reinforcing their relevance for analyzing credit supply dynamics and default risk (European Central Bank, 2025).

SME dataset is collected on quarterly basis which further being interpolated to monthly frequency. The key modelling data covers time period from 2013Q3 to 2019Q1. To highlight, the data is collected on the loan-level, meaning all information fields are specific for a particular loan (including financials of an associated firm); at the same time, we do not observe unique identifier for a firm, thus cannot say if any two loans belong to the same firm.

[2] Netherlands, Belgium and Portugal datasets lack financial performance information (such as delinquency) which is essential for research questions. Germany and France datasets have a large proportion of observations (85% and 55%, respectively) with non-amortized loans (bullet or revolving), thus cannot be assessed for repayment performance.

This distinguishes the study from other key literature in the field which focuses on firm-level outcomes. Such data structure allows us to analyze the data from perspective of a portfolio of loans which individually represent an asset with associated quality. Thus, the constructed risk model will reflect the probability with which asset (i.e., loan) will deteriorate in quality.

To investigate the proposed questions from multiple angles, the models are constructed to cover 3 types of explanatory factors:

1. loan-specific time-invariant characteristics.
2. loan-specific time-variant characteristics.
3. financial variables (for the associated firm).

The first group covers continuous variables and discrete factors. For continuous we work with loan terms in months, number of employees, company age in months. Factors are included as fixed effects and represent legal form of the associated firm, NACE industry group, amortization type, NUTS region group, and SME size class. Time-variant characteristics (group 2) intends to cover loan behavioral variables that might change over time. Those would include outstanding balance, interest rate, and loan progression rate. As a separate group we identify financial variables that would be relevant for the associated firm performance. The financials are time-invariant and assessed at application time; this information is reflected through leverage ratio, return on equity (RoE) ratio, short-term to total liabilities ratio, and debt to equity ratio.

The choice of covariates is motivated from Lin et al. (2012):

- Interest Rate: loan interest rate.
- Loan Balance: current exposure of remaining balance to be paid.

- Loan Term: loan duration, calculated as difference between loan maturity and loan origination dates in months.
- Number of Employees: reported number of employees at loan origination date.
- Loan Progression Ratio: ratio of difference between loan origination date and current observation date towards Loan Term.
- Return on Equity (ROE): ratio of firm's net profit towards equity.
- Short-term to Total Liabilities (ST-TL): ratio of firm's short-term liabilities from total liabilities.
- Debt to Equity (DE): a ratio of firm's debt to equity.
- Leverage (Leverage): a ratio of firm's debt to assets.
- Legal Form: a qualitative factor, defined as legal form of associated firm.
- NACE: a qualitative factor, industry classification group on a general scale (18 values).
- Amortization type: a qualitative factor, defined as type of loan amortization schedule (French or Linear).
- NUTS: a qualitative factor, defined as regional classification group on NUTS3 scale.
- SME class: a qualitative factor for fixed effects, defined as small, medium, or micro based on EC criteria:
  - Micro: 10 or less employees and turnover of 2 million EUR or less.
  - Small: 50 or less employees and turnover of 10 million EUR or less.
  - Medium: 250 or less employees and turnover of 50 million EUR or less.

Table 1. Data Coverage

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Interest Rate	4233339	3.932	2.119	0	2.221	5.278	19.5
Loan Balance	4233339	125050	798800	0	8282	66666	240000000
Loan Term	4233339	76.579	57.153	3	37	84	552
Number of Employees	4233339	8	19	1	1	10	250
Loan Progress	4233339	0.561	0.261	0	0.344	0.778	1
Return on Equity	1163765	0.894	3.289	0	0.145	0.786	277
Short-term to Total Liabilities	2281230	0.416	0.34	0	0.102	0.689	1
Leverage	2271136	0.636	0.308	0	0.414	0.909	1
Turnover	4233339	1495308	4052627	1	7479	1132711	50000000

2.3.2 Subset to evaluate binning and machine learning

For the Chapter 2, only Spanish SMEs data piece is being used for empirical analysis covering the time period from 2013 Q3 to 2019 Q1. The credit risk event is defined through 90 days past due delinquency on the payment (classical default flag). The single-country dataset has been used because machine-learning techniques are considerable better at capturing cross-country heterogeneity, thus findings might be as not representative for structural robustness assessment. Spain dataset contained larger number of default instances, thus were chosen to have better data coverage.

The dataset has been split into 3 pieces: train, test, and out-of-time (OOT) validation. The train and test datasets have been created as 70/30 split with controlling for NUTS and NACE for the timeframe 2013 Q3 – 2018 Q1. The OOT dataset has been split for testing model robustness on different time periods than modelling dataset and contains information about the latest year (Q2 2018 – Q1 2019). The combination of prediction outcomes on test and validation sets gives full picture on expected model performance and results robustness after initial estimation.

Table 2. Summary Statistics for train, test, and out-of-time samples

Train sample

Statistic	Mean	St. Dev.	Median	Min	Max
Interest Rate	4.56	1.74	4.62	0	15.00
Loan Balance	114,804	353,046	50,000	581	12,000,000
Loan Term	54.30	30.25	60.00	5.00	378.00
Number of Employees	24.01	28.27	10.00	1.00	250.00
Loan Progress	0.76	0.27	0.92	0.06	1.00
Return on Equity	0.47	2.17	0.23	0.00	196.00
Short-term to Total Liabilities	0.38	0.31	0.30	0	1.00
Debt to Equity	2.34	6.51	0.96	0.00	241.00
Leverage	0.49	0.24	0.49	0.00	1.00
Turnover	3,194,145	5,422,129	1,410,000	1,655	49,890,000
Default_flag	0.02	0.15	0	0	1.00

Test sample

Statistic	Mean	St. Dev.	Median	Min	Max
Interest Rate	4.57	1.72	4.61	0	14.00
Loan Balance	117,311	449,033	50,000	1,399	25,000,000
Loan Term	54.32	30.01	60.00	5.00	363.00
Number of Employees	24.89	30.25	10.00	1.00	250.00
Loan Progress	0.76	0.27	0.91	0.06	1.00
Return on Equity	0.45	1.44	0.23	0	79.00
Short-term to Total Liabilities	0.38	0.31	0.30	0	1.00
Debt to Equity	2.35	6.50	0.95	0.00	193.85
Leverage	0.49	0.24	0.49	0.00	0.99
Turnover	3,360,093	5,767,088	1,470,000	14,207	49,850,000
Default flag	0.02	0.14	0	0	1.00

Out-of-time sample

Statistic	Mean	St. Dev.	Median	Min	Max
Interest Rate	3.70	2.01	3.50	0	15.00
Loan Balance	160,854	511,056	60,000	1,270.00	22,000,000
Loan Term	72.36	46.57	60.00	7.00	366.00
Number of Employees	23.89	27.83	10.00	10.00	250.00
Loan Progress	0.79	0.20	0.82	0.12	1.00
Return on Equity	0.47	1.35	0.25	0.00	79.00
Short-term to Total Liabilities	0.34	0.30	0.30	0	1.00
Debt to Equity	2.47	10.11	1.00	0	701.00
Leverage	0.49	0.24	0.50	0.00	1.00
Turnover	3,237,259	5,601,517	1,400,000	10,000	49,720,000
Default flag	0.02	0.15	0	0	1.00

The calculation has been performed in RStudio 4.0.5 with reliance on base R functionality as well as classical libraries for data manipulation (data.table, tidyverse, lubridate), modelling data split and model estimation/ training (caret, fixest), evaluation (pROC), and reporting (stargazer). More detailed guidance on modelling in R could be inspired by Finch et al. (2019) and Wickham and Grolemund (2016).

2.3.3 Defining default

The dependent variable reflects the realization of credit risk event and, according to the described data, could be defined for each loan with a monthly frequency. Chapter 2 focuses on classical Basel II default event definition based on delinquency: 90+ days past due (as discussed in previous sections and can be also found in literature, e.g., Wagner, 2016). Formally, it can be formulated as:

$$DE_{it} = \begin{cases} 1, & if\ delinquency_{it} \geq 90 \\ 0, & otherwise \end{cases} \quad (3)$$

Important note for the default definition is that only primary default (first instance for each individual obligor) is considered, meaning recovery, curing and secondary defaults are excluded from the analysis. As a limitation, such approach creates a censorship bias since this truncates the observed time-at-risk and excludes cured exposures that may subsequently re-default, leading to downward-biased estimates of multi-period default probabilities and transition dynamics. However, the first-default approaches are often preferable due to isolation of the SME's initial transition into distress from posterior events by avoiding contamination from restructuring, cure, and re-default dynamics. It is important as those are highly dependent on lender practices, legal processes, and regulatory definitions rather than solely on underlying SME credit quality (EBA, 2016).

2.4 Empirical evidence

For the binning process, the rule of thumb of 0.02 IV is used (Siddiqi, 2006) for variable to be chosen. We follow the Siddiqi (2017) in terms of the algorithm setting; the number of quantiles is chosen to be 20; the benchmark outcomes with 10-quantile showed no drastic difference in general outcome and resulting IV values; confidence intervals for non-overlapping bins are set to 10%. The final model uses WoE for each variable as inputs; thus, negative sign of coefficients is expected to be observed by construction as direction of the impact will be reflected directly in the binning outcome. Since variables have different dispersion in the WoE values, the coefficient magnitude can't be directly interpreted as in the case of the logistic regression approach.

For the machine learning representation, XGBoost approach has been chosen. As an example of regularised gradient-boosted tree ensemble, it captures non-linearities and interaction effects and incorporates sparsity-aware split finding with learned default directions for missing values (Chen & Guestrin, 2016). In empirical credit risk studies, XGBoost has been reported to outperform logistic regression on predictive discrimination (Zhu et al., 2023), thus has its basis to demonstrate better accuracy for the SME credit risk as well.

2.4.1 Performance metrics

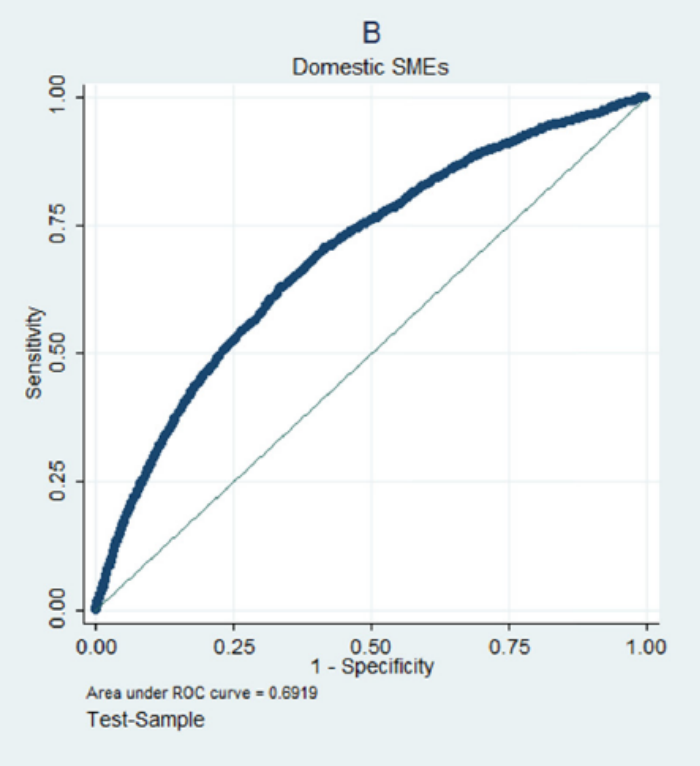
Measuring predictive performance is the key quality assessment of credit scoring models. The empirical studies have focused on an established set of performance metrics or tools. The first group focuses on overall classification accuracy. Given a cut-off value for probability of default, all outcomes from test dataset are classified in either defaulted or non-defaulted state. This set of parameters would include:

- Confusion Matrix: A breakdown of predictions into a table showing correct predictions (the diagonal) and the types of incorrect predictions made (what classes incorrect predictions were assigned). The classical examples of the confusion matrix derived analysis are Type I/Type II errors and Cohen's kappa.
- Precision: A measure of a classifier exactness, which is defined as share of correctly predicted events out of number of observations being predicted as “events”.

- Recall: A measure of a classifier completeness, which is defined as share of correctly predicted events out of number of all events in the dataset.
- F-score: A weighted average of precision and recall.
- Accuracy rate: Defined as share of correctly predicted events and non-events in the dataset.

Another approach to quantify performance of a model is to assess the correctness of ordering observed cases in terms of their riskiness. The commonly used tool for this type of analysis is the receiver operating characteristic (ROC) curve. The ROC curve plots true-positive rate against false-positive rate comparing model sensitivity and specificity. The advantage of using the ROC curve is its graphical illustration and direct link to the Gini coefficient and the Kolmogorov-Smirnov (KS) statistic. The KS statistic evaluates the difference between the events and non-events distributions at the optimal cut-off point and comparable to the Gini coefficient (Anderson, 2007). The determining cut-off point for default probability can be extended into a separate methodology discussion. Thus, for the empirical example only Gini evaluation is presented.

Figure 2. Receiver operating characteristic curve example (Gupta et al., 2014)



2.4.2 Logistic regression

As a part of logistic regression analysis, the performance of financial information could be tested through performance comparison between models with only general firm/loan factors and with both general and financial characteristics:

$$Y_{it} = L^{-1}[\alpha + \beta X_i + \gamma Z_{it}] \quad (4)$$

$$Y_{it} = L^{-1}[\alpha + \beta X_i + \gamma Z_{it} + \theta F_i] \quad (5)$$

The equation above, characterizes α intercept and fixed-effects terms for relevant qualitative factors, Y_{it} stands for the default flag, X_i stands for time-invariant factors (group 1) and Z_{it} stands for time-variant characteristics (group 2). Expanding the basic specification with financial data F_i brings the financial specification. For the second specification, as we know

both number of employees and turnover information, we can define an SME class as a qualitative factor based on European Commission criteria:

- Micro: 10 or less employees and turnover of 2 million EUR or less.
- Small: 50 or less employees and turnover of 10 million EUR or less.
- Medium: 250 or less employees and turnover of 50 million EUR or less.

2.4.3 First outcomes: adding financial information

The observed effects are universally in line with the expectations (Table 3) which are based on similar studies which employed same or equivalent drivers within their specifications. The list includes Calabrese et al. (2019), Lin et al. (2012), Altman and Sabato (2007) and Dietsch and Petey (2004). The outcomes of logistic regression evaluation are presented in Table 4.

Table 3. Expected effects for logistic regression analysis

Specification	Expected sign	Outcome	Significance
Interest Rate	Positive	Intuitive	***
Loan Balance	Positive	Intuitive	***
Loan Term in months	Positive	Intuitive	***
Number of Employees	Undefined	Negative	***
Loan Progression Ratio	Undefined	Strong Positive	***
Return on Equity Ratio	Negative	Intuitive	**
Short-term to Total Liabilities Ratio	Positive	Intuitive	
Debt to Equity Ratio	Positive	Intuitive	***

Table 4. Estimation results for logistic regression

	Default flag	
Logistic regression	**(1)**	**(2)**
Interest Rate	1.3110***	1.1460***
	(0.2066)	(0.2087)
Loan Balance	0.3750***	0.3759***
	(0.0266)	(0.0269)
Loan Term in months	0.8491***	0.7527***
	(0.1547)	(0.1570)
Number of Employees	-0.3428***	
	(0.1050)	
Loan Progression Ratio	4.2380***	4.2070***
	(0.5578)	(0.5634)
Return on Equity Ratio		-0.5898**
		(0.2357)
Short-term to Total Liabilities Ratio		0.4699
		(0.3017)
Debt to Equity Ratio		0.6254***
		(0.1044)
Observations	12878	12878
Squared Correlation	0.0789	0.0872
Pseudo R2	0.1916	0.2073
BIC	2918	2911
Fixed Effects		
ObligorLegalForm	✓	✓
NACE group	✓	✓
Amortization Type	✓	✓
NUTS	✓	✓

SME Size Category	✓
*** stands for 0.1% significance level ** stands for 1% significance level	

Both specifications are comparable in terms of magnitude and significance of observed effects. The positive, significant coefficient of interest rate reflects correct pricing of risk – ex-ante evaluation of banks is confirmed with repayment behaviour. The positive sign of loan balance is in line with expectations as larger loans usually imply larger financial burden and limit firm's agility and diversification capacity While longer term loans expose higher risk which might relate to repayment burden: the longer term is chosen to minimize the instalment amount, but also subject to higher uncertainty and may suffer from typical SME issues like weaker amortization discipline and vulnerability to credit cycle. The negative connection between default rate and number of employees brings a conclusion of negative relationship between size of the firm and probability of default. This is the first evidence within the thesis that size segmentation is present for intra-SME segment as posted in research questions E5. At the same time, strong positive relationship of the loan progression is a critical insight as it means that refinancing pressure is quite high for Spanish SMEs. Since it is loan-level indicator, this highlights importance of data granularity.

All financial variables effects are in line with overall business intuition. Negative sign of coefficient and its strong significance of return on equity justifies that even for smaller firms, the fundamental financial health is relevant.

At the same time the maturity structure of the liabilities has been marked as insignificant for the risk prediction. Theoretically, more short-term debt should lead to higher refinancing risk but in practice SMEs are heavily reliant on short-term financing and impact might be washed by low variability of the ratio. At the same time, debt-to-equity provides strong confirmation of capital structure importance. Overall, improvement from moving towards financials specification is observable but not substantial. However, predictive performance assessment requires testing performance not just on modelling dataset but also test and OOT validation samples.

2.4.4 Binning: bucketing improves robustness

To proceed with further analysis, the results from Gini evaluation are presented in Table 5. The Gini evaluation demonstrates the quality of each model's performance in terms of ranking within each sample. Values in the “Training” column represent in-sample capabilities, while test and out-of-time validation demonstrate predictive efficiency for forecasting purposes.

Table 5. Gini coefficients across all data samples

Specification	Training	Test	Out-of-time Validation
Logistic firm-/loan-only	0.6867	0.6665	0.3458
Logistic with financial	0.7118	0.6825	0.4035
Logistic after binning	0.7514	0.7409	0.7204
XGBoost	0.8750	0.7429	0.6542

Universally, in-sample Gini is higher than test Gini which, at the same time, outperforms OOT sample metric. However, the gap is not the same. For classical logistic regression, the difference between training and test is just few

points which could be explained by the fact that samples are not identical, and models are expected to have higher in-sample metric values. While the parameters are tailored to the training dataset, the test outcomes theoretically can outperform if the corresponding sample has better variation explainability through the chosen specification. At the same time, out-of-time sample has a considerably different default outline, which leads to significant drop in Gini. As a conclusion, the pure logistic specifications bring certain limitations in terms of time robustness.

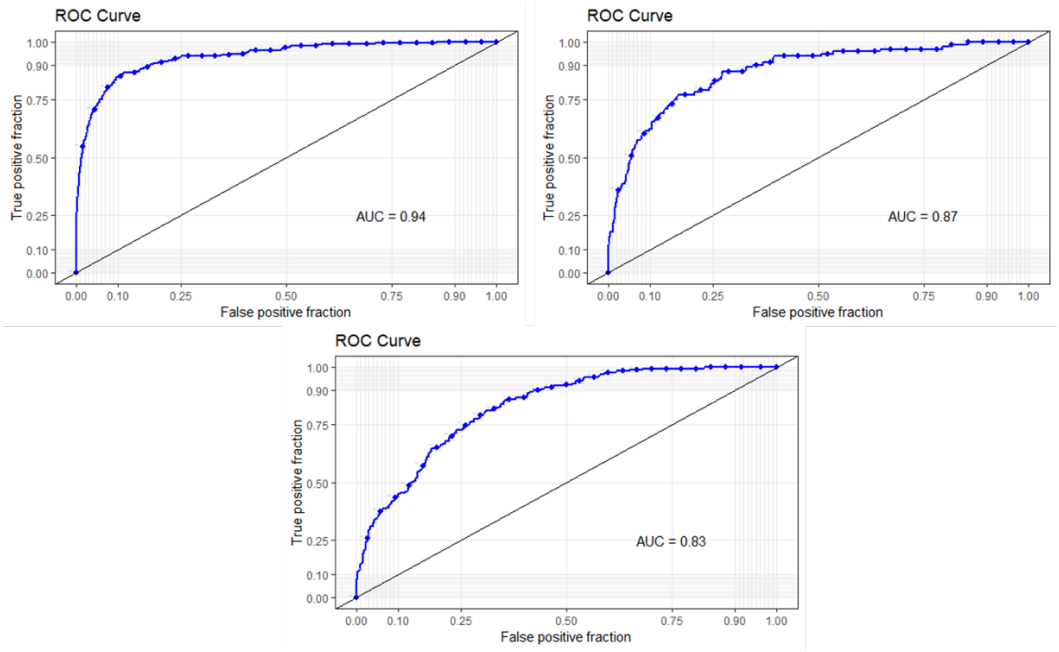
Interestingly, the application of binning drastically changes the situation. While the Gini coefficients are larger for training and test pieces, the performance on validation set stands out of the picture. This finding signals the existing of outliers or outbursts in the validation set which are potentially smoothed over the binning process. Also, improvement over all metrics indicates that application of the log transformation for RHS variables can be misleading and different specification/model setup should be tested.

2.4.5 XGBoost: higher in-sample fit, lower robustness

Finishing the analysis, the outcomes from XGBoost are pushing in-sample fit to its higher bounds. As this exercise omitted fine tuning of the algorithm and used the automated search, the boosting might have overfitted the training data. The overfitting is a standard consideration for machine learning algorithms which can be defined as considering an excessive number of explanatory variables than the data can justify. It results into such model parameters that correspond to the modelling data with high precision but would fail to fit another dataset. The overfitting problem has been widely studied

within the computer studies field (Roelofs et al., 2019) and the consequences of overfitting can be observed through deterioration of model performance when used for different dataset. The algorithm creates model so customized that it is become not accurate even with minor characteristic deviations. While test and validation Gini characteristic values are comparable with binning numbers, the drop from in-sample metric is substantial. Also, it is worth mentioning that boosting outcome on validation sample is lower, meaning that model is less robust when used over certain period of time. To overcome this issue, one can perform tunning of input parameters of XGBoost which would increase versatility of the model across data at a cost of in-sample ranking quality.

Figure 3. ROC curves for XGBoost (modelling - upper left, test - upper right, validation - bottom)



Using the connection of the Gini coefficient with the ROC curve, the performance of ranking can be depicted in a series of plots. The outcomes for

binning and boosting models are presented in Figure 3. The profiles help with analyzing which part of the sample is harder to order. This might help with tunning the algorithm or using additional preparation steps (subsampling, resampling, extra data transformations) to improve the final outcome.

2.5 Chapter conclusion

This chapter addresses the methodological dimension of SME credit risk modelling, focusing on research questions M1 and M2 by evaluating whether machine learning techniques, represented by XGBoost, provide improvements over conventional econometric models such as logistic regression. The results indicate that while machine learning models can deliver higher predictive accuracy in-sample, this advantage is conditional and not uniformly sustained when broader criteria are considered. This finding is particularly relevant in the SME context, where data limitations and heterogeneity constrain model performance, which has its implications in line with proposed agency and stakeholder frameworks.

Taken together, the findings suggest that model performance is not driven by methodological complication, but by the interaction between modelling approach and the informational environment in which it is applied. Therefore, the comparison between logistic regression and XGBoost highlights a broader insight: gains from more complex algorithms depend critically on the quality and structure of the underlying data. This conclusion provides a direct link to the next stage of the analysis. If model outcomes are sensitive to how risk is represented, then the definition of the dependent variable becomes a central modelling choice.

3 Assessing Model Performance under Various Definition of Risk Event

A substantial piece of corporate-oriented literature utilized the bankruptcy status as the main default indicator. However, the Basel Committee on Banking Supervision release of the capital framework (proposed in 2001, published II accord in 2006) has boosted research oriented towards default definition and credit risk realization (Hayden, 2003; Lin et al., 2012).

To motivate further, the actual banking market rarely observes individual bankruptcies with capturing risk patterns through a delinquency or missed payment information. While 90 days of non-payment being standard within existing standards, the lower delinquency bands (like 30 or 60 days) may contain an early warning indication of a borrower to face the default status. At the same time, the higher range (180 days) could serve as an additional insight for further delinquency and potential recover purposes. To summarize, the exact default definition could be set variously creating an extra space for model differentiation. Absorbing features of both corporate and retail banking, the SMEs loans credit risk could be captured in multiple ways while benchmarked with standard IFRS9 definition.

Empirical literature on measuring SME credit risks leans towards the bankruptcy definition due to limited lending data availability and lower frequency (annual or biannual) of data collection. As an alternative, the credit risk model can be driven by attribute aggregated statistic with a qualitative overlay (like hierarchical model in Roy & Shaw, 2023), however such approach requires an expert assessment of factors as an input and might be suffering

from a biased evaluation. Thus, the large part of policy-oriented research is mostly restricted to specific regions or data-dependent designs due to obstacles above. As a result, a larger scale and multi-country analysis based on would be in a great interest of lending organisations and banks.

This research extends SME credit risk assessment by building a panel-style model across multiple definitions of risk. Similarly to Lin et al. (2012), multiple definitions of distress are analyzed; in contrast to their study, this work focuses on individual loans and their repayment as the key indicator of risk quality. With testing multiple specifications and information treatment approaches, the understanding of SME credit risk can be improved by redefining the risk event variable and providing analysis beyond classical bankruptcy and default definitions which suffer from rare occurrences and being underrepresented within modelling set (Lin et al., 2012). The special attention is paid towards robustness of assessment to determine which definition is more prone to changes in data distribution and provides stable model performance over time.

The empirical analysis demonstrates that nature or quantitative definition of a credit event does not affect model performance in the standard setup: while overall level of GINI is higher for the classical default definition, the difference across specifications is comparable. However, the robustness of models on subsamples demonstrates mixed results. The next stage would be to add coarse classing to check whether it improves model stability and extend the set of methodologies with machine learning techniques like XGBoost (Chen et al., 2016). The preliminary testing of those methods has been performed within empirical illustration of the handbook chapter paper.

3.1 Which factors are critical for SME credit risk modelling?

While SME borrowers' quality has perpetually been in the interest of financial institutions, the literature remains relatively limited. Nonetheless, improvements have been seen over past years as data coverage and collection develops and regulators apply better practices and initiatives in respect to SME sector transparency. Example of such initiative is AnaCredit that was initiated in 2011 by ECB. The data collection is definitive for model construction and decision maker, thus higher information disclosure would reduce information asymmetry and reduce opaqueness penalty on SME, as described, for instance, in Song et al. (2016). In the meantime, this paper considers key contributions to the field as well as specific research investigation of default definition for SME.

3.1.1 Justifying research method: credit risk definitions

The default event has its classical definition within credit risk framework. It associated with bad performance of the borrower who is failing to maintain their liabilities. However, exact interpretation of "failure" varies and there are multiple views and ways on its translation into a particular indicator. For instance, bankruptcy-identified (or other reasons) flags are easier to be tracked as they (i) might be available from financial reports and (ii) have clear connection with liquidity and solvency principles which are key indicators for repayment risk formulation. However, the legal status of bankruptcy is not necessarily provided, even if a firm experienced financial distress (Balcaen & Ooghe, 2006). Thus, while regulatory guidelines have formal proposed

definitions, several studies have their own rationale behind other options and their relevance to risk assessment.

In response to Basel capital adequacy frameworks, Hayden (2003) questions surface-laying hypothesis whether performance and/or structure of credit scoring models would depend on definition of default. Based on example of Austrian data set covering 1987-1999, the author compares three most standard definitions: bankruptcy, restructuring, and payment delinquency. Through univariate analysis and building of a scoring model, Hayden demonstrates that bankruptcy-definition model is not significantly inferior in predicting events of rescheduling and delay-in-payment defaults when compared to models which are built specifically for each definition. As a result, focussing on more granular definition (like latter two) may not be a solid improvement over standard bankruptcy definition.

Lin et al. (2012) also investigates how banks could extend their analysis of firm failures. The study expands classical bankruptcy with few more definitions based on “financial distress levels”. To overcome general problem of default events insufficiency, authors include proxies for bankruptcies to increase validity of risk models. Another innovation comes from acknowledgment of difference between modelling approaches for retail and corporate exposures while SME might demonstrate features of both classes. Correspondingly, authors use financial ratios as well as coarse classification leverage on both methodology sets. As a result, default definition variation is responsible for distinctions in model composition, while influencing sets of drivers and sensitivity to them. Based on that, one could think of building more

sophisticated model including data expansion and various estimation techniques to check the robustness of findings.

Di and Pattison (2023) analyse the U.S. SBA loan data focusing on loan outcomes such as charge-offs and default-related performance metrics. Their framework treats credit events through realized loan performance rather than explicit definitions of default or deterioration. The paper demonstrates that industry-specialized lenders exhibit improved loan performance and screening efficiency, suggesting that better information can reduce default incidence. However, the study consider credit risk primarily via post-default charge-offs rather than a structured sequence of credit events, thereby limiting its applicability for modelling early-stage deterioration or transition dynamics in SME portfolios.

Bertoni et al. (2023) examine EU-guaranteed SME loans and introduce a broader concept of credit events by identifying firms experiencing a significant increase in leverage (more than 5% liabilities-to-assets ratio). This proxy captures balance-sheet stress and potential deterioration, complementing traditional performance indicators such as growth, productivity, and survival. The study shows that guaranteed loans improve firm outcomes and reduce failure risk over a long horizon. However, the definition of a “credit event” remains indirect and accounting-based, reflecting financing structure changes rather than clear credit-risk triggers such as delinquency, restructuring, or insolvency, which limits comparability with regulatory or modelling definitions of default.

Acebo et al. (2026) focus on European micro, small and medium enterprises (MSMEs) during COVID-19 and define financial distress through

the concept of “zombie firms,” operationalized as firms with an interest coverage ratio (ICR) below 1 for three consecutive years. While not a credit distress definition itself, zombification can be interpreted as a high credit risk state, capturing structural SME distress that precedes and partially explains observed credit events. This approach captures persistent distress rather than discrete default events, and the study further introduces a multi-criteria definition of recovery (“de-zombification”) based on profitability, leverage, and growth. The findings highlight that targeted public guarantees can facilitate recovery, particularly for small firms. Nevertheless, the reliance on accounting-based distress proxies introduces potential misclassification, as temporary liquidity shocks may be indistinguishable from structural insolvency, and the framework does not map directly to firm survival or credit events.

Across the reviewed literature, a consistent limitation is the absence of a harmonized and granular definition of SME credit events. Studies rely on various proxies capturing different aspects of credit risk but are not exactly comparable. As a result, it is challenging to formulate a unified framework linking early deterioration, credit distress, and subsequent default in a single flow. At the same time, due to disparity of definitions in studies, research findings might be incomparable.

3.2 Research design

3.2.1 Credit risk event definition

In addition to the default definition, which was introduced in Section 2.3.2, there are two extra credit risk event flags (CREF) which are being considered:

1. 12-month forward looking default event. In this case, the CREF is calculated as rolling maximum of 12 forwards of default flag defined in the first point. Such idea is derived from Basel III and IFRS9 regulations (IASB, 2014). The CREF as defined evaluates whether account will default within upcoming year.

$$12M_{DEF_{it}} = \begin{cases} 1, & if\ \max(DE_{it+1}, \dots DE_{it+12}) = 1 \\ 0, & otherwise \end{cases} \quad (6)$$

As the definition interpretates forward looking nature, it is crucial to have window of at least 12 months for a certain loan to be observed.

2. Missing payment delinquency definition: 30+ days past due

$$DEL_t = \begin{cases} 1, & if\ delinquency_t \geq 30 \\ 0, & otherwise \end{cases} \quad (7)$$

The one-missed-payment definition could be used as *early warning indicator* (EWI) for incoming default. Such event is equivalent of significant increase in credit risk within IFRS 9 regulation (IASB, 2014).

3.2.2 Specification

The analysis starts with classical logistic panel regression approach which is in line with the literature described above. The general specification ("basic specification") could be presented as:

$$\log\left[\frac{p(y_{it})}{1-p(y_{it})}\right] = \alpha_0 + \alpha_1^{FE} + \cdots \alpha_k^{FE} + \beta_1 \ln(1+x_{i1}) + \beta_2 \ln(1+x_{i2}) + \cdots + \gamma_1 \ln(1+z_{it1}) + \gamma_2 \ln(1+z_{it2}) + \cdots, \quad (8)$$

or, after combining factors and reverting logistic function:

$$Y_{it} = L^{-1}[\alpha + \beta X_i + \gamma Z_{it}] \quad (9)$$

As of the equation above, α characterizes intercept α_0 and fixed-effects terms (α^{FE}) for relevant qualitative factors, Y_{it} stands for CREF binary variable outcome, X_i stands for time-invariant factors (group 1) and Z_{it} stands for time-variant characteristics (group 2). Expanding the basic specification with financial data (F_i) brings the full financial specification ("full specification"):

$$Y_{it} = L^{-1}[\alpha + \beta X_i + \gamma Z_{it} + \theta F_i] \quad (10)$$

Alternatively, financial-information-only specification ("financial specification") includes only factor that related to firm (X_i^*) and its financial performance without inclusion information on loan details:

$$Y_{it} = L^{-1}[\alpha + \beta X_i^* + \theta F_i] \quad (11)$$

Important to mention, that panel structure of data is rarely tested – the observed default frequencies are considerably low, thus it creates a challenge to evaluate statistically significant patterns as default becomes a very rare event. This nuance emphasizes research question of whether delinquency definition is more resolving.

3.3 Results

The calculation has been performed in RStudio 4.0.5 with reliance on base R functionality as well as classical libraries for the data manipulation (data.table, tidyverse, lubridate), modelling data split and model estimation/ training (caret, fixest), evaluation (pROC), and reporting (stargazer). More detailed guidance on modelling in R could be inspired by Finch et al. (2019) and Wickham and Grolemund (2016).

The estimation sample has been restricted to keep observations with non-missing values for each variable; thus, there is a difference in sample size

between specifications. The financial information is available for a considerably lower subsample; thus, the results of basic specification are provided both for full sample and financial-information-available subsample (cut sample). To check to comparability of the outcome, Wilcoxon-Mann-Whitney and Kolmogorov-Smirnov tests were conducted, which resulted into significant difference between full sample and financial-information-available subsample. Performance evaluation for models is done using commonly used the receiver operating characteristic (ROC) which is transformed into Gini coefficient using the standard formula:

$$Gini = 2 \times AUC - 1, \quad (12)$$

where AUC is area under the ROC curve. The original ROC curve plots true-positive rate against false-positive rate comparing model sensitivity and specificity; example of its application could be found in Gupta et al. (2014).

3.3.1 Information completeness

Thus, as starting point, the general specification is estimated both for full and cut samples. The specification is the same; however, the estimation sample is different as for the cut version the additional restriction on non-missing financial variables information is imposed. The results for those multi-country datasets (both Spain and Italy) are presented in the Table 6. The former presents estimation outcome for the baseline specification across 3 definitions. The interest rate and number of employees effects are only to be universal across all 6 regressions while the rest factors present mixed evidence. This is especially true for loan progression ratio which is switching the sign of coefficient for both 12-month and usual default flags keeping 1% significance

in both estimation sets. Similar picture is observed for loan balance within delinquency definition. This indicates that data distribution experience important shift when the dataset is cut.

At the same time, the latter table highlights that even the same set of drivers demonstrate considerably better outcome for dataset with populated financial information. One can imply that there is a correlation between observability and explainability of risk events. The finding is confirmed across all tested CREFs with a lower difference its performance for the default flag. As a result, to interpret effects meaningfully, only the dataset with available financial data is used in further analysis.

Table 6. Estimation results for Spain and Italy

	12-month default		Delinquency flag		Default flag	
	(Cut)	**(Full)**	**(Cut)**	**(Full)**	**(Cut)**	**(Full)**
Interest Rate	0.8752***	0.5674***	0.9889***	0.5244***	0.8438***	0.3069***
	(0.0194)	(0.0052)	(0.0242)	(0.0056)	(0.0520)	(0.0101)
Loan Balance	0.0745***	0.0196***	0.0773***	-0.0157***	0.0106	-0.1939***
	(0.0055)	(0.0016)	(0.0069)	(0.0017)	(0.0138)	(0.0026)
Loan Term	-0.0439***	0.2201***	0.2634***	0.4957***	0.2232***	0.5052***
	(0.0161)	(0.0065)	(0.0205)	(0.0078)	(0.0436)	(0.0164)
Number of Employees	-0.1032***	-0.0543***	-0.1389***	-0.0897***	-0.0543**	-0.0243***
	(0.0086)	(0.0030)	(0.0109)	(0.0037)	(0.0222)	(0.0082)
Loan Progression	0.3983***	-0.4785***	2.010***	0.7401***	1.408***	-0.8453***

	(0.0557)	(0.0232)	(0.0717)	(0.0283)	(0.1526)	(0.0617)
Observations	1,081,790	4,221,464	1,085,538	4,222,873	1,079,401	4,221,464
Squared Correlation	0.0195	0.0096	0.0147	0.0049	0.0038	0.0031
Pseudo R2	0.0747	0.0415	0.0837	0.0390	0.0741	0.0678
BIC	209,742	936,657	145,705	693,534	42,353	185,083
Fixed Effects						
Obligor Legal Form	✓	✓	✓	✓	✓	✓
NACE group	✓	✓	✓	✓	✓	✓
Amortization Type	✓	✓	✓	✓	✓	✓
NUTS	✓	✓	✓	✓	✓	✓

Table 7. Gini performance indicators for pooled data (Spain and Italy)

Specification	Cut	Full
12-month forward-looking default	0.4676	0.3477
Delinquency flag	0.5159	0.3647
Default flag	0.5377	0.5065

3.3.2 Effects:

3.3.2.1 General and financial information

To account for general and financial information separately, the next part analyses financial-only and financial specification. The results are more consistent across the board (Table 8 and Table 9) when compared to pre-cut outcomes.

Table 8. Estimation results for 12-month default and delinquency flag

	12-month default			Delinquency flag		
	(1)	(2)	(3)	(4)	(5)	(6)
Interest Rate	0.8709***		0.8054***	0.9975***		0.9587***
	(0.0233)		(0.0238)	(0.0289)		(0.0295)
Loan Balance	0.0727***		0.0815***	0.0714***		0.0754***
	(0.0066)		(0.0067)	(0.0082)		(0.0083)
Loan Term	-0.0649***		-0.1100***	0.2599***		0.2273***
	(0.0193)		(0.0201)	(0.0245)		(0.0254)
Number of Employees	-0.1089***		-0.1423***			
	(0.0103)		(0.0130)			
Loan Progression	0.3618***		0.3958***	1.927***		1.997***
	(0.0666)		(0.0673)	(0.0856)		(0.0867)
Return on Equity		-0.1753***	-0.2362***		-0.1981***	-0.2658***
		(0.0205)	(0.0209)		(0.0249)	(0.0254)
Short-term to Total Liabilities		1.222***	1.047***		1.104***	0.9378***
		(0.0366)	(0.0382)		(0.0453)	(0.0474)

Debt to Equity		0.1093***	0.1538***		0.1242***	0.1712***
		(0.0154)	(0.0156)		(0.0184)	(0.0187)
Leverage		2.760***	2.633***		2.550***	2.404***
		(0.1038)	(0.1039)		(0.1284)	(0.1284)
Observations	757,503	757,503	757,503	760,152	760,152	760,152
Squared Correlation	0.0198	0.0245	0.0268	0.0148	0.0171	0.0192
Pseudo R2	0.0753	0.0906	0.1001	0.0846	0.0921	0.1048
BIC	147,463	145,057	143,624	102,469	101,652	100,321
Fixed Effects						
Obligor Legal Form	✓	✓	✓	✓	✓	✓
NACE group	✓	✓	✓	✓	✓	✓
Amortization Type	✓		✓	✓		✓
NUTS	✓	✓	✓	✓	✓	✓
SME Size		✓	✓		✓	✓

(1) & (4) are basic specification; (2) & (5) are financial specification; (3) & (6) are full specification

Table 9. Estimation results for default flag

	Default flag		
	(1)	**(2)**	**(3)**
Interest Rate	0.8521***		0.7994***
	(0.0613)		(0.0628)
Loan Balance	-0.0109		
	(0.0159)		(0.0161)
Loan Term	0.2328***		0.1784***
	(0.0513)		(0.0535)
Number of Employees	-0.0497*		
	(0.0264)		
Loan Progression	1.230***		1.306***
	(0.1790)		(0.1816)
Return on Equity		-0.2811***	-0.3517***
		(0.0522)	(0.0529)
Short-term to Total Liabilities		1.341***	1.139***
		(0.0966)	(0.1007)
Debt to Equity		0.1088***	0.1554***
		(0.0386)	(0.0389)
Leverage		3.533***	3.392***
		(0.2932)	(0.2919)
Observations	753164	753164	753164

Squared Correlation	0.00381	0.00592	0.0059
Pseudo R2	0.07293	0.0892	0.09619
BIC	30707	30200	30050
Fixed Effects			
Obligor Legal Form	✓	✓	✓
NACE group	✓	✓	✓
Amortization Type	✓	✓	✓
NUTS	✓	✓	✓
SME Size Category	✓	✓	✓

1) is basic specification; (2) is financial specification; (3) is full specification

First, the significance and signs of interest rate are in line with previously discussed literature. The positive coefficient of interest rate across all definitions and demonstrate that riskier loans are usually associated with higher rates. The higher magnitude here can be simply connected to the fact that delinquency frequency is higher.

The financial information variables are highly significant and in line with expected intuition. Similarly to Chapter 2 outcomes, the coefficients follow the same logic, confirming the previously described patterns. Additionally, leverage has been added, which is has significantly higher magnitude. As logistic regression is utilized, the coefficient can be interpreted as percentage increase, implying that 1% increase in leverage would cause 2.633% increase in frequency of 12-month defaults, 2.404% for delinquency flag, and 3.392% for classical default (full specification outcomes). By far, leverage is highest sensitivity factor, demonstrating how critical indebtedness is important for SMEs.

High reliance on short-term liabilities is perceived to be a signal of higher risk, while better performance and lower reliance on debt are improving creditworthiness. This finding is in line with assumption that structure of debt

(short-term vs. long-term) might serve as an important indicator of company's credit quality (Gopalan, Song, & Yerramilli,2014) Company size is also improving credit risk outputs, meaning larger companies are less likely to experience credit risk event.

With respect to loan term, the evidence is mixed. This might be connected to the fact that 12-month default definition is forward-looking and accounting for the events happening in future rather contemporaneous nature of default and delinquency flag, thus the timing of event is different. The combination of the negative sign for the 12-month default definition and the positive sign for the rest means that quite a few defaults happen within first few quarters. This might be also a reason, why coefficient for loan progression is lower when compared to default and delinquency flags. Loan balance has minor positive effect on the credit event likelihood for 12-month default and delinquency definitions but insignificant for default definition, concluding minor impact of this factor.

The implications of these findings are significant, especially in the current macroeconomic environment. With rising interest rates globally, businesses must maintain strong financial health to secure favorable loan terms and avoid high capital cost pressure. Larger companies benefit from lower credit risk estimate, suggesting that growth and expansion can be strategic goals for reducing financial vulnerability. Thus, growth opportunities, such as cost-efficiency innovations described in De Blick et al. (2024), are expected to provide better risk resilience and ensure better access to finance.

3.3.2.2 Size Effects

Size of the SME can play crucial role for multiple business operations and both policymakers and financial regulators must take it into consideration when designing support mechanisms. For instance, credit guarantee schemes or financial assistance programs may need to target micro-SMEs more aggressively due to their higher credit risk propensity. As another example, capital requirements for loans to micro-SMEs could be adjusted to account for their higher likelihood of default.

To answer the question, the financial specification has been estimated with direct inclusion of SME class into the regression. The “medium” class has been picked as base group. The results are presented in Table 10.

Table 10. Estimation results for SME size effects

	Without SME effects factor			With SME effects factor		
	Default	**Delinquency**	**12-month default**	**Default**	**Delinquency**	**12-month default**
Interest Rate	0.8019***	0.9744***	0.8341***	0.7940***	0.9510***	0.8100***
	(0.0530)	(0.0245)	(0.0198)	(0.0533)	(0.0247)	(0.0199)
Loan Balance	0.0135	0.0624***	0.0653***	0.0191	0.0818***	0.0834***
	(0.0134)	(0.0065)	(0.0053)	(0.0140)	(0.0069)	(0.0056)
Loan Term	0.1758***	0.2526***	-0.0626***	0.1676***	0.2281***	-0.0866***
	(0.0451)	(0.0211)	(0.0166)	(0.0454)	(0.0213)	(0.0168)
Loan Progress	1.459***	1.998***	0.3599***	1.485***	2.083***	0.4347***
	(0.1536)	(0.0718)	(0.0557)	(0.1547)	(0.0726)	(0.0563)
Return on Equity	-0.3503***	-0.2755***	-0.2333***	-0.3496***	-0.2758***	-0.2337***
	(0.0445)	(0.0212)	(0.0174)	(0.0446)	(0.0213)	(0.0175)
Short-term to Total Liabilities	1.108***	0.9070***	1.042***	1.118***	0.9407***	1.071***
	(0.0852)	(0.0396)	(0.0319)	(0.0853)	(0.0396)	(0.0319)
Debt to Equity	0.1942***	0.1959***	0.1721***	0.1922***	0.1909***	0.1668***
	(0.0324)	(0.0155)	(0.0129)	(0.0325)	(0.0155)	(0.0129)

Leverage	3.022***	2.293***	2.532***	3.021***	2.284***	2.527***
	(0.2427)	(0.1066)	(0.0864)	(0.2426)	(0.1066)	(0.0864)
SME Class: Micro				0.0787	0.3873***	0.3152***
				(0.1025)	(0.0516)	(0.0379)
SME Class: Small				-0.0130	0.1810***	0.1290***
				(0.1043)	(0.0525)	(0.0383)
Observations						
	1,079,401	1,085,538	1,081,790	1,079,401	1,085,538	1,081,790
Squared Correlation	0.00571	0.01927	0.0267	0.0057	0.01918	0.0266
Pseudo R2	0.09609	0.10336	0.09891	0.09617	0.10406	0.09953
BIC						
	41,434	142,669	204,343	41,458	142,587	204,231
Fixed Effect						
Obligor Legal Form	✓	✓	✓	✓	✓	✓
NACE group	✓	✓	✓	✓	✓	✓
Type of Amortization	✓	✓	✓	✓	✓	✓
NUTS	✓	✓	✓	✓	✓	✓

The findings advocate for the unique risk profiles of each SME size category, extending the findings presented in the Chapter 2. The results show that size has effect on facing 12-month default and delinquency events with micro-SME demonstrating highest amplitude followed by small-sized firms. This evidence has implication for size heterogeneity, as larger SMEs tend to be more resilient to shocks. It also confirms the necessity to identify SME classes as separate groups. At the same time, classical default event specification has no significant size effect, suggesting the implications being relevant only for certain credit risk events.

3.3.3 Robustness

3.3.3.1 Out-of-time evaluation

Risk evolution over time is key principle to be considered in the longitudinal studies, thus, the performance and model robustness are important aspects of the credit risk analysis. The drop in performance over time or over different dataset might reveal model limitation or lack of relevance outside of development sample. To comment on the performance, the associated tables for train, test and OOT sets are presented in Table 11

Table 11. Gini performance indicators for train, test, and OOT datasets

Specification	Sample	12-month default	Default flag	Delinquency flag
Basic		0.4676	0.5377	0.5159
Financial	Modelling	0.5112	0.5807	0.5378
Full		0.5365	0.6063	0.5717
Basic		0.4603	0.5413	0.5059
Financial	Test	0.5059	0.5764	0.5264
Full		0.5306	0.5989	0.5626
Basic		0.4493	0.4161	0.4039
Financial	OOT	0.3622	0.3159	0.3219
Full		0.5262	0.4925	0.4984

The estimation using modelling dataset demonstrates that financial-only specification is outperforms basic specification across all definitions with a subsequent improvement in a form of the full specification. Unsurprisingly, financial information has dominant contribution for the model ranking abilities. The outcomes for the test dataset are generally the same, standing within +/- 1.5 points range of the modelling GINI values implying that findings are robust outside estimation set.

The out-of-time dataset reveals more insights on model robustness as relative number of defaults in OOT sample is 2 times lower (0.37% vs 0.74% of observations). First observation is that financial specification evaluation significantly drops over all 3 considered definitions. This might be connected to the fact that specification doesn’t contain any time-varying factors, thus there

are no factors that could potentially explain a default rate drop, subsequently decreasing accuracy.

While initial in-sample GINI is lower for the 12-month default definition, it makes the most robust representation, losing less than 1 point for test and OOT samples for the full specification. Despite higher representation of events in the data (number of delinquent observations is about 2.2 times higher than number of default events), the delinquency definition doesn't allow model to reach higher performance outcome with an about 3-points gap for modelling and test sets and on par values for OOT estimation. This finding can be interpreted in several ways. First, it justifies defining risk event over horizon, as its robustness increases dramatically, meaning model outcomes remain valid over longer period and for different set of exposures. At the same time, delinquency and default definitions are less sensitive to different set of obligors rather than time-driven patterns. Inclusion of time-sensitive variables such as macroeconomic variables might mitigate this gap. Also, adding more risk drivers, especially on qualitative side, can improve the results.

The rarity of default events in the OOT sample highlights the challenges of predicting low-frequency outcomes. While forward-looking definitions, such as the 12-month default indicator, showed better robustness, they also require sufficient observation periods and high-quality data to maintain accuracy.

3.3.3.2 After COVID-19

The previous section showed that out-of-time assessment has very meaningful implication for risk definitions. As the pandemic has changed operations of corporates with digital innovations (Holl & Rama, 2024), credit risk patterns

have been adjusted, leading to a natural question whether model need to be changed or their outcomes are still valid. Thus, to extend out-of-time evaluation, we extract additional dataset for Italy and Spain to calculate performance metrics and relevance of the model after COVID-19.

Table 12. Post-COVID-19 sample summary statistics

Statistic	N	Mean	St. Dev.	Min	Max
Interest Rate	172,732	2.795	1.897	0	15
Loan Balance	172,732	110,926	413,980	-	18,000,000
Loan Term	172,732	86.718	62.982	6	481
Number of Employees	172,732	23	36	1	250
Loan Progress	172,732	0.604	0.239	0.017	0.996
Return on Equity	172,732	0.32	1.241	-64	125
Short-term to Total Liabilities	172,732	0.312	0.285	0	1
Debt to Equity	172,732	2.498	7.961	0	824
Leverage	172,732	0.495	0.251	0	1
Turnover	172,732	3,450,236	6,473,598	10,000	49,880,000

The updated dataset covers period from April 2020 to June 2023 and has the same data preparation as main sample. Indeed, the summary statistic concludes that there is shift in the distribution, showing lower interest rates and lower return on equity which is expected by us due to appearing new loans after COVID-19 stress. Interestingly, the leverage (lower) and term structure of liabilities (lower share of short-term liabilities) are moving away from stress which might be associated with development of guarantees program during 2015-2019 and additional government support shortly after pandemic as described in Corredera-Catalán et al. (2021) for Spain and Casanova et al. (2021) and Falagiarda et al. (2021) generally for Euro area. The latter highlights the switch to longer term of lending in first half of 2020 as well as role of public loan guarantees that supported such lending to SMEs. Both Spain and Italy are among top 5 countries in terms of guarantees issued as % of GDP among advanced economies (Casanova et al.,2021, graph 2) which confirms more pronounced switch in the sample statistics. This has its reflection in the dataset statistics shift as the event frequency in short-term period after the COVID demonstrates drop in defaults (2.65% vs 4.06% for ever defaulted, 1.4% vs 2.3% for classic default flag).

Table 13. Gini performance indicators for after-COVID-19 sample

Specification	12-month default	Default flag	Delinquency flag
Basic	0.4339	0.4119	0.4409
Financial	0.4079	0.3896	0.3797
Full	0.548	0.5408	0.557

To evaluate performance of models, we predict the event probability as done for OOT testing but with after COVID-19 sample. The GINI metrics is presented in Table 13. The outcomes expand previous OOT testing: while numbers are slightly better overall, the full specification is dominating more than 0.1 point when compared against basic one and up to 0.15 when compared with financial only. The latter specification performs considerably better implying financial variables could explain more variation in the after COVID-19 sample. At the same time, the variation across event definitions is not as pronounced and cannot justify choosing one to be superior of other.

Interestingly, results for after-COVID-19 sample are better than previous OOT. Technically, this could mean that originally estimated coefficients work better in new reality, which can be translated into absence of structural shift in risk patterns. While macroeconomic shocks and credit cycle can alter sensitivities to certain drivers (e.g., making more reactive to interest rates or term structure during higher-pressure periods), modelling set is long enough to have minimum requirement for time variance. Also, the stability is partially achieved through versatile fixed effects structure, accounting for regional and sectoral specifics.

3.3.4 Single-country models

While the presented specification accounts for region-specific effects, the sensitivity to factors is assumed to be the same for all loans irrespective of country of origination. To reduce dependency on this assumption, we estimated the full specification separately on Italian and Spanish subsets and evaluated performance and robustness. The results are presented in Table 14-Table 16.

Table 14. Delinquency flag estimation results for Italy, Spain, and pooled samples

Delinquency flag	Full	IT	ES
Current Interest Rate	0.9587***	0.8509***	1.352***
	(0.0295)	(0.0333)	(0.0762)
Outstanding Balance	0.0754***	0.0937***	0.0493***
	(0.0083)	(0.0098)	(0.0148)
Loan Term in months, log tr.	0.2273***	0.1160***	0.4705***
	(0.0254)	(0.0293)	(0.0608)
Loan Progression Ratio	1.997***	1.728***	2.454***
	(0.0867)	(0.0994)	(0.1966)
Return on Equity Ratio	-0.2658***	-0.2733***	-0.0497
	(0.0254)	(0.0265)	(0.0959)
Short-term to Total Liabilities Ratio	0.9378***	1.037***	0.5632***
	(0.0474)	(0.0519)	(0.1206)
Debt to Equity Ratio	0.1712***	0.1860***	-0.2053**
	(0.0187)	(0.0196)	(0.0802)
Leverage Ratio	2.404***	2.405***	3.447***
	(0.1284)	(0.1453)	(0.3627)
Observations	760,152	383,684	376,460
Squared Correlation	0.0192	0.01529	0.00545
Pseudo R2	0.10482	0.05881	0.07608
BIC	100,321	79,972	20,287
Fixed Effects			
ObligorLegalForm	✓	✓	✓
NACE group	✓	✓	✓
Amortization Type	✓	✓	✓

NUTS	✓	✓	✓
SME Size Category	✓	✓	✓

Table 15. Default flag estimation results for Italy, Spain, and pooled samples

Default flag	**Full**	**IT**	**ES**
Current Interest Rate	0.7994***	0.6603***	1.576***
	(0.0628)	(0.0704)	(0.1838)
Outstanding Balance	-0.0028	0.0097	-0.0314
	(0.0161)	(0.0193)	(0.0287)
Loan Term in months, log tr.	0.1784***	0.1073*	0.3675***
	(0.0535)	(0.0608)	(0.1399)
Loan Progression Ratio	1.306***	1.131***	1.111**
	(0.1816)	(0.2056)	(0.4330)
Return on Equity Ratio	-0.3517***	-0.3394***	-0.6049**
	(0.0529)	(0.0543)	(0.2498)
Short-term to Total Liabilities Ratio	1.139***	1.283***	0.2863
	(0.1007)	(0.1089)	(0.2793)
Debt to Equity Ratio	0.1554***	0.1624***	-0.2315
	(0.0389)	(0.0404)	(0.2014)
Leverage Ratio	3.392***	3.462***	4.668***
	(0.2919)	(0.3280)	(0.9216)
Observations	753,164	382,555	370,609
Squared Correlation	0.0059	0.00501	0.00154
Pseudo R2	0.09619	0.05623	0.06995
BIC	30,050	24,746	5,480
Fixed Effects			

ObligorLegalForm	✓	✓	✓
NACE group	✓	✓	✓
Amortization Type	✓	✓	✓
NUTS	✓	✓	✓
SME Size Category	✓	✓	✓

Table 16. 12-month default flag estimation results for Italy, Spain, and pooled samples

12-month default	Full	IT	ES
Current Interest Rate	0.8054***	0.6176***	1.466***
	(0.0238)	(0.0287)	(0.0531)
Outstanding Balance	0.0815***	0.0801***	0.0842***
	(0.0067)	(0.0082)	(0.0124)
Loan Term in months, log tr.	-0.1100***	-0.1675***	-0.0467
	(0.0201)	(0.0246)	(0.0422)
Loan Progression Ratio	0.3958***	0.3456***	-0.1695
	(0.0673)	(0.0830)	(0.1267)
Return on Equity Ratio	-0.2362***	-0.2401***	-0.1468**
	(0.0209)	(0.0221)	(0.0661)
Short-term to Total Liabilities Ratio	1.047***	1.213***	0.5155***
	(0.0382)	(0.0435)	(0.0821)
Debt to Equity Ratio	0.1538***	0.1839***	-0.1679***
	(0.0156)	(0.0166)	(0.0548)
Leverage Ratio	2.633***	2.464***	3.811***
	(0.1039)	(0.1224)	(0.2532)
Observations	757503	382621	374882
Squared Correlation	0.02682	0.02464	0.0111
Pseudo R2	0.10013	0.06969	0.0765

BIC	143624	105577.6	37658.1
Fixed Effect			
ObligorLegalForm	✓	✓	✓
NACE group	✓	✓	✓
Amortization Type	✓	✓	✓
NUTS	✓	✓	✓
SME Size Category	✓	✓	✓

Eventually, the model estimation results are quite similar for both countries. While sensitivity to interest rate and leverage are higher for Spain, the term structure of liabilities is less pronounced across all definitions. With respect to other factors, there are ambiguous patterns from one definition to another. The outstanding balance is more important in the Italian model when delinquency factor is being used as target but is comparable in 12-month default setup (default event regression concluded this factor to be insignificant across the board). Return on equity is less important for Spanish SMEs with delinquency and 12-month default specifications but unexpectedly jumps in the default one. Loan term, loan progression and debt to equity ratio demonstrate unintuitive sign switch for Spain which make the model less transparent for explanation. Such abnormalities can be connected to the low-frequency nature of events, making results sensitive to the size of modelling sample. Indeed, when we move to GINI performance metrics, both Spain and Italy are performing worse when evaluated on validation data sample (Table 17), confirming the combined multinational model to maintain better robustness, irrespective of event definition. This is in line with assumption that both economies are quite similar

in their credit market state including high fragmentation (Moscalu et al., 2020) and low judicial efficiency (Mc Namara et al., 2020). Both studies incorporate SAFE survey data which demonstrate convergence in SME access to finance for both regions as well.

Table 17. Gini performance indicators for Spain and Italy samples

Specification	Country	12-month default	Default flag	Delinquency flag
Basic		0.4345	0.487	0.4825
Financial	Spain	0.4285	0.462	0.4211
Full		0.5046	0.5572	0.5371
Basic		0.3253	0.3265	0.3121
Financial	Italy	0.4141	0.4274	0.3766
Full		0.4318	0.454	0.4148

The implications of this finding are twofold. On one side, this suggests that sensitivities to risk drivers are quite similar and suggests that both regions can be pooled together. Moreover, there is a clear benefit from it as higher number of defaults in combined modelling set make estimates less volatile and closer to “true” value. Another finding can be made around sufficiency of regional fixed effect to account for variation in risk patterns. This has important implication for policymaking – if sensitivities are not expected to be different, the policymakers might expect similar sensitivities to risk factors, and design polices in more unified way while accounting for different default rates in different regions. As example, a marginal effect of subsidized interest rates should have similar reduction in credit risk events probabilities, while absolute level might be lower or higher depending on the region or country.

This can be especially relevant in the light of recent changes in insolvency regulations. Both Spain and Italy underwent regulatory changes in 2022 (implementation of EU Restructuring Directive 1023, 2019) to stimulate SMEs restructure instead of liquidation. However, countries chose different paths for its implementation: Italian Business Crisis and Insolvency Code (Decree-Law No. 7/2022) suggested more duty-forward including proactive stance when both firms and lenders are legally obliged to embed early default detection in corporate governance by additional reporting and mechanisms to alert potential distress (such as internal warning reports). At the same time, the Spanish reform (Law 16/2022) has more voluntary and incentive-based, emphasizing debtor initiative. While both policies are expected to provide higher recovery values and faster resolution of distressed debt as well as improve unemployment hikes due to SME bankruptcies, actual efficiency may vary due to mandatory or voluntarily nature. Since this directive replaces previous mid-20 century outdated regulations, actual impact evaluation is of immense value, provided multi-region analysis can demonstrate difference in policy outcome through change in regional effects.

3.4 Discussion of potential extensions

3.4.1 Application of alternative estimation techniques

While classical statistical models have been studied for decades, the machine learning techniques and AI technologies become increasingly more adopted in the field of credit risk. Development of credit scoring systems has become very popular in recent years. Since even small improve in those models grant banks with high profits (Hand & Henley, 1997), credit scoring attracts

researchers from both academia and public sector, especially provided such development is connected to increase of available data as well as technological advancement.

To start with, the binning procedure has been tested as basic technique that helps to control the robustness. In such setup, loan-level characteristics could undergo a classification or binning procedure as a part of data preparation. The process of coarse and fine classing to create bins is a quite standard technique for the scorecard creation (Lin et al., 2012) as it allows a modeler to focus on key patterns within one explanatory variable. Moreover, it makes model less sensitive to selection bias or outliers (Calabrese et al., 2019). Overall, the binning process algorithm is extensively described in Siddiqi (2017) which iteratively defines cut-points by merging quantiles into non-overlapping groups with a target of IV maximization. This provides classic econometric approach with perfect interpretation and explainability; however, when the modelling data become more sophisticated and richer, the advanced scoring techniques could be preferred.

As an example of a study that provides comparison analysis, Abdou et al. (2016) evaluated conventional regression, classification and regression tree (CART) and cascade correlation neural network (CCNN) to get better picture how far can machine learning and NN reach in terms of the model performance. They concluded that neural networks demonstrate higher accuracy and could be considered “to be of critical interest to bankers” (p.102). The similar observations were confirmed by Ala'raj and Abbod (2016), Abellan and Castellano (2017), and various other studies which focus on credit scoring systems. While robustness is one of key elements of evaluation among those

studies, analysing them over various credit risk event definitions can add extra dimension to the picture.

3.4.2 Macroeconomic models

Usually scoring systems assess default risks of each entity independently and focus on classification power. At the same time, these models can also be used in portfolio or segment-focussed models when the conditionality of defaults within a certain group is essential factor to be accounted. One of the most practical ways to model conditional defaults is a mixture model. In such model default probabilities are expected to be dependent not only on a set of individual characteristics, but also on a list of common economic factors including macroeconomic variables. Given this fact, interdependence between defaults is modelled as sensitivity to those macroeconomic factors (Frey & McNeil, 2003). The importance of macroeconomic variables inclusion was also highlighted by Filipe et al. (2016) as they identify “significant regional variations”. Karas (2022) developed a Cox proportional hazard model for EU-28 SMEs (2014-2019) that combined firm financials with macro indicators. The study identified the employment rate (inversely related to unemployment) and the interest rate to be significant determinants of SME survival. Higher employment was associated with longer survival and lower default probability, while higher interest rates reduced survival time. Including these macro factors added a slight boost in predictive power (higher AUC) compared to using only accounting ratios, implying that accounting the macroeconomic environment in SME default models can be beneficial for the analysis.

Macroeconomic models can have different application and reveal patterns that would be hard to explore independently. For example, they might simulate the impact of a severe economic downturn on the default rates of loans across different sectors: to accounting for the cascading effects of defaults in one sector on other sectors, providing a more intertwined systemic risk representation. This level of analysis might be particularly illustrative for small and medium enterprises (SMEs), which are often more vulnerable to economic fluctuations.

While traditional credit scoring models often focus on individual borrower characteristics, they may not fully capture the broader economic factors that can influence credit risk. This is where multi-layer macroeconomic credit risk models come into play complementing the existing traditional methodologies to provide holistic view (Stein, 2012). Also, central banks may utilize stress-testing exercises to stimulate or restrict lending and risk appetite (Shapiro & Zheng, 2024). Thus, these models are quite important for both regulators and lending institutions.

The construction of such combined model might be not so straightforward, as variable selection and data preparation are critical to avoid model inconsistency (Ferrari et al., 2011). To account for macroeconomic factors, the corresponding element can be embedded directly or through modelled central default tendency (CDT). Example of the first approach is presented by Tinoco and Wilson (2013). In the context of model specification, the region-specific effects should be replaced with lagged time-variant macroeconomic component (M_{it-1}), resulting into:

$$Y_{it} = L^{-1}[\alpha + \beta X_i + \gamma Z_{it} + \theta F_i + \eta M_{it-1}] \quad (13)$$

The macroeconomic component could be inserted directly as macroeconomic factor like GDP or unemployment; however, the sensitivity to those variables may vary from one region to another, thus this approach is more suitable for single-region studies. In case of Tinoco and Wilson (2013), the Retail Price Index and UK 3-month T-bill rate cover inflation and interest rate on macro level.

Stress testing on individual-level data is not common for stress testing literature — most of the studies are oriented towards evaluation of the whole banking sector (e.g., Covas, Rump, and Zakrajsek, 2014) or specific portfolios (Basurto & Padilla, 2006). Therefore, the CDT modelling as a separate step has more potentially, especially for multi-region studies when different set of variables or their lag structure can be potentially used for different regions separately. The construction of such model and fitting it into the loan-level specification ("the satellite model") can follow multi-step regression analysis where first step is to estimate CDT for each country/region as

$$CDT_t = L^{-1}[\alpha + \mu M_{t-l}] \quad (14)$$

Where lag structure *l* allows modeler to control for delayed effects in the market when changes in default rate levels are not following the environment adjustments immediately. Then, the CDT will be used in specification, similar as above.

$$Y_{it} = L^{-1}[\alpha + \beta X_i + \gamma Z_{it} + \theta F_i + \eta CDT_{it-1}] \quad (15)$$

The advantage of such setup is that the CDT can be constructed individually for each country/region, while transaction-level model can be pulled across all regions, establishing advanced robustness as showed in the previous

sections. Example of such approach can be found in Cihák (2007) and Foglia (2009).

Accounting for macroeconomic factors provides stress-testing feature to the analysis while also improving robustness over longer timeframe datasets. However, the multi-step nature of approach might complicate the evaluation of the model as errors from the first step can bias final outcomes and cause results deterioration.

3.5 Concluding remarks for empirical analysis

This study expands the assessment of credit risk for Small and Medium Enterprises (SMEs) by exploring various definitions of credit risk events and their impact on model performance, particularly focusing on model robustness across various timeframes and regions. It answers the predefined methodology (M3) and empirical (E1-E5) questions by testing whether SME credit-risk models remain stable when the dependent variable is redefined across alternative forms of credit deterioration. The analysis concluded that while traditional default definitions provide clear insight and can demonstrate adequate performance, alternative definitions may offer better insights into early warning signs of credit risk, especially when robustness of model is considered. While incorporating financial data significantly enhances model performance, thorough evaluation of credit risk (including additional factors) is important for SMEs' assessment.

The findings demonstrate inter-SME size effects across multiple definitions suggesting that micro and small SMEs experience higher credit risk compared to medium-sized firms. The use of multiple definitions enhances the

confidence of cross-definition confirmed effects such as importance of financial factors, importance and strong magnitude of interest rates and leverage indicators, while question findings which face controversy across the definitions boards such as loan term. The robustness evaluation using in-time testing sample, out-of-time and after-COVID-19 samples demonstrated that robustness of one definition (12-month default) has its dichotomy with overall performance when compared to other definitions (standard default and delinquency). Those metrics have primary theoretical connection to confidence in estimates, which can translate into reduced lending when model power drops. Thus, one can explore whether considering multiple approaches and combining them in a more systematic framework can help with better decision-making.

These findings support idea that model outputs more relevant for monitoring and intervention, not just ex post classification. Recent empirical studies point in the same direction: bank–firm relationship variables improve SME default prediction beyond accounting data while non-financial information improves collective assessments when used together with financial variables; and event-based or network-based models can identify the propagation of negative signals through firm ecosystems (Modina et al, 2023).

The robustness of the models after the COVID-19 pandemic is less expected: while credit risk patterns are supposed to be different (accounting for change in financial health and lending behaviors among SMEs), it had limited impact on previously concluded findings, highlighting the validity of outcomes. However, it brings an important result of redundancy of model reevaluation and

highlighting the fact that fundamental information is reliable even under changing circumstances.

Cross-country pooling improves inference when events are infrequent, provided that pooling can be interpreted as a trade-off between statistical power and heterogeneity. The results for Spain and Italy are therefore valuable not only because they enlarge the sample, but because they show that shared SME risk patterns can be estimated more reliably when event counts are low. Cheraghali & Molnar (2024) similarly argue that SME default research needs more cross-country designs and more explicit treatment of institutional and macroeconomic heterogeneity, rather than if single-country evidence can be generalized into region-level evidence by theoretical justification.

As the last point, one can argue that the results do not state a single, “best” definition of event or approach. Instead, there is evidence of a benefit from adding extra factors and combining financial and loan characteristics. This leads toward a fact that further advances in risk assessment are likely to stem from expanding and improving the data environment itself rather than refining model architecture alone. Naturally, it motivates a broader methodological inquiry into how additional, unstructured or weakly observable information can be systematically incorporated into credit risk frameworks. The next section extends the analysis by examining existing limitations and the role of Generative AI as an enabling mechanism for transforming such information into usable modelling features. The discussion positions GenAI not as a corrective tool, but as a continuation of the data-centric perspective emerging from the empirical results.

4 Discussion

4.1 Limitations of empirical part

This study achieves its primary objectives of enhancing the robustness of predictive of SME credit risk assessment through more granular credit event definitions and enriched modelling approaches. Nevertheless, several limitations and areas for further refinement remain in place, which should be considered when interpreting the results and extending the framework in future research.

One of the previously mentioned reservations lies in reliance on first default and discrete event frameworks. While the study intentionally explores multiple definitions, these constructs remain more focused on signals of underlying credit deterioration rather than fundamental processes evolution and business survival. The extension through of a macro-financial framework can help with accounting for structural breaks and evolving risk drivers. At the same time, hazard models can address the limitations of discrete event frameworks by capturing the continuous evolution of default risk over time.

The use of first-default approaches, although methodologically consistent with regulatory practices, introduces potential survivorship bias by excluding post-default dynamics and repeat distress cycles that are particularly relevant for SMEs. This may limit the ability of the models to capture cyclical vulnerability and recovery patterns, especially in setups characterized by restructuring or forbearance.

Another limitation relates to data coverage and feature construction. Due to data constraints, the empirical analysis has necessarily been simplified,

limiting the extent to which the study can fully test the hypothesis that SMEs exhibit structurally distinct risk characteristics compared to retail and large corporate exposures. In particular, the available data does not allow for a sufficiently granular segmentation across firm sizes and corresponding financing nuances, which restricts better validation of the proposed modelling framework.

A key suggestion for further empirical research would be a comparison with micro-enterprises, which are typically more reliant on retail-type (i.e., owner-driven) lending dynamics. An important question in this context is whether the transition from micro firms to SMEs is gradual and continuous, or whether it exhibits a structural break in risk drivers and model behavior due to different driving factors. A similar analysis could be extended to the transition between medium-sized firms and large corporates. However, conducting such assessment in a robust manner would require more granular and higher-quality datasets, which were not available within the scope of this study.

These limitations point toward the potential role of Generative AI-enabled data augmentation as a promising extension. By systematically extracting, structuring, and enriching information from unstructured sources, GenAI approaches can help mitigate data sparsity and improve the observability of SME factors. In this context, the integration of GenAI-driven feature engineering offers a solution to bridge existing data gaps, enabling more comprehensive testing of structural hypotheses and enhancing the robustness of SME credit risk models.

4.2 Generative AI in SME credit risk assessment: a European perspective

Generative AI (GenAI), particularly large language models (LLMs) like GPT-4, has emerged in the last two years as a transformative technology. These models are oriented towards processing unstructured data from financial reports, news and other sources to generate human-like outputs, offering new ways to evaluate creditworthiness beyond traditional quantitative models. Financial institutions and fintech companies are exploring how GenAI and agentic AI[3] might enhance credit risk modelling.

While the GenAI adoption keep increasing, the recent survey from KPMG (2024) showed that only 12% of organizations are building GenAI solutions on their own. At the same time, the use cases can cover the full credit life cycle – from prospecting, then to underwriting, and finally, to portfolio monitoring and risk management. However, the exact implementation of GenAI received data is still in early stages and require more fundamental review.

LLMs have a great improvement for SME lending, where speed and efficiency often break a loan's profitability and affect SMEs access to finance. Shi et al. (2023) proposes a hybrid model that leverages LLMs to extract relevant information from unstructured textual data to assess SME creditworthiness, combined with traditional scoring. While larger banks are exploring these capabilities as process automation (co-pilots, chat bots), more fintech lenders start utilizing LLMs to analyze open banking data such as

[3] AI agents are AI-backed programs that "operate autonomously, perceive their environment, persist over a prolonged time period, adapt to change, and create and pursue goals" (Russell & Norvig, 2021, p.4)

transaction descriptions, cash-flow patterns and industry reports about an SME's sector, yielding features for credit models that were previously unavailable.

This paper proposes a framework for deploying GenAI in SME credit risk assessment, focusing on conceptual quantitative modelling approach. The goal is to demonstrate how GenAI can be harnessed for more accurate, efficient, and inclusive SME lending by academic researchers, fintech executives, and policy professionals – while managing the attendant risks and compliance requirements. The proposed modelling structure can be used to evaluate multiple potential effects such as impact of news, innovation, supply chain disruption, political exposure and other through a new lens which offers high potential to improve research and understanding of SME credit risk phenomena.

The paper is structured in a following way. First, we start with a literature review of the ongoing development of GenAI in SME credit risk assessment and evaluate integration of these technologies into practice. Consequently, we comment on European regulatory and ethical considerations from perspective of GDPR and AI Act. Finally, we suggest a quantitative framework to augment GenAI-infused data into modelling exercise to ensure high reliability and transparency that would (a) provide better performance; (b) remain compliant with existing regulations; (c) allow researchers and analyst comment on effects of both classical data fields as well as proxied via LLMs factors.

4.3 Motivation

4.3.1 Challenges of SME credit risk assessment in Europe

Assessing SME credit risk has long been recognized as one of the most challenging tasks for lenders and policy makers. Unlike large corporates, SMEs often lack audited financial statements or credit ratings, resulting in information asymmetry (Calabrese et al., 2021). Limited financial disclosure and data collection gaps are key challenges for SME risk evaluation due to those incomplete or short-history financial records, especially when economic conditions are not favorable (Ghulam et al., 2025). Opaqueness means banks have difficulty estimating their true creditworthiness. In practice, credit assessors rely on proxies (for example, a credit score or a collateral value) or manual due diligence. Thorough SME credit underwriting means evaluating business plans and financial projections as well as conducting site visits which are labor-intensive. Consequently, the cost of manual assessment for small loan amounts can be disproportionately high, leading to many SMEs being rationed out through direct decline or charging extremely high rates because lenders can't efficiently analyze them. A World Bank study noted that because of information gaps, many banks struggle to assess SME credit risk, often resorting to relationship-based lending and collecting soft information from client interactions (Abraham & Schmukler, 2017)

Another important difficulty is associated with SMEs heterogeneity. Those firms can be a highly diverse group represented by small local shops, high-growth start-ups, or medium size regional enterprises. Standardized scoring models can be less effective for such heterogeneity. Sector, region, and

firm-specific factors can significantly impact risk, making one-size models less predictive (Dietsch and Petey, 2004). Lenders face challenges incorporating alternative data such as industry outlook, online sales, and supplier payments to capture these nuances.

While smaller size can be perceived as advantage due to firms' flexibility, it often turns its back in a form of vulnerability when SMEs are more affected by macroeconomic swings and shocks. Recent data shows a rise in European SME insolvencies, for instance, the UK sector has risen by 16% year-on-year in mid-2024 as shown by the Insolvency Service (2024) due to higher input costs and higher interest rates. Volatility usually brings challenges for risk modelling – default probabilities can change rapidly with the economic cycle, requiring dynamic risk assessment.

All those are only amplified by regulatory pressure due to capital constraints. Under Basel and EU banking regulations, banks must hold capital against SME loans based on risk. Uncertainty or lack of data often forces conservative risk estimates and higher capital charges. This in turn reduces banks' willingness to lend to opaque SMEs or providing lenders excessive market power, contributing to the well-documented SME financing gap (Wang et al.,2020).

In summary, European SMEs face a data disadvantage in credit markets – they are risky in perception partly due to lack of transparent information. The result is a finance gap where otherwise viable businesses struggle to obtain credit. These challenges set the stage for AI-driven solutions: if new technologies can better utilize available data (including unstructured data) and produce more accurate, explainable risk insights, they could

materially improve SME financing. Also, it might help policy makers to have better understanding of SME's access to finance, define target groups for polices and support programs, and more efficiently monitor their outcomes.

4.3.2 Advancements in Generative AI and LLMs for risk modelling

Application of AI technologies is not exactly earthshaking – credit risk methodologies have been gradually absorbing new algorithms and computational advantages over past decades. Inclusion of machine learning and AI models has shown its benefits through performance gains and automation of processes, outperforming the traditional models in predicting SME credit risk. For instance, Bitetto et al. (2024) studied Italian SMEs to compare traditional probit regression with machine learning alternatives. The study clearly demonstrates how fintech lenders can rely on such scoring for small business credit decisions, even in imperfect conditions. At the same note, Stevenson et al. (2021) introduced unstructured textual data into credit risk assessment of SMEs. Utilizing more than 60'000 loan reports that were manually filled, they built a model that were using textual and structured financial data. While unstructured part alone was quite efficient predicting defaults on its own, it didn't bring much of improvement when combined on top of financial metrics model, implying that additional qualitative information might be redundant in presence of robust data.

Recent breakthroughs in AI – especially generative models – offer powerful new tools to the challenges. Large language models such as OpenAI's GPT-3.5 and GPT-4, Google's PaLM, and domain-specific models like

BloombergGPT (a large language model trained on financial data, Bloomberg, 2023) have demonstrated remarkable capabilities in natural language understanding and generation. As credit risk analysis involves a mix of numbers and narrative, these LLMs can boost evaluation by analyzing unstructured text and extracting insights, which can further be used for automated report generation. These novelties provide new data sources that were not present previously and could not be utilized by researchers to evaluate potential impact or relevance.

As part of this study, there are five key aspects that contribute to improve credit assessment of SMEs: efficiency, automation, augmentation, specialization, and variability. The former one is the most Known due to increase in efficiency through natural language understanding. Modern LLMs can read and comprehend complex documents and distill key facts about companies whether it is annual reports, news alerts, or business plans. This effectively enables extra channel of passing qualitative information within risk assessment models.

One of the examples is the after-COVID19 supply chain disruption. The shortage of semiconductors was affecting SMEs that were using those in their processes. Traditional models usually ignore such qualitative information or consume it too late, when the damage has already propagated through financial performance indicators. Evaluating supply chain risks requires deep assessment of management practices and comprehensive analysis of vulnerabilities without which it won't be possible to predict disruptions (Xiong et al., 2024). Generative AI systems use LLMs to combine, summarize, and analyze unstructured data and output the results in natural language or

structured forms (Devlin et al., 2019). This means a GPT-based system might go through all information much faster than a human-involving system (such as manual qualitative overlay), flagging any concerns that would be relevant for the evaluation.

As the second contribution stands, the application of agentic AI for data retrieval supports modelling with more automated and complete data collection. In this context, agentic AI refers to AI systems that can perform multi-step reasoning or actions autonomously without replicating simple steps manually. In practice, LLMs have been augmented with use of tools – the ability to call external APIs, run code, or query databases (Gautam, 2023). In a risk context, an LLM could automatically extract the latest data and perform a quick calculation as part of its analysis to reduce errors and ensure the narrative is backed by most relevant computations. The design of such workflows brings extra benefits for those processes that are repetitive, simplifies tasks of data retrieval from various sources and its pooling – being essential enhancement for SME data collection, where relevant data might be scattered.

Third piece is related to data transformation as augmentation techniques can help with deriving better proxies for missing information. Generative AI isn't limited to text; models like generative adversarial networks (GANs, as described in Goodfellow et al., 2014) and diffusion models (Song et al., 2020) can create synthetic data oriented towards enhancing scarce training data for SME credit models, for example, through augmentation of limited default examples for a new SME segment. The UK's Financial Conduct Authority has noted that synthetic data can simulate customer scenarios, allowing risk models to be trained and tested under more varied conditions,

including hypothetical scenarios and stress-testing analysis to generate how an SME's finances might look in a recession.

Accounting for domain-specific context helps a lot with specialized research, thus forth benefit is formed around localization. Finance-specialized AI models have been in focus of both market leaders and financial institutions. BloombergGPT, for instance, has demonstrated superior performance on financial NLP tasks by training on financial news, filings, and market data (Bloomberg, 2023). Similarly, open-source projects like Financial GPT aim to democratize LLMs for finance by tuning models on domain data (Liu et al., 2023). These models are better at understanding finance jargon, interpreting numerical context, and adhering to the precise language needed in financial documents. In credit risk modelling, such domain-specific LLMs can more accurately interpret an SME's context.

All previous angles combined enable the variability, the fifth and final element of the sequence. By handling multimodal inputs covering text, images, audio and other types of information, the latest AI systems are expanding types of data that can be used. From perspective of data collection, it improves frequency, coverage and recency of input factors as well as improve chances to account for all critical factors and avoid omission bias in the model. Additionally, Generative AI provides an inductive complement to traditional deductive methods. Traditional credit models apply predefined formulas to data; generative AI inductively derives insights from patterns and language.

4.3.3 How GenAI methods compare to existing methodologies?

The combination of both conventional and GenAI techniques could yield a more holistic risk assessment. Table 18 compares key considerations

of the traditional method and GenAI-enhanced approach to illustrate this synergy. The traditional SME assessment section is driven by literature review, that has been previously summarized in Pleshkevich & Han (2025).

Table 18. Traditional vs. Generative AI-augmented approaches to SME credit risk assessment.

Aspect	Traditional SME credit assessment	GenAI-enhanced SME credit assessment
Data Scope	Primarily structured data such as financial ratios, credit scores, collateral values. Unstructured data largely excluded or manually assessed.	Enables unstructured textual data beyond financial statements to derive credit sentiment from news, reports, and other sources. Can be extended into synthetic data.
Analysis Method	Statistical models on numeric features; deterministic rules for policy checks. Qualitative information propagated through overlays that are largely calibrated through expert judgement.	LLMs for text understanding and summarization; ability to find complex patterns and interactions in data and translate them into quantitative metric. Agentic AI automates analysis by multi-source data gathering and interpretation.
Model Output	Credit score or default probability. Explainability is driven by input factors which are usually inspired by classic risk literature and relevant from compliance angle.	Traditional output can be accommodated by narrative output: analytics in natural language, including rationale and references to source data. Additional engines augment

		input with “human-like” logic to justify results.
Bias & Fairness	Quantitative models rely on carefully selected variables to avoid bias, potential human bias in qualitative assessment.	Reduced subjective bias by evaluating textual data through the same LLM. Maintaining same LLM over time might be a challenge. Requires careful tuning and robustness checks to avoid learned biases. Requires built-in bias detection and debiasing techniques are applied (excluding protected attributes in prompts). (Chen et al., 2024)
Compliance & Validation	Requires standardized validation exercise oriented towards model performance and robustness. Usually, performed as per guidelines and policy documents.	Requires extensive validation and accuracy monitoring to establish objectivity and reliability of model. Governance guidelines are very limited. Compliance is a grey zone and requires extra caution as “hallucinated” results might bring economic losses.
Scalability & Adaptability	New data sources or rules require manual model redevelopment. Scaling qualitative analysis requires extra human resources.	Highly scalable via automation. New data pieces can be ingested by LLM-based systems with minimal

	Monitoring can be done by automated triggers but requires human involvement.	reprogramming (prompt engineering). GenAI agents can continuously monitor data streams (news feeds, transactions) and adapt risk flags in near-real-time.

While the comparison obviously highlights the technological benefits by incorporating a significantly large amount of new information into credit assessment, there are 2 key limitations that should be treated with utmost care. Firstly, the use cases are still emerging, and adoption requires a large amount of trained data scientist and AI experts to design efficient and transparent systems. Secondly, the use of LLMs requires additional governance from compliance perspective which is still evolving and might be subject to change within short period of time. The next section reviews the regulatory nuances of LLM use in modelling.

4.4 Proposed framework for GenAI-powered SME credit assessment

Building on the literature and use cases above, we propose a conceptual framework for applying Generative AI to SME credit risk assessment. The goal is to leverage GenAI's strengths – especially handling unstructured data – while maintaining standards of accuracy, robustness, and compliance. The framework consists of three layers: data augmentation, quantitative model, and output interpretation. Next subsections describe each piece sequentially.

4.4.1 Data layer

Data ingestion and preparation are fundamental to aggregate all relevant information about the SME. Foundation of the system requires a robust data pipeline that would organize various types of data, transform it into model-ready format, test statistical properties to check relevance of created metrics, and then re-estimate the previous, GenAI-agnostic models to check the impact and change in sensitivities of all considered factors.

Generally, the data can be classified into two large domains: structured and unstructured. Financial statements, credit bureau data, associated account transaction data, and any performance history with the bank would fall under structured data. This data is already formatted and can be used for univariate or/and multivariate analysis. Unstructured data can be associated with credit bureau or company registry info, news articles, sectorial reports, supplier credit data, social media sentiment, financial accounts outlines, business plans and strategy reports, tax returns, invoices, communication between the enterprise and the bank. To make it machine-readable, one would need to use OCR and translate information into text that can be run through GenAI preparatory step. At this stage, it is important to keep all data entries timestamped to account for data appearance in the storage and unify for credit sentiment analysis including classifying whether information is credit-relevant, type of information (credit improving or deteriorating) and tags each piece with metadata (type, date, source) for context and relevance.

The data layer part of the framework focuses on formulating information that is required for model construction. Like classically observed data X_i, for each firm i one can extract a set of GenAI-derived features $Z_i =$

$\{Z_1, Z_2, .. Z_J\}$. Since information is being derived and not observed, we assume that there is a level of uncertainty with which this information appears to be true. Based on confidence that information corresponds to firm i, one can define confidence levels $C_i = \{C_1, C_2, .. C_J\} \in [0,1]$. The confidence level can be characterized by reliability of initial source of information, cross-matching information from multiple sources, and prompt self-consistency across augmentations (i.e., language model accuracy for specific feature information collection). To define this more formally, the feature generator can be set as: #(1)

$$Z_i = \tilde{Z}_i + \epsilon_i = G(T_i, \psi) + \epsilon_i \quad (16)$$

Where Z_i is an estimate of latent true $\bar{Z}_i$, and can be denoted as function G (LLM) of raw textual inputs T_i (as defined within the section above) and parameters ϕ. Rewriting from as a conditional probability:

$$\mathrm{Z_i} \sim \mathrm{P(\,Z \mid T_i; G, \psi\,)} \quad (17)$$

Similarly, confidence level can be defined as estimated posterior about informativeness of features Z_i:

$$\mathrm{C_i} \sim P(\,Z_i\ is\ informative \mid T_i; Z_i; G, \psi\,) \quad (18)$$

Estimating confidence levels is not a straightforward step. If historical information or manually verified data subsample is available for derived data, this can be evaluated statistically by building meta-model from Z_i and $\tilde{Z}_i$. However, this is extensive exercise and won't be available in general setup. Instead, one can use LLM own outputs to put a label on received sentiment from each textual piece T_{ik}, classifying it into credit positive or credit negative, evaluate entropy across observed sentiments and translate it into sample consistency:

$$C_i: P\left(label_{ij} \mid H(T_i, Z_{ij})\right) \rightarrow (0,1) \quad (19)$$

Confidence levels are very important as lack of trust is one of classical criticism of GenAI applications (Magesh et al., 2025). Additionally, one can use the combination of confidence levels to evaluate the observability by constructing observability score (W_i) as a function of observed information X_i for actual information and confidence levels C_i for derived fields:

$$W_i \sim P(\, i \; is \; observable \mid X_i, T_i, C_i, G, \psi\,) \quad (20)$$

The very basic definition of observability in digital context can be linked to the fact that a sufficient number of features can be informative (i.e., demonstrate high confidence levels) for this type of information to appear relevant. Additionally, inclusion of actual information is important as factual and firmographic information like sector, region, size, and coverage flags may inform whether firm is expected to be observed in the first place. For example, smaller firms from less developed countries would probably be less covered than larger firms from countries with better media coverage. Later, observability scores can be used as part of the model to define whether derived information is material for credit risk.

Textual information plays key role in making features and observability scores relevant for the quantitative model. To make the evaluations robust, the information pieces need to be transformed accordingly and checked for empirical relevance. We are summarizing potential additional GenAI-derived synthetic data with the empirical evaluation strategies in the Table 19.

Table 19. GenAI-augmented data metrics that can be used for SME credit risk assessment

Feature Name	Data Source Example	Description	Empirical Evaluation
Reputational Risk Index	News articles, social media, review platforms	Sentiment score over time (for instance, 12-month average and volatility)	1. Test Granger causality with default events. 2. Include in logistic/ Cox model. 3. Evaluate AUC ROC when included.
Business Model Complexity Score	Company websites, whitepapers, brochures	Embedding-based LLM assessment of product or service variety as well as value chain depth	1. Use embeddings (like Sentence-BERT as described by Reimers & Gurevych, 2019) to score complexity. 2. Correlate with default events in multivariate regression model and test incremental predictive power
Stakeholder Communication Clarity	Media texts such as interviews, website official pages, regulatory filings	Measures coherence, readability, and consistency in SME's public narrative	1. Use Natural Language Processing tools (such as Coh-Metrix, as in Graesser et al., 2004) and GenAI to evaluate. 2. Compare across defaulted and healthy firms. 3. Use model selection methods like LASSO to check if it should enter final model

Litigation Signal	Legal databases, press coverage	Binary or count of legal disputes or regulatory actions	1. Test as lagged binary indicator. 2. Validate with default data using survival analysis or out-of-time AUC ROC lift
Digital Visibility Index	Website age, frequency of content change, social engagement	Proxy for market activity and digital reputation	1. Construct time-series signals through Principal Component Analysis 2. Use model selection methods like LASSO to check if it should enter final model 3. Test robustness to sector and size heterogeneity
Supply Chain Disruption Mentions	News, supplier reviews, industry reports	Proxy for delivery disruptions and reliability risks	1. Track frequency via Named-entity recognition (NER) 2. Test predictive power of disruptions prior to distress using lagged variable (short-term, up to one year)
Innovation Signal Score	Patent filings, press releases, R&D mentions	Proxy for resilience via product pipeline and differentiation	1. Include as binary/continuous factors 2. Test predictive power of innovations to reduce distress using lagged variable (short- or mid-term)

Importantly, GenAI augmented data can be used not just to inform explanatory factors but also to detect credit distress. The enhanced indicator would be accounting not just for default realization or change of legal status

within database, but also to news articles (mention of bankruptcy filing, support enquiry, or legal action) and litigation signals. Both contributions are expected to be more forward-looking and provide indication before actual default realization.

Once data is prepared, the LLM application is the final part of data transformation. LLM analysis processes the unstructured data to retrieve specific relevant chunks of text from a vector database when answering queries, ensuring it grounds its output in actual data. Notably, the LLM in this architecture should operate within a controlled environment, limiting its assessment on the provided information with minimum hallucinating with exogeneous facts and it follows strict credit-specific prompts. This can be achieved by setting LLM temperature to 0, meaning it will be fully deterministic in its outputs. It also implies that agentic behavior is permitted only for chosen tools and within set bounds of information search.

Similarly to classical data distribution analysis, the preliminary analysis and descriptive statistics of derived GenAI features is an important contribution to the field on its own. The correlation between traditional and augmented data fields can demonstrate whether new data variables bring extra potential explanatory power. Also, if the information is available over a multiple time periods (panel data structure), one can test whether GenAI derived data is more reactive and whether it can be translated into change of actual performance ratios. This can help to identify and reduce cross-correlation within future model design.

The LLM outputs several items including (i) extracted features – key risk indicators gleaned from text, such as sanctions and politically exposed

persona flags, supplier chain concentration, management experience, cyber security evaluation and other, and (ii) a narrative risk assessment – a summary of processed information regarding SME. These can be turned into quantitative features that feed into the credit model via direct translation into qualitative score or used as a flag for specific events. This type of inputs is specifically useful for coarse classing type of models. Next section discusses the exact specifications of the modelling.

4.4.2 Modelling layer

Due to anticipated complexity and requirement to ensure transparency of GenAI infusion, we outline a multi-stage model that can separate LLM analysis from classical econometric model by creating hybrid "LLM + traditional" models. At the core of the hybrid modelling is an approach that combines generative AI components with traditional predictive models within the specification.

At first, GenAI derived data can be viewed simply as additional source of extracted data and can be tackled with a straightforward regression analysis to evaluate direct impact of GenAI features:

$$P(Y_i = 1) = F(X_i, Z_i) \tag{21}$$

This can be, for instance, classical logit or probit function and can be selected through classical factor-selection algorithms like stepwise regression or LASSO. Altman et al. (2023) proposed an improved framework called Omega Score, which combines more than 160 variables (qualitative, financial, management-related, and other) into one metric. Like Altman's Z score (1968), this approach oriented towards defining best set of weights across factors to

maximize accuracy. Adding additional synthetic features can add more variation within explanatory factors and improve the Omega score with extra data points.

At the same note, the derived features can contain additional noise, be unavailable or less trusted for observations. Thus, the proposed credit risk model will use the structured approach by combining financial data with prepared GenAI proxies into a traditional credit risk model. This could be an individual model like logistic regression scorecard or an ensemble machine learning model with multiple basic learners; the standard solution would suggest using ML models as they are easier to interpret while providing solid improvement over standard scorecards (Bitetto et al., 2024).

4.4.3 Integrated risk evaluation

The outputs of the scorecard model and the LLM engine can be combined through ensemble model. In this case the setup can look like:

$$P(Y_i = 1) = g(p_i{}^{Scorecard}, p_i{}^{GenAI}, W) = w_i \cdot F_1(X_i) + (1 - w_i) \cdot F_2(Z_i), \quad (22)$$

where $p_i{}^{Scorecard}$ is classical scorecard model outcomes, $p_i{}^{GenAI}$ is LLM-focused model outcomes, and w_i are the associated weights that are calibrated using observability inference.

The multi-stage approach requires splitting modelling dataset in multiple subsets to train the base learners, to train ensemble model, and to validate the final outputs. Thus, this approach is quite sensitive to number of events in the data and require enough observations.

Alternatively, integration could be achieved through a simple rules-based overlay or a second-level model. It can be used as downgrading

mechanism, when if the LLM's sentiment scores are negative, the model adjusts the final risk grade downward even if financials looked okay. The calibration is done based on sensitivity of unexplained default frequencies towards GenAI-augmented factors.

$$P(Y_i = 1) = \alpha_0 + \hat{Y}_i'\alpha_1 + \widehat{\Delta Y}_i'\alpha_2 \quad (23)$$

$$Y_i = logit^{-1}(\beta_0 + X_i'\beta) \quad (24)$$

$$\Delta Y_i = Y_i - \hat{Y}_i = logit^{-1}(\gamma_0 + Z_i'\gamma) \quad (25)$$

In this case, regression analysis takes the traditional PD and selected NLP-derived features to output a final PD. The specification would assume LLM fields as additional factors to improve overall accuracy within unexplained variation.

Those specifications assume that all firms that are coming from traditional data sources can be efficiently analyzed by GenAI tools. However, this is a strong assumption as one of the key challenges related to SME data is its observability. Thus, unlike sectorial or pooled information, firm-specific augmentation will be relevant only to a certain subset of entities. As a result, observability (by GenAI tools) is extra layer that should be incorporated into modelling.

4.4.3.1 Multi-regime model

To account for this challenge and expand the previous model, one may suggest using switching regression approach with a latent indicator. Let us define $S_i \in \{0,1\}$ as an unobserved indicator:

- $S_i = 1$: firm is observable → GenAI-augmented model
- $S_i = 0$: firm is not observable → Baseline model

For SME i, we observe (Y_i, X_i, Z_i), but Z_i may be meaningless if $S_i = 0$. Thus, the distress probability would be state-dependent and appear like:

$$P(Y_i|X_i, Z_i, S_i) = \begin{cases} P_1(Y_i|X_i, Z_i) \; if \; S_i = 1 \\ P_0(Y_i|X_i) \; if \; S_i = 0 \end{cases} \tag{26}$$

where P_k are logit models for each state:

$$P_1(Y_i = 1) = logit^{-1}\left(\beta_0^{(1)} + X_i'^{\beta^{(1)}} + Z_i'^{\gamma^{(1)}}\right) \tag{27}$$

$$P_0(Y_i = 1) = logit^{-1}\left(\beta_0^{(0)} + X_i'\beta^{(0)}\right) \tag{28}$$

Assume S_i depends on an observability score W_i which can be described as likelihood of firm's information to be present in the digital space:

$$\pi_i: P(S_i = 1|W_i) = logit^{-1}(\delta_0 + W_i'\delta) \tag{29}$$

Then we can define mixture model marginal log-likelihood over n firms as:

$$L = \sum_{i=1}^{n} log[\pi_i * P_1(Y_i|X_i, Z_i) + (1 - \pi_i) * P_0(Y_i|X_i)] \tag{30}$$

As in case of classical regime switching regression, EM algorithm (Dempster et al.,1977) can be applied iteratively to estimate the parameters. In that paradigm, we refer to X as observed data (financials, loan information, macroeconomic or sectoral indicators), while Z is asset of "latent" data that can be derived from digital sources, provided firm is sufficiently disclosed in those sources. We estimate $\Theta = \left(\beta^{(1)}, \gamma^{(1)}, \beta^{(0)}, \delta\right)$ by maximizing L.

E-step is to compute the expected regime probabilities (posterior of S_i), given current parameters $\Theta^{(t)}$:

$$\hat{s}_i^{(t)} = \frac{\pi_i^{(t)} * P_1\left(Y_i\middle|X_i, Z_i; \Theta^{(t)}\right)}{\pi_i^{(t)} * P_1\left(Y_i\middle|X_i, Z_i; \Theta^{(t)}\right) + \left(1 - \pi_i^{(t)}\right) * P_0\left(Y_i\middle|X_i; \Theta^{(t)}\right)} \tag{31}$$

M-step is to maximize the weighted complete-data log-likelihood:

$$\sum_i \left[\hat{s}_i^{(t)} * logP_1(Y_i|X_i, Z_i) + \left(1 - \hat{s}_i^{(t)}\right) * logP_0(Y_i|X_i)\right] + \\ + \sum_i \left[\hat{s}_i^{(t)} * log\pi_i + \left(1 - \hat{s}_i^{(t)}\right) * log(1 - \pi_i)\right] \quad (32)$$

This involves (i) logistic regression of Y_i on (X_i, Z_i) weighted by $\hat{s}_i$ to get $(\beta(1), \gamma(1))$; (ii) logistic regression of Y_i on X_i weighted by $1 - \hat{s}_i$ to get $\beta^{(0)}$; (iii) logistic regression of $\hat{s}_i$ on W_i to get δ. E and M steps should be repreated until convergence is reached for preset threshold.

4.4.3.2 Defining hypothesis

There are a list of more detailed questions that empirically can justify the viability of methodology. To answer research question M5 there are a list of principles that need to be tested: whether augmenting data improve catch rate of credit risk signals, whether it creates additional noise, whether it can recognize signals better and faster, whether fundamental factors which are there in conventional models are actually useful for early distress. Table 20 demonstrates relevant regime-switching tests and associated hypothesis.

Table 20. Tests and expected outcomes for empirical GenAI-augmented study

Hypothesis	Property to be tested	Regime 1 (Observable)	Regime 0 (Shadowed)
H1	Sensitivity (TPR)	Higher as Z component detects alleged distress signals	Typically, lower
H2	Specificity (TNR)	Potentially lower as may overreact to noise	Typically, higher

		from GenAI component	
H3	Model Accuracy (AUC, ROC)	Possibly higher if Z is informative	More stable but less informative
H4	Lead Time of Prediction	Longer: GenAI component is expected to provide early warnings	Shorter due to lagged reaction of indicators
H5	Coefficient Estimates	$\gamma^{(1)} \neq 0; \beta^{(1)}$ may shrink as more potential explanatory variables	$\beta^{(0)}$ larger to compensate the unobserved effects
H6	Frequency of Distress Events	Higher due to additional information sources	Lower: limited to actual defaults

4.4.3.3 A Bayesian switching

While classical regime-switching models do account for multiple latent states, the regime assignment is a fixed process and lack robust mechanisms for parameter uncertainty or rare event handling in a practice. A Bayesian approach, by contrast, introduces a latent regime variable with full posterior inference, allowing the model to dynamically allocate firms into an "augmented" regime. This not only enables conditional model complexity but also prevents overfitting by shrinking parameters through informative priors (Baltodano López et al., 2024; Redner & Walker, 1984).

In the Bayesian setup, priors' distributions are required to be set – that is where confidence levels can be utilized to define higher variance σ_j for

higher confidence C_j of feature j. This hierarchical approach implies looser priors and higher variance of potential parameters, so that model can utilize stronger features more. Similarly, if confidence levels are estimated through meta-model, hyperparameters can be informed by empirical estimation of variance from the subsample.

The Bayesian framework is particularly valuable for SMEs as it can accommodate informational asymmetries and better capture the episodic nature of default clusters (Berentsen et al., 2022). Moreover, this type of estimation offers extra tools for updating beliefs as new signals emerge which make the model more adaptive to the volatile realities of SME environments (Ptak-Chmielewska & Kopciuszewski, 2022). These properties also enhance the early warning system potential of the framework as Bayesian models generate predictive distributions, not just point forecasts, allowing users to flag firms with rising uncertainty even before sharp distress indicators appear (Berloco et al., 2023).

From a regulatory standpoint, this approach aligns closely with emerging expectations under the EU AI Act. Bayesian regime-switching enables explicit modelling of uncertainty, allowing feature-level transparency through probabilistic attribution (Weng et al., 2025), and facilitates auditability by exposing the full reasoning path – from prior assumptions to posterior outputs. Unlike opaque black-box models, this structure supports compliance with AI governance standards on fairness, explainability, and risk-based classification. As Umavezi (2025) emphasizes, models that yield probabilistic forecasts with transparent parameter uncertainties are better suited for critical decision-making domains, including credit allocation and stress testing. Thus,

the Bayesian regime-switching approach offers both a methodological advantage for risk modelling under data heterogeneity and a governance advantage under the tightening lens of regulatory scrutiny.

4.5 Regulatory and ethical considerations in Europe

4.5.1 Compliance side of GenAI-modelling

Compliance side of the credit risk assessment is an essential point for any lending institution. Expectedly, any use of AI in credit risk assessment must contend with strict regulations and ethical expectations, especially in Europe. Credit decisions affect both individuals' and businesses' access to finance, raising concerns about fairness, transparency, and accountability. The European regulatory landscape in 2024 is being fundamentally shaped by two major frameworks: the General Data Protection Regulation (GDPR, 2024) and recent EU Artificial Intelligence Act (AI Act, 2025). Additionally, financial regulators and central banks are continuously monitoring modelling space and providing guidance on AI governance.

GDPR Article 22 gives individuals (which can include sole proprietors or very small businesses) the right not to be subject to decisions based solely on automated processing that have significant effects on them, unless certain exceptions apply. In a landmark December 2023 case involving the German largest credit bureau Schufa, the Court of Justice of the EU (CJEU) ruled that generating a credit score – a probability of default – constitutes an automated decision subject to Article 22 (CJEU, 2023). The case confirmed that if a credit score materially influences a lending decision, the subject has rights to inquire information about the logic involved and to contest the decision. The CJEU

emphasized that calculating a score qualifies as decision instead of a preparatory step for taking one, thus sufficient explanation of factors is required, and that trade secrecy cannot override data subject rights in such cases (CJEU, 2023).

For lenders, this means any AI-driven credit scoring or GenAI recommendation that leads to denying an SME credit must be accompanied by an explanation and a human review upon request. It does not mean that all automated scoring is outright banned, but it does mean controllers must: (i) avoid making decisions without human involvement, and (ii) if it is automated, ensure it can be classified as exception under Article 22(2) like the automated decision is necessary for a contract, or explicitly consented to, or authorized by law with safeguards. While SMEs as entities are not directly protected under GDPR like individuals, SME loans may involve personal guarantees or data of owners which fall under GDPR – and ethically, similar principles are being applied to SME lending to ensure fairness and transparency.

This case demonstrates that AI compliance is only under development. In fact, the EU AI Act is the first comprehensive regulation that specifically oriented on artificial intelligence. Originally formulated in 2024, it is expected to fully take effect in 2025-2026 (EU, 2024). The idea of regulation is to have risk-based approach to formulate requirements for each type of AI risk: unacceptable, high, limited, or minimal. AI systems for credit scoring are explicitly classified as “high-risk” meaning it is permitted but subject to strict obligations and oversight.

In the proposed Annex III, creditworthiness assessment is listed alongside things like employment and biometric ID as areas where AI has

significant impact on citizens' lives. For banks, the AI Act's most significant aspect is that AI credit scoring systems will be red flagged due to potential for discrimination. High-risk AI systems will face stringent requirements as they must undergo conformity assessments before deployment, ensuring transparency in how they work, sufficient accuracy, and human oversight. The Act will require credit AI models to have technical documentation to demonstrate explainability-by-design so that outputs can be interpreted. Importantly, the AI Act also mandates measures to prevent and test for discriminatory outcomes, meaning if a lender uses an LLM for credit scoring, they will need to show that it doesn't systematically disadvantage applicants without a valid justification. These requirements are imposing restriction which data can be used and how it is treated by credit risk models.

However, an important nuance in case of the Annex III linked to the point that creditworthiness provision refers to natural persons. This means AI models assessing the credit risk of companies or SMEs (legal persons) are not explicitly listed as high-risk under the Act's final text and might be classified as limited one. While this might make SME's loans not formally qualify under "high-risk AI" case, there are two other considerations from the legislation. First, the regulation aims to prevent automated discrimination and controversial outcomes in lending, suggesting that application of Annex is universal according to the spirit of the law when it comes to natural persons. Consequently, if the assessment includes evaluation of management or sole entrepreneurship, this system will fall under Annex III cases. Second, in a previous discussion around SMEs and fraud detection systems, the Council of the EU was proposing exclusion of SME internal tools to reduce regulatory

burden (Council of the EU, 2022) while European Parliament were disagreeing with the exemption as regulation focus is oriented towards safety measures while agreed with excluding fraud detection systems from high-risk ultimately in the financial sector (European Parliament, 2023). Based on this proposition, which was published in the final version (EU, 2024), the coverage of the regulation is expected to be universal. This case supports the vision where stricter legalization methods are applied, even if they put additional scrutiny.

The complexity and disparity of both regulations suggest that compliance procedures will involve new governance setups. Under the Act, financial-sector AI applications may be supervised by either national AI authorities or existing financial regulators. It raises the possibility of dual oversight when bank's AI models might require approval from both an AI regulator and an existing financial supervisor. Such complex oversight structure will require coordination to avoid conflicting guidance. Additionally, AI functionalities are constantly evolving and requires continuous incremental changes to the regulation, which can be efficient only if actors and stakeholders which are affected by the regulation are actively participating (Justo-Hanani, 2022). European bankers have voiced concern that compliance burdens might slow AI innovation or limit model performance due to very conservative constraints (Zamfir, & Pototschnig, 2023). Nonetheless, the AI Act aims are targeted to balance innovation with fundamental rights – expecting lenders to build accountability and fairness into their AI credit systems.

Both GDPR and the AI Act effectively force a high degree of transparency in AI decisions. For GenAI in credit risk, this acts as a double-edged sword – LLMs are often considered to be untraceable "black boxes",

albeit they can also be used to generate additional context and justification to support the applied logic. Indeed, practitioners are exploring using GenAI to draft more pronounced explanation notes to fulfill adverse action notice requirements, while still grounded in the model's actual factors. The AI Act will likely require that such explanations be provided in an understandable form to the user. The challenge is ensuring the explanation is faithful to the model and not just a plausible explanation.

The ethical use of AI in credit mandates avoiding discrimination. In Europe, this is not only a legal risk but also a reputational one. GenAI systems are expected to be trained and tested to ensure they do not redline or inadvertently use proxies for protected characteristics. To address this, robust model validation is needed with relevant evaluation of the applicability of GenAI and checks of the outputs for bias and relevance. Techniques like Shapley values or LIME (local interpretable model-agnostic explanations) have been used in traditional ML for credit to check bias; similar approaches can be applied to GenAI, albeit with adaptation (Melsom et al., 2022; Chen et al., 2024). Financial regulators like the European Banking Authority have also highlighted that model risk management principles (governance, documentation, testing) apply equally to AI/ML models – meaning banks should extend their validation frameworks to generative models (EBA, 2023).

In summary, Europe's regulatory space is oriented towards innovation but remains firm on guardrails. GDPR enforcement (as seen in the Schufa case) and the new AI Act framework together ensure that GenAI in credit risk is and will be subject to rigorous scrutiny. Lenders implementing these technologies for SME credit scoring or underwriting will need to invest in

compliance including model validation, bias testing, and building explainability and human oversight tools for their AI systems. On the beneficial side, those efforts have potential to expand credit access in a fair and efficient manner. The next section builds on how exactly financial institutions can integrate Generative AI into SME credit risk assessment in quantitative analysis.

4.5.2 Model governance and validation

While regression methods for events prediction have been continuously researched from perspective of validation (for example, extensive study by Tantithamthavorn et al., 2016), the GenAI component brings extra complexity to the model governance process. Indeed, even we defined empirical evaluation strategy for various metrics in the data section, this would only be relevant for model selection and robustness but not its validation over time.

The observed effects and their ability to persist are sensitive not just to the derived data but also to process of its acquiring. Thus, the validation process should account for (i) validating the language model, and (ii) change in underlying unstructured data. Both spaces are quite extensive and go far beyond the credit risk modelling context, thus formulating validation principles around Z_i features and language function $G(.)$ remain a challenge for the future research.

Choice of LLM model is another crucial consideration as choice of a particular option might be associated with favorable outcomes. Testing and justifying various model to establish a benchmark is another possible extension of the incorporating GenAI data into modelling assessment. Specifically, comparing the potential advantage of domain-focused models (like FinGPT) to

more generalized language models might of great contribution. Human oversight and override should remain the final layer of credit risk assessment, especially for high-risk decisions. Moreover, if the AI's confidence is low or it triggers a compliance rule (e.g. possible bias detected), it should flag for human review.

The compliance team should be involved in designing the system on top of existing governance procedures, ensuring that regulatory requirements (like GDPR's rights) are met. This would cover the cases when an SME requests an explanation or human intervention and support processes of providing documented details. The AI Act's documentation requirements can be satisfied by creating a "technical file" for the AI system – describing its design, training data, risk assessment, and periodic audits, possibly by external experts.

The successful deployment of GenAI in SME credit assessment could have positive policy implications – it aligns with regulators' objectives of broader SME finance and inclusive lending. By using more data and sophisticated analysis, lenders may approve more SMEs that were previously declined due to lack of info, thus improving access to credit for underserved segments. Moreover, if the AI can identify risks earlier, it can help SMEs avoid default, contributing to financial stability. Policymakers will, however, keep a close eye on outcome whether these AI-driven decisions fair and SMEs receive understanding of the underlying reasons while demanding raising industry standards for AI explainability, especially when it drives decision making (Seeber et al., 2020).

In Europe specifically, coordination between data protection regulators, AI regulators, and financial supervisors will be key. The framework proposed ensures that any GenAI system is responsible by design, which will ease supervisory acceptance. Banks that pioneer these techniques could share best practices through industry bodies or pilot programs (the European Banking Federation or EBA might initiate something like a regulatory sandbox for AI in credit, to gather learnings in a controlled environment).

4.6 Conclusion

Generative AI holds a large promise to revolutionize SME credit risk assessment in Europe by addressing longstanding information gaps and efficiency bottlenecks. The proposed GenAI-augmented framework demonstrate how newly derived features can enhance credit decision accuracy, speed, and fairness – while remaining compliant with existing and upcoming regulations.

Crucially, successful adoption will depend on trust – by lending institutions in the AI's enhancements, by regulators that the AI is controlled and compliant, and by SMEs that they are treated fairly. This trust can be earned through the diligent application of the safeguards: transparency and explainability, human oversight, continuous validation, and alignment with regulatory principles of fairness and accountability.

The assessment of existing regulations demonstrated that current approach is balancing between adoption of new innovations and maintaining high standards of data manipulation and compliance. The most recent AI Act and GDPR put additional burden on lenders in terms of information treatment

(which was demonstrated during Schufa case) but act only in interest of borrowers and their fair treatment by lending organization. The adoption of the GenAI tools in respect to SMEs might face certain grey zones, which will be only clarified once regulation is fully operational.

In conclusion, the fusion of Generative AI with SME credit risk assessment is a paradigm shift in the making. The current state of research and practice shows that we have the building blocks to make credit evaluation more inclusive, data-driven, and intelligent than ever before. By innovating within a sound governance framework, lenders can unlock new opportunities: not only achieving better risk management but also granting better access to finance for SMEs growth across Europe. The proposed modelling framework demonstrates that using a combination of traditional and GenAI-derived data provides a better understanding of the SME credit risk phenomena, especially when for the information that might be changing rapidly. Testing the proposed modelling framework will bring a large number of opportunities for impact testing, SME sector stress-testing, and early distress recognition.

5 Thesis Conclusion

This thesis has aim to address one of the central problems in SME credit risk assessment: lenders must make decisions about opaque firms using incomplete, uneven, and often low-frequency information, while the event to be predicted is not exactly neutral. The described research challenges and the analytical framework together with the answers to proposed questions M1-M5 and E1-E5 converge on one general conclusion that better SME credit risk modelling depends on the joint design of outcome definition, explanatory information, and estimation strategy, rather than on any single, isolated methodological innovation. For practical implications, the thesis shows that accuracy is not sufficient as the dominant benchmark. In the context, information asymmetry and high economic costs of model instability, a useful model is one that remains interpretable from economic and business points, while empirically dependable when the target event changes, when the sample moves out of time, and when the borrower population becomes heterogeneous.

The first thesis-level contribution is methodological. The comparative evidence developed earlier in the thesis indicates that innovation in credit modelling should not be read as a simple progression from simpler econometric models to more complex machine-learning tools. The finding that XGBoost can generate stronger in-sample fit, while logistic regression combined with specialized preprocessing and binning demonstrates greater robustness and interpretability, is important as it shifts the evaluative criterion from raw discrimination to stability under realistic credit-risk conditions. This reframes robustness from a diagnostic purpose into a design principle. Within the logic of agency theory, that matters because unstable models increase uncertainty

for lenders and may add precautionary credit rationing under lack of information. The broader credit-scoring literature supports this emphasis: recent evidence shows that when datasets become more imbalanced, the stability of interpretability tools (like SHAP and LIME) deteriorates, implying that robustness must apply not only to predictions but also to their meaning (Chen et al., 2024).

The second thesis-level contribution is empirical and provides extra evidence for the definition of credit risk itself. The thesis demonstrates that the dependent variable is not a simple technical choice rather a structural determinant of model behaviour. When the target is expanded from standard default to more granular versatile indicators of deterioration, the resulting models can capture more stable relationships across specifications. This addresses the research challenge embedded in M3 and E1-E4: whether robustness can be improved by redefining credit events instead of changing algorithms. Delinquency definitions reveal that credit deterioration is a process rather than a single endpoint, and that different explanatory variables might be more or less important at different stages of that process. In particular, non-financial and structural variables gain relevance when risk is defined more broadly than default alone. This is consistent with the thesis's stakeholder perspective, because distress affects not only the lender-borrower contract but also the wider network of suppliers, employees, owners, and regions before default formally materialises.

A third contribution follows from the treatment of heterogeneity. The Spain-Italy evidence shows that pooling data across countries can improve inference when low-frequency events make single-country estimation

unbalanced. However, it is essential to repeat that such gain in statistical power is not equivalent to assuming SME homogeneity. At the same time, size effects are present: micro, small, and medium enterprises exhibit distinct patterns of credit behaviour, implying that segmentation within the SME class is necessary from analytical perspective. This finding responds directly to E3 and E5 and speaks back to the original research challenge of generalization of findings. The contribution here is not that pooled models can work, but that they work best when pooling is enhanced by economic comparability and combined with meaningful intra-SME segmentation. In other words, the thesis shows the importance of balance between generalisation and granularity properties when applied to SMEs.

The data challenge which is highlighted through the whole thesis can be interpreted as a bridge to its most forward-looking implication. Once the empirical results show that risk measurement improves when event definitions are more granular and when models incorporate more than standard balance-sheet variables, the next question is no longer only which technique performs the best. It becomes how the information set itself can be expanded in a traceable way. Recent SME research supports this idea of moving towards a data-centric perspective. For instance, Jiang et al. (2023) show that information extracted from firms' official websites improves SME credit-risk prediction and helps reduce information asymmetry between SMEs and financial institutions, especially through content-based and dynamic features. That evidence matters for this thesis because it validates the underlying logic of Chapter 4: richer risk measurements depend not just on new algorithms, but on new forms of feature construction from unstructured information. Consequently, the transition to

GenAI is not a compensation or another technical improvement; it is a logical extension of empirical results which are established prior in the thesis, including the fact that information depth and event granularity improve modelling robustness.

The methodological significance of the GenAI discussion is defined by how carefully it should be interpreted. The thesis does not argue that LLMs should replace conventional credit modelling, nor that unstructured data can substitute conventional financial analysis. Instead, it claims that GenAI may become justified strategy where it performs a very specific function by transforming textual, qualitative, and weakly observable SME information into quantitative features which can be monitored and governed within existing credit-risk frameworks. That position is strengthened by the emerging literature, which recognises that generative AI can enhance predictive capability and decision support, but simultaneously identifies bias, transparency, and regulatory accountability as first-order concerns (Ali et al., 2025).

Taken together, the overall contributions of the thesis can be restated clearly. Conceptually, it positions SME credit risk as a problem of representation under asymmetry of information, where robustness is a theoretically grounded requirement. Empirically, it shows that model conclusions depend on how deterioration is defined, that non-financial information matters more when risk is observed more granularly, that cross-country pooling can improve models when credit risk events are sparse, and that size heterogeneity within SMEs is meaningful. Methodologically, it offers an extension path from interpretable classical models and robustness-oriented validation towards a broader data architecture in which richer features,

including those derived from GenAI-enabled processing, can be incorporated without reducing transparency, interpretability and accountability. Nonetheless, the GenAI framework remains conceptual and is pending future empirical testing. The next stage of research should therefore extend the thesis's core argument: wider-country validation, multi-state or survival-based event modelling, and formal evaluation of GenAI-derived variables under confidence, stability, and governance criteria would deepen the contribution while maintain the thesis's central theme that reliability in SME credit assessment starts with robust problem formulation.

References

Abdou, H. A., Tsafack, M. D. D., Ntim, C. G., & Baker, R. D. (2016). Predicting creditworthiness in retail banking with limited scoring data. Knowledge-Based Systems, 103, 89-103.

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

Abraham, F., & Schmukler, S. L. (2017). Addressing the SME finance problem. World Bank Research and Policy Briefs(120333).

Akerlof, G. A. (1970). The market for “lemons”: Quality uncertainty and the market mechanism. Quarterly Journal of Economics, 84(3), 488-500.

Ala'raj, M., & Abbod, M. F. (2016). A new hybrid ensemble credit scoring model based on classifiers consensus system approach. Expert Systems with Applications, 64, 36-55.

Ali, H., Zafar, M. B., & Aysan, A. F. (2025). Generative AI in finance: Replicability, methodological contingencies, and future research directions. Finance Research Letters, 108797.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589-609.

Altman, E. I., & Sabato, G. (2007). Modelling credit risk for SMEs: Evidence from the US market. Abacus, 43 (3), 332-357.

Altman, E. I., Balzano, M., Giannozzi, A., & Srhoj, S. (2023). The Omega Score: An improved tool for SME default predictions. Journal of the International Council for Small Business, 4(4), 362-373.

Altman, E. I., Esentato, M., & Sabato, G. (2020). Assessing the credit worthiness of Italian SMEs and mini-bond issuers. Global Finance Journal, 43, 100450.

Altman, E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589-609.

Altman, E.I., Sabato, G., Wilson, N. (2010). The value of non-financial information in small and medium-sized enterprise risk management. The Journal of Credit Risk, 6(2), 1-33.

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

Angilella, S., & Mazzù, S. (2015). The financing of innovative SMEs: A multicriteria credit rating model. European Journal of Operational Research, 244(2), 540-554.

Arslan, Ö., & Karan, M. B. (2009). Credit risks and internationalization of SMEs. Journal of Business Economics and Management, 10(4), 361-368.

Balcaen, S., & Ooghe, H. (2006). 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. The British Accounting Review, 38(1), 63-93.

Baltodano López, S., Götz, T., & Scholz, J. (2024). Modeling corporate CDS spreads using Markov switching regressions. Studies in Nonlinear Dynamics & Econometrics, 28(2), 271-292. https://doi.org/10.1515/snde-2022-0084

BCBS (2015) Guidance on credit risk and accounting for expected credit losses. Web page, accessed 20-February-2023.

Beck, T., Demirgüç-Kunt, A., & Maksimovic, V. (2008). Financing patterns around the world: Are small firms different?, Journal of financial economics, 89(3), 467-487.

Bekhet, H. A., & Eletter, S. F. K. (2014). Credit risk assessment model for Jordanian commercial banks: Neural scoring approach. Review of Development Finance, 4(1), 20-28.

Berentsen, G. D., Meyer, B. D., & Scholtes, J. (2022). Modelling clusters of corporate defaults: Regime-switching models significantly reduce the contagion source. Journal of the Royal Statistical Society: Series C (Applied Statistics), 71(1), 1-24.

Berge, L. (2018). Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm. CREA Discussion Papers.

Berger, A. N., & Frame, W. S. (2007). Small business credit scoring and credit availability. Journal of Small Business Management, 45(1), 5-22.

Berger, A. N., & Udell, G. F. (1998). The economics of small business finance: The roles of private equity and debt markets. *Journal of Banking & Finance*, 22(6-8), 613-673.

Berloco, A., Caldarola, E. G., & Petrella, A. (2023). Forecasting short-term defaults of firms in a commercial network via Bayesian spatial and spatio-temporal methods. International Journal of Forecasting, 39(3), 1065-1077.

Bernanke, B. S., Gertler, M., & Gilchrist, S. (1999). The financial accelerator in a quantitative business cycle framework. Handbook of macroeconomics, 1, 1341-1393.

Bertoni, F., Colombo, M. G., & Quas, A. (2023). The long-term effects of loan guarantees on SME performance. Journal of Corporate Finance, 80, 102408.

Black, F., & Cox, J. C. (1976). Valuing corporate securities: Some effects of bond indenture provisions. The Journal of Finance, 31(2), 351-367.

Bloomberg. (2023). Introducing BloombergGPT, Bloomberg's 50-billion parameter large language model, purpose-built from scratch for finance. URL: (Accessed 22-05-2025)

Boubaker, S., Cellier, A., Manita, R., & Saeed, A. (2020). Does corporate social responsibility reduce financial distress risk?. Economic Modelling, 91, 835-851.

Calabrese, R., Andreeva, G., & Ansell, J. (2019). "Birds of a Feather" fail together: Exploring the nature of dependency in SME defaults. Risk Analysis, 39(1), 71-84.

Calabrese, R., Girardone, C. & Sclip, A. (2021). Financial fragmentation and SMEs' access to finance. Small Bus Econ, 57, 2041-2065 .https://doi.org/10.1007/s11187-020-00393-1

Calabrese, R., Girardone, C., & Sclip, A. (2021). Financial fragmentation and SMEs' access to finance. Small Business Economics, 57(4), 2041-2065.

Calabrese, R., Marra, G., & Angela Osmetti, S. (2016). Bankruptcy prediction of small and medium enterprises using a flexible binary generalized extreme value model. Journal of the Operational Research Society, 67, 604-615.

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. The Journal of finance, 63(6), 2899-2939.

Casanova, C., Hardy, B., & Onen, M. (2021). Covid-19 policy measures to support bank lending. BIS Quarterly Review, 20. Accessed online 14-Apr-24.

Castagnolo, F., & Ferro, G. (2014). Models for predicting default: towards efficient forecasts. The Journal of Risk Finance, 15(1), 52-70.

Chan-Lau, J. A. (2006). Fundamentals-based estimation of default probabilities: a survey. International Monetary Fund Working Paper

Chen, H., Lu, W., Song, R., & Ghosh, P. (2024). On learning and testing of counterfactual fairness through data preprocessing. Journal of the American Statistical Association, 119(546), 1286-1296.

Chen, Y., Calabrese, R., & Martin-Barragan, B. (2024). Interpretable machine learning for imbalanced credit scoring datasets. European Journal of Operational Research, 312(1), 357-372.

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

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Zhou, T. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1-4.

Chen, Y., Calabrese, R., & Martin-Barragan, B. (2024). Interpretable machine learning for imbalanced credit scoring datasets. European Journal of Operational Research, 312(1), 357-372.

Cheraghali, H., & Molnár, P. (2024). SME default prediction: A systematic methodology-focused review. Journal of Small Business Management, 62(6), 2847-2905.

Ciampi, F., Giannozzi, A., Marzi, G., & Altman, E. I. (2021). Rethinking SME default prediction: a systematic literature review and future perspectives. Scientometrics, 126, 2141-2188.

Corredera-Catalán, F., di Pietro, F. & Trujillo-Ponce, A. Post-COVID-19 SME financing constraints and the credit guarantee scheme solution in Spain. Journal of Banking Regulation, 22, 250-260 (2021).

Council of the European Union. (2022, December 6). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) – General Approach (Document ST 14954/22). https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf

Court of Justice of the European Union. (2023, December 7). OQ v. Schufa Holding AG and Land Hessen (Case C-634/21). URL: (Accessed 17-07-2025)

Danielsson, J., Valenzuela, M., & Zer, I. (2018). Learning from history: Volatility and financial crises. The Review of Financial Studies, 31(7), 2774-2805.

Decree-Law No. 7/2022 of January 26, 2022, implementing Directive (EU) 2019/1023 concerning the Corporate Crisis and Insolvency Code.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the

association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

Dietsch, M., & Petey, J. (2002). The credit risk in SME loans portfolios: Modeling issues, pricing, and capital requirements. Journal of Banking & Finance, 26(2-3), 303-322.

Dietsch, M., & Petey, J. (2004). Should SME exposures be treated as retail or corporate exposures? A comparative analysis of default probabilities and asset correlations in French and German SMEs. Journal of Banking & Finance, 28(4), 773-788.

Dowle, M., & Srinivasan, A. (2021). data.table: Extension of `data.frame`. R package version 1.14.2. URL https://CRAN.R-project.org/package=data.table

Duan, J. C., Kim, B., Kim, W., & Shin, D. (2018). Default probabilities of privately held firms. Journal of Banking & Finance, 94, 235-250.

European Banking Authority. (2016, September 28). Guidelines on the application of the definition of default. EBA/GL/2016/07.

European Banking Authority. (2023, August 4). Followup report on machine learning from the consultation on the discussion paper on machine learning for IRB models. EBA/REP/2023/28.

European Central Bank. (2018). Guide to internal models: Risk-type-specific chapters (credit risk). Web page, Accessed on 08 April 2026.

European Central Bank (2020) What is AnaCredit?. Web page, Accessed on 20 February 2023.

European Parliament. (2023, June 14). Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)). URL: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html (Accessed 24-07-2025)

European Union. (2024, July 12). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 21 May 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2019/2144 and (EU) 2019/881 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (OJ L, 2024/1689). URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj (Accessed 24-07-2025)

Eurostat. (2025, May 16). Quarterly registrations of new businesses and declarations of bankruptcies – statistics. European Commission. URL: (Accessed 27-05-2025)

Falagiarda, M., & Köhler-Ulbrich, P. (2021). Bank Lending to Euro Area Firms What Have Been the Main Drivers During the COVID-19 Pandemic? 49. European Economy, (1), 119-143.

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

Fearnley, S., & Hines, T. (2007). How IFRS has destabilised financial reporting for UK non-listed entities. Journal of Financial Regulation and Compliance, 15(4), 394-408.

Fedorova, E., Gilenko, E., & Dovzhenko, S. (2013). Bankruptcy prediction for Russian companies: Application of combined classifiers. Expert systems with applications, 40(18), 7285-7293.

Ferrari, S., Van Roy, P., & Vespro, C. (2011). Stress testing credit risk: Modelling issues. Financial Stability Review, 9(1), 105-120.

Fidrmuc, J., & Hainz, C. (2010). Default rates in the loan market for SMEs: Evidence from Slovakia. Economic Systems, 34(2), 133-147.

Filipe, S. F., Grammatikos, T., & Michala, D. (2016). Forecasting distress in European SME portfolios. Journal of Banking & Finance, 64, 112-135.

Finch, W. H., Bolin, J. E., & Kelley, K. (2019). Multilevel modelling using R. CRC Press.

Frey, R., & McNeil, A. J. (2003). Dependent defaults in models of portfolio credit risk. Journal of Risk, 6, 59-92.

Fulmer, J., Moon, J., Gavin, T., & Erwin, M. (1984). A bankruptcy classification model for small firms. Journal of Commercial Bank Lending(7)., pp. 25-37

Gautam, S. (2023, October). Bridging multimedia modalities: enhanced multimodal AI understanding and intelligent agents. In Proceedings of the 25th International Conference on Multimodal Interaction (pp. 695-699).

Ghulam, Y., Hakro, A. N., & Naumani, O. (2025). SMEs' Access to Bank Financing During the Financial Crises in Europe. Journal of Small Business Strategy, 35(1), 74-96

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. Behavior research methods, instruments, & computers, 36(2), 193-202.

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. Journal of Statistical Software, 40, 1-25.

Gupta, J., Wilson, N., Gregoriou, A., & Healy, J. (2014). The effect of internationalisation on modelling credit risk for SMEs: Evidence from UK market. Journal of International Financial Markets, Institutions and Money, 31, 397-413.

Hallucination-free? Assessing the reliability of leading AI legal research tools. Journal of Empirical Legal Studies, 22(2), 216-242.

Hayden, E. (2003). Are credit scoring models sensitive with respect to default definitions? Evidence from the Austrian market. Evidence from the Austrian Market (April 2003).

Hillegeist, S. A., Keating, E. K., Cram, D. P., & Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. Review of Accounting Studies, 9, 5-34.

Hlavac M (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables. Social Policy Institute, Bratislava, Slovakia. R package version 5.2.3, https://CRAN.R-project.org/package=stargazer.

Hörisch, J., Freeman, R. E., & Schaltegger, S. (2014). Applying stakeholder theory in sustainability management: Links, similarities, dissimilarities, and a conceptual framework. Organization & Environment, 27(4), 328-346.

Huang, J. Z., & Huang, M. (2012). How much of the corporate-treasury yield spread is due to credit risk?. The Review of Asset Pricing Studies, 2(2), 153-202.

IASB (2014) IFRS 9 Financial Instruments, accessed 20-Feb-2023 .

Insolvency Service. (2024, August 20). Commentary – Company Insolvency Statistics July 2024. GOV.UK. URL: (Accessed 17-05-2025)

Jacobson, T., Lindé, J., & Roszbach, K. (2005). Credit risk versus capital requirements under Basel II: are SME loans and retail credit really different?. Journal of Financial Services Research, 28(1-3), 43-75.

Jiang, C., Yin, C., Tang, Q., & Wang, Z. (2023). The value of official website information in the credit risk evaluation of SMEs. Journal of Business Research, 169, 114290

Jiménez, G., & Saurina, J. (2006). Credit cycles, credit risk, and prudential regulation. *International Journal of Central Banking*, 2(2), 65-98.

Justo-Hanani, R. (2022). The politics of Artificial Intelligence regulation and governance reform in the European Union. Policy Sciences, 55(1), 137-159.

Kalemli-Ozcan, S., Sorensen, B., Villegas-Sanchez, C., Volosovych, V., & Yesiltas, S. (2015). How to construct nationally representative firm level data from the Orbis global database: New facts and aggregate implications (No. w21558). National Bureau of Economic Research.

Katsinis, A., Lagüera-González, J., Di Bella, L., Odenthal, L., Hell, M., & Lozar, B. (2024). Annual Report on European SMEs 2023/2024. Publications Office of the European Union, Luxemburg. doi:10.2826/355464.

KPMG LLP. (2024, August). GenAI survey: Risk mitigation priorities in enterprise adoption. KPMG LLP. Retrieved from . Accessed 17-06-2025

Kuhn, M. (2022). caret: Classification and Regression Training. R package version 6.0-91. URL

Law 16/2022, amending the Insolvency Law for the transposition of Directive (EU) 2019/1023 of the European Parliament and Council of June 20, 2019, on restructuring and insolvency.

Lehmann, B. (2003). Is it worth the while? The relevance of qualitative information in credit rating. Working Paper presented at the EFMA 2003 Meetings, Helsinki, p. 1-25.

Leland, H. E., & Toft, K. B. (1996). Optimal capital structure, endogenous bankruptcy, and the term structure of credit spreads. The Journal of Finance, 51(3), 987-1019.

Lin, S. M., Ansell, J., & Andreeva, G. (2012). Predicting default of a small business using different definitions of financial distress. Journal of the Operational Research Society, 63(4), 539-548.

Liu, X. Y., Wang, G., Yang, H., & Zha, D. (2023, July). Data-centric fingpt: Democratizing internet-scale data for financial large language models. In NeurIPS Workshop on Instruction Tuning and Instruction Following.

Longstaff, F. A., & Schwartz, E. S. (1995). A simple approach to valuing risky fixed and floating rate debt. The Journal of Finance, 50(3), 789-819.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025).

Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G., & Ziegler, A. (2012). Probability machines: consistent probability estimation using nonparametric learning machines. Methods of Information in Medicine, 51(1), 74.

Marques, A. I., García, V., & Sánchez, J. S. (2013). A literature review on the application of evolutionary computing to credit scoring. Journal of the Operational Research Society, 64, 1384-1399.

Mc Namara, A., O'Donohoe, S., & Murro, P. (2020). Lending infrastructure and credit rationing of European SMEs. European Journal of Finance, 26(7-8), 728-745

McCann, F., & McIndoe-Calder, T. (2012). Determinants of SME loan default: the importance of borrower-level heterogeneity. Research Technical Paper (No. 06/RT/12). Central Bank of Ireland.

Melsom, B., Vennerød, C. B., de Lange, P. E., Hjelkrem, L. O., & Westgaard, S. (2022). Explainable artificial intelligence for credit scoring in banking. Journal of Risk, 25(2), 1-25.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. The Journal of Finance, 29(2), 449-470.

Modina, M., Pietrovito, F., Gallucci, C., & Formisano, V. (2023). Predicting SMEs' default risk: Evidence from bank-firm relationship data. The Quarterly Review of Economics and Finance, 89, 254-268.

Moscalu, M., Girardone, C., & Calabrese, R. (2020). SMEs' growth under financing constraints and banking markets integration in the euro area. Journal of Small Business Management, 58(5), 974-1005

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 109-131.

Perera, D., & Chand, P. (2015). Issues in the adoption of international financial reporting standards (IFRS) for small and medium-sized enterprises (SMES). Advances in Accounting, 31(1), 165-178.

Ptak-Chmielewska, E., & Kopciuszewski, T. (2022). New definition of default—Recalibration of credit risk models using Bayesian approach. Risks, 10(1), 16.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL .

Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. SIAM Review, 26(2), 195-239.

Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3982-3992).

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics, 12(1), 1-8.

Roelofs, R., Shankar, V., Recht, B., Fridovich-Keil, S., Hardt, M., Miller, J., & Schmidt, L. (2019). A meta-analysis of overfitting in machine learning. Advances in Neural Information Processing Systems, 32.

Roy, P.K. and Shaw, K. (2023). A credit scoring model for SMEs using AHP and TOPSIS. International Journal of Finance & Economics, 28(1), pp.372-391.

Schulze Brock, P., Katsinis, A., Lagüera González, J., Di Bella, L., Odenthal, L., Hell, M., Lozar, B., & Secades Casino, B. (2025). *Annual report on European SMEs 2024/2025: SME performance review* (JRC142263). Publications Office of the European Union.

Seeber, I., Waizenegger, L., Seidel, S., Morana, S., Benbasat, I., & Lowry, P. B. (2020). Collaborating with technology-based autonomous agents: Issues and research opportunities. Internet Research, 30(1), 1-18.

Shapiro, J., & Zeng, J. (2024). Stress testing and bank lending. The Review of Financial Studies, 37(4), 1265-1314.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. The Journal of Business, 74(1), 101-124.

Siddiqi, N. (2006): Credit Risk Scorecards: developing and implementing intelligent credit scoring. Hoboken, NJ: John Wiley & Sons, Inc.

Siddiqi, N. (2016) Intelligent credit scoring: Building and implementing better credit risk scorecards (2nd ed., pp. 119-130). Hoboken, NJ: John Wiley & Sons.

Song, H., Yu, K., Ganguly, A., & Turson, R. (2016). Supply chain network, information sharing and SME credit quality. Industrial Management & Data Systems, 116(4), 740-758.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2020). Score-based generative modelling through stochastic differential equations. Proceedings of the 37th International Conference on Machine Learning (ICML).

Stevenson, M., Mues, C., & Bravo, C. (2021). The value of text for small business default prediction: A deep learning approach. European Journal of Operational Research, 295(2), 758-771.

Stiglitz, J. E., & Weiss, A. (1981). Credit rationing in markets with imperfect information. *American Economic Review*, 71(3), 393-410.

Tantithamthavorn, C., McIntosh, S., Hassan, A. E., & Matsumoto, K. (2016). An empirical comparison of model validation techniques for defect prediction models. IEEE Transactions on Software Engineering, 43(1), 1-18.

Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). Credit scoring and its applications. Philadelphia: Society for Industrial and Applied Mathematics.

Tucker, J., & Lean, J. (2003). Small firm finance and public policy. Journal of Small Business and Enterprise Development, 10(1), 50-61.

Umavezi, A. T. (2025). Bayesian deep learning for uncertainty quantification in financial stress testing and risk forecasting. International Journal of Research Publications and Reviews, 6(5), 6540-6555.

URL

Wagner, H. (2016). Default definition under Basel. In N. Siddiqi (Ed.), Intelligent credit scoring: Building and implementing better credit risk scorecards (2nd ed., pp. 119-130). Hoboken, NJ: John Wiley & Sons.

Wang, X., Han, L., & Huang, X. (2020). Bank competition, concentration and EU SME cost of debt. International Review of Financial Analysis, 71, 101534.

Weng, H., Yan, H., Wang, Y., & Guo, Z. (2025). Class imbalance Bayesian model averaging for consumer loan default prediction: The role of soft credit information. Research in International Business and Finance, 74, 102722.

Wickham, H. (2021). dtplyr: Data Table Back-End for 'dplyr'. R package version 1.1.0.

Wickham, H., & Grolemund, G. (2016). R for data science: import, tidy, transform, visualize, and model data. O'Reilly Media, Inc.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., ... & Yutani, H. (2019). Welcome to the Tidyverse. Journal of Open Source Software, 4(43), 1686.

Xiong, W., Wu, D. D., & Yeung, J. H. (2024). Semiconductor supply chain resilience and disruption: Insights, mitigation, and future directions. International Journal of Production Research, 1-24.

Zamfir, I., & Pototschnig, P. (2023). AI in financial services: Between opportunity and over-regulation (CEPS Policy Brief No. 2023/08). Centre for European Policy Studies.

Zhu, Y., Xie, C., Wang, G. J., & Yan, X. G. (2017). Comparison of individual, ensemble and integrated ensemble machine learning methods to

predict China’s SME credit risk in supply chain finance. Neural Computing and Applications, 28, 41-50.

Zhu, X., Chu, Q., Song, X., Hu, P., & Peng, L. (2023). Explainable prediction of loan default based on machine learning models. Data science and management, 6(3), 123-133.

