UNIVERSITY OF TWENTE.

MSc. Industrial Engineering and Management
Financial Engineering and Management

Estimating Dutch SME EBITDA from public data: A predictive framework for early-stage lead sourcing

S. G. Jeurissen

Master Thesis (MSc.)

June, 2025

Supervisors University:

B. Roorda
H. Kroon

Faculty BMS
University of Twente
The Netherlands

Supervisors Marktlink:

J. Scholtens
J. Blom

Marktlink
Deventer
The Netherlands

Management Summary

Valuing Small and Medium-sized Enterprises (SMEs) when sourcing potential deals remains a major challenge in the Dutch Merger and Acquisition (M&A) market. Most SMEs disclose only limited financial information, typically restricted to balance sheets, leaving crucial income statement data, such as Earnings Before Interest, Taxes, Depreciation, and Amortization (EBITDA), unavailable. This research, conducted in collaboration with Marktlink, addresses this gap by developing a method to accurately estimate SME EBITDA using only publicly available financial statement data.

The research followed a structured process. After reviewing academic literature on SME valuation methods and profitability drivers, a dataset of Dutch and Belgian SMEs was extracted from Bureau van Dijk's Orbis database. The data underwent cleaning, including Winsorization to mitigate the impact of outliers and log transformations to address skewed distributions. Exploratory data analysis, including sector-specific correlation testing and backtesting with 2019 data, revealed and confirmed the stability of relationships underlying features across time and industries.

Regression models were then developed separately for three high-volume sectors, business services, wholesale, and construction, reflecting the sectoral differences in financial structures. Two modeling spaces were considered: normal space for absolute euro-level predictions and log-transformed space for proportional accuracy. Multiple Linear Regression (MLR), Random Forest (RF), and XGBoost were compared across different feature sets. A visual summary of the modeling process is shown in Figure 1, outlining the key steps from data collection to model validation.

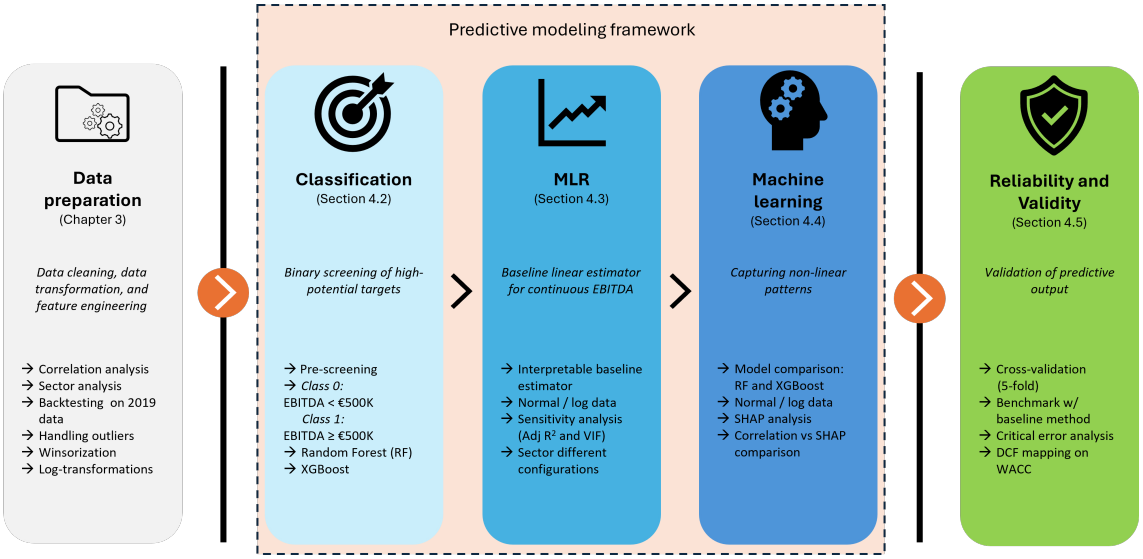


FIGURE 1: Overview of the modeling process

Among the models tested, XGBoost with log-transformed features delivered the strongest results, reaching R^2 values up to 0.71, and SMAPE as low as 4.35%, in the wholesale sector on log-transformed data. The current rule-of-thumb method produced SMAPE values exceeding 9%, in log-space, and 80%, in normal space. Therefore the developed models reduced prediction errors, both in proportional and absolute terms. Notably, MLR also proved highly competitive, achieving R^2 scores up to 0.66, in log space with minimal bias, and standing out in normal space for its ability to deliver well-centered euro-level predictions. Thanks to its transparency and simplicity, MLR remains a valuable option where model explainability is critical.

To interpret the results of XGBoost, SHapley Additive exPlanations (SHAP) analysis was applied. This revealed that liquidity and leverage-related variables, particularly total assets, equity, and current liabilities, were the strongest predictors of EBITDA across sectors, aligning with established financial theory. 5-fold cross-validation further confirmed the models' stability and generalizability.

An error analysis highlighted a minor blind spot around an EBITDA of €730,000, where some firms were underestimated and occasionally misclassified across the €500,000 screening boundary. Although this did not materially affect overall model quality, future research could address it through hybrid approaches or other targeted adjustments. Finally, mapping predicted EBITDA values into a stylized discounted cash flow (DCF) framework yielded enterprise values and implied sector-specific weighted average cost of capital (WACC) figures consistent with market benchmarks, confirming both the statistical and financial validity of the models developed in this research.

This research offers a practical, scalable tool to improve early-stage deal sourcing by enabling more accurate prioritization of high-potential acquisition targets. To the best of the author's knowledge, it is one of the first studies to demonstrate that sector-specific EBITDA estimates can be derived from balance sheet data alone. By bridging the gap between theoretical research and practical application, this study delivers both academic insights and actionable relevance for M&A practitioners like Marktlink.

Preface

This thesis marks the completion of my Master's in Industrial Engineering and Management at the University of Twente, with a specialization in Financial Engineering and Management. It also symbolizes the end of my time as a student, an incredibly formative chapter of my life. Throughout these years, I have been fortunate to embrace a wide range of opportunities, achieve meaningful milestones, and, most importantly, enjoy an unforgettable experience. I now truly understand the feeling that "your student days are the best days of your life." This journey has shaped me both personally and professionally, and I feel well-prepared with both a Bachelor's and Master's degree in Industrial Engineering and Management to take the next step in my career.

I would like to express my sincere gratitude to everyone who contributed to my development during my time at the University of Twente.

First and foremost, I would like to thank my first academic supervisor, Berend Roorda, for his invaluable guidance throughout this research project. Our discussions consistently challenged me to push my thinking further, significantly enhancing the quality of this thesis. It has been a pleasure working with Berend, both during this project and in an earlier role as a teaching assistant, his expertise and mentorship have been incredibly valuable to me. I also extend my thanks to my second supervisor, Henk Kroon, whose in-depth knowledge and thoughtful feedback further enriched this thesis.

Furthermore, I am also grateful to Marktlink for the opportunity to conduct my graduation project within their organization. In particular, I would like to thank my first supervisor at Marktlink, Jeffrey, for warmly welcoming me into the company and generously sharing his experience in the dynamic world of Mergers and Acquisitions. I also want to thank Jaron, who acted as my buddy during this period, his willingness to answer all kinds of questions and involve me in real deal-making activities made my experience at Marktlink both insightful and enjoyable. Finally, I thank all my colleagues at Marktlink for their support and for creating such a positive working environment.

Last but certainly not least, I want to thank my family and friends for their constant support, not just during the final stages of writing this thesis, but throughout my entire journey as a student. Whether in good times or challenging ones, their presence has meant a great deal to me.

With this chapter coming to a close, I look forward with excitement and optimism to what lies ahead in my professional career and I am more than ready to take on the next steps.

I hope this thesis offers valuable insights and is as enjoyable for you to read as it was for me to write.

Sven Jeurissen
Enschede, June 2025

Contents

List of Figures

List of Tables

Glossary

CAPEX Capital Expenditures

CCA Comparable Company Analysis

CTA Comparable Transaction Analysis

DCF Discounted Cash Flow

EBITDA Earnings Before Interest, Taxes, Depreciation, and Amortization

EPS Earnings Per Share

EV Enterprise Value

FCF Free Cash Flows

M&A Mergers and Acquisitions

MLR Multiple Linear Regression

MPSM Managerial Problem-Solving Method

PV Present Value

RF Random Forest

ROA Return on Assets

ROE Return on Equity

SBI Standaard Bedrijfsindeling

SHAP SHapley Additive exPlanations

SMAPE Symmetric Mean Absolute Percentage Error

SME Small and Medium-sized Enterprise

VIF Variance Inflation Factors

WACC Weighted Average Cost of Capital

Chapter 1

Introduction

Contents

This chapter introduces the research by outlining its relevance, context, and objectives. Section 1.1 highlights the significance of the study, emphasizing the need for accurate SME valuations based on public financial data. In Section 1.2, a description of Marktlink is provided to give insight into the company's role in SME Mergers and Acquisitions (M&A) transactions. Section 1.3 elaborates on the problem context, explaining the challenges faced in estimating SME valuations. Section 1.4 identifies and structures the key problems, leading to the definition of a core problem. Based on this, Section 1.5 establishes the research design and corresponding research questions that guide this study. Finally, Section 1.6 defines the scope of the research, outlining its focus areas.

1.1 Significance

Estimating the revenue and Earnings Before Interest, Taxes, Depreciation, and Amortization (EBITDA) of SMEs based on publicly available financial data is becoming increasingly crucial in the Dutch M&A landscape, where financial transparency is often limited. In the Netherlands, financial disclosure requirements for SMEs are generally restricted to the publication of a balance sheet, with no accompanying income or cash flow statements. As a result, direct estimation of revenue and EBITDA is not feasible from these disclosures alone.

Accurate EBITDA estimates are essential for investors, advisors, and business owners to make well-informed decisions regarding valuations, deal structures, and investment strategies. In M&A advisory, estimating a company's financials before contact is highly valuable. It helps firms quickly identify and prioritize high-potential acquisition targets. Recognizing this need, M&A firms are increasingly adopting data-driven methodologies to improve the accuracy of SME valuations based on publicly available financial information.

In response, this research aims to develop an EBITDA estimation model specifically designed to enhance early-stage deal screening and lead generation. By providing more reliable financial insights before client engagement, this study contributes to improving the efficiency and effectiveness of the lead generation process within the Dutch SME M&A market.

Traditional financial analysis relies on the interconnected use of three core financial statements: the balance sheet, the income statement, and the cash flow statement. Together, these offer a comprehensive view of a company's performance, liquidity, and profitability. However, in the Dutch context, only the balance sheet is typically disclosed by SMEs, making it challenging to extract the financial depth needed for early-stage valuations. This challenge is visualized in Figure 1.1, which shows the three core financial statements that together form the basis of company valuations, of which SMEs typically only publish the balance sheet.

FIGURE 1.1: Interdependence of the three financial statements in corporate finance

This research addresses the challenge of estimating EBITDA using only the publicly available balance sheet. By doing so, it aims to close a critical gap in early-stage lead sourcing for SME transactions.

1.2 About Marktlink

Founded in 1996, Marktlink is a leading M&A advisory firm specializing in SME transactions across Europe. Over the years, the firm has facilitated more than 1,200 deals, solidifying its position as a key player in the SME M&A market. Marktlink provides expert advisory services to entrepreneurs, business owners, and investors, offering tailored guidance throughout the acquisition and sales process.

With offices in multiple countries, including the Netherlands, Belgium, Croatia, Ger-

many, Denmark, Poland, Switzerland, and the United Kingdom, Marktlink has established itself as a market leader in SME transactions. Its extensive presence across Europe enables the firm to facilitate cross-border deals and leverage regional expertise to maximize client value. Since 2018, Marktlink has been recognized as the leading firm in terms of SMEs transactions completed across Europe, further reinforcing its industry position (M&A Community, 2024). The firm offers a range of specialized services:

- **Exit ready:** Preparing businesses for a future sale by optimizing their financial, operational, and strategic positioning.
- **M&A advisory:** Managing the entire transaction process, from initial valuation and deal structuring to negotiations and finalization, supported by in-house legal specialists, debt advisory experts, and a dedicated transaction services team.
- **Marktlink capital:** Providing a platform for entrepreneurs to reinvest capital and explore new investment opportunities.

This research is conducted within Marktlink's M&A Advisory division in the Netherlands, specifically in Deventer, which is primarily responsible for facilitating transactions in the SME market.

1.3 Problem context

Accurately valuing SMEs prior to engaging with them is essential for effective lead generation, which serves as the initial stage of the acquisition process for an M&A firm like Marktlink. Providing Marktlink with a tool capable of pre-assessing company value before initiating contact offers a competitive advantage in identifying and evaluating potential acquisition targets. However, company valuation remains challenging when access to detailed financial data is limited, necessitating reliance on publicly available financial disclosures and market benchmarks. This challenge is primarily caused by financial disclosure regulations, which typically mandate the publication of only the balance sheet for most SMEs, restricting access to key financial metrics such as revenue and EBITDA (Kamer van Koophandel (KvK), 2025). Understanding the commonly used valuation methods in the SME M&A sector highlights why public financial data alone is often insufficient for accurately estimating a company's value.

1.3.1 How are Dutch SMEs typically valued in the industry?

Valuing a SME is a fundamental component of M&A transactions, as an accurate valuation enables buyers, sellers, and financial advisors to determine a company's fair market value and engage in well-informed negotiations. In the Dutch SME market, two primary valuation methods are widely applied: the EBITDA-multiple method and the Discounted Cash Flow (DCF) method. While both approaches provide valuable insights into a company's value, the EBITDA multiple method is generally preferred due to its simplicity and practical applicability in transactions (Brookz, 2025; Corporate Finance Institute (CFI), 2025).

EBITDA-multiple method

The EBITDA-multiple method is one of the most widely applied valuation techniques for Dutch SMEs (Blom, 2024; Stokkers, 2024). This approach determines a company's

Enterprise Value (EV) by multiplying its EBITDA by an industry-specific multiple, derived from comparable transactions or sectoral benchmarks (Blom, 2024).

This method is particularly popular due to its ability to provide a rapid, market-driven estimate of EV based on industry norms. Since EBITDA reflects a company's operational profitability, it serves as an effective representative for cash flow before financing and tax considerations. The EBITDA-multiple itself is influenced by various factors, including industry growth rates, market conditions, company size, and risk profile (Bagna & Ramusino, 2017; Mauboussin, 2018). One of the key reasons for the dominance of EBITDA-multiples in SME valuation is the availability of comparable transaction data, referring to historical M&A deals where EBITDA-multiples were used to determine valuations (Brookz, 2025). This data allows for benchmarking SMEs against sector-specific multiples derived from past acquisitions in the industry.

By combining EBITDA with an industry multiple that reflects sectoral valuation trends, this method provides a widely accepted and practical approach to estimating a company's EV (Nissim, 2024). However, while the EBITDA-multiple method provides a standardized and market-driven approach to valuation, its reliability depends on a clear understanding of EBITDA itself. Since EBITDA is subject to various adjustments and does not fully represent actual cash flows, it is essential to examine its role as a core financial metric in SME valuation (Damodaran, 2006; Palepu & Healy, 2012).

Discounted Cash Flow method

The DCF method is another well-known valuation approach, which involves forecasting a company's future free cash flows and discounting them to their present value using an appropriate discount rate, typically the Weighted Average Cost of Capital (WACC) (Damodaran, 2006).

While the DCF method offers a theoretically sound valuation framework, it is less frequently applied to Dutch SMEs due to its complexity and reliance on uncertain future cash flow projections (Corporate Finance Institute (CFI), 2025). Small businesses often lack the financial predictability necessary for accurate DCF modeling, and assumptions regarding growth rates, discount rates, and long-term sustainability can cause significant valuation uncertainty. Although the DCF method is a well-established valuation technique, its practical applicability in the SME sector remains limited due to these challenges (Corporate Finance Institute (CFI), 2025).

Why the EBITDA-multiple is the preferred method for Dutch SME valuation

In practice, the EBITDA-multiple method is more commonly used than the DCF model for valuing Dutch SMEs due to several key advantages:

- **Simplicity and efficiency:** The EBITDA-multiple approach requires fewer assumptions and can be applied more quickly than the data-intensive DCF method.
- **Market-driven valuation:** Multiples are derived from actual transaction values, ensuring that valuations align closely with current market conditions.
- **Lower forecasting uncertainty:** Unlike the DCF model, which relies on long-term financial projections, the EBITDA-multiple method is based on current operating performance, thereby reducing valuation uncertainty.

As a result, the majority of Dutch M&A transactions involving SMEs rely on the EBITDA-multiple method as the primary valuation approach, with DCF analysis often

serving as a supplementary tool rather than the main valuation technique (Magnimetrics, 2025).

1.3.2 EBITDA as core concept

Since the EBITDA-multiple method is widely used in SME valuations, it is essential to define EBITDA itself as a concept. EBITDA is a financial metric that reflects a company's core operating profitability by excluding financing costs, taxation, and non-cash accounting expenses such as depreciation and amortization. Despite not being a standardized reporting metric under either GAAP or IFRS, EBITDA is widely adopted in corporate finance and valuation practice due to its ability to isolate operational earnings from capital structure, tax environments, and accounting policies (Nissim, 2024).

In M&A practice, EBITDA is often adjusted over multiple periods rather than taken at face value for a single year. This type of EBITDA is referred to as "normalized EBITDA". Financial analysts and advisors normalize earnings to account for anomalies, extraordinary expenses, and expected future performance, making EBITDA a somewhat subjective measure depending on the adjustments applied (Damodaran, 2016; Palepu & Healy, 2012). However, this research adopts an objective approach by deriving EBITDA directly from the available balance sheet data under investigation without any adjustments or normalizations.

Although EBITDA serves as a key profitability indicator, it does not fully represent the actual cash available to investors. A more precise measure of a company's financial flexibility is Free Cash Flows (FCF), which adjusts EBITDA for taxes, Capital Expenditures (CAPEX), and changes in working capital. Since FCF represents the cash available for reinvestment after fulfilling operational and capital expenditure requirements, it is widely regarded as a more precise indicator of a company's financial health (Damodaran, 2016).

1.3.3 Challenges in SME valuations using public financial data

The lack of financial transparency presents a structural challenge in SME M&A transactions, significantly impacting M&A advisors' ability to accurately assess a company's true value. While the multiple in the EBITDA-multiple method can be derived from industry standards, estimating EBITDA, the other key component of the valuation, is considerably more complex (Bagna & Ramusino, 2017; Mauboussin, 2018). Unlike publicly listed companies, which disclose detailed financial statements, most Dutch SMEs are only required to publish a balance sheet. However, a subset of SMEs, known as *"controleplichtige MKB'ers"* (SMEs subject to audit requirements), must disclose full financial statements if they meet at least two of the following three criteria for two consecutive years: (1) a balance sheet total exceeding €7.5 million, (2) revenue above €15 million, or (3) more than 50 employees (Dagblad, 2025). Since most SMEs fall below these thresholds, EBITDA estimation from balance sheet data becomes essential for early-stage valuations (Overheid.nl, 2025).

As balance sheets lack critical financial metrics such as revenue, gross margin, operating expenses, and net profit, direct EBITDA calculation is not feasible (Overheid.nl, 2025). This absence of readily available financial information complicates the initial valuation process and increases information asymmetry between Marktlink and their acquisition targets (Overheid.nl, 2025). To mitigate this challenge, Marktlink currently relies on a set of assumptions to approximate EBITDA. One commonly applied "rule of thumb" is estimating revenue using a multiplier of 5.5 times accounts receivable (debtors), from which EBITDA can then be derived using averages for the EBITDA margin. This approach provides a practical benchmark that can later be compared to the results of the developed

model to assess its accuracy. A high-level overview of these structural challenges, relating to limited data access, outdated reporting, and dependence on private disclosures, is presented in Figure 1.2.

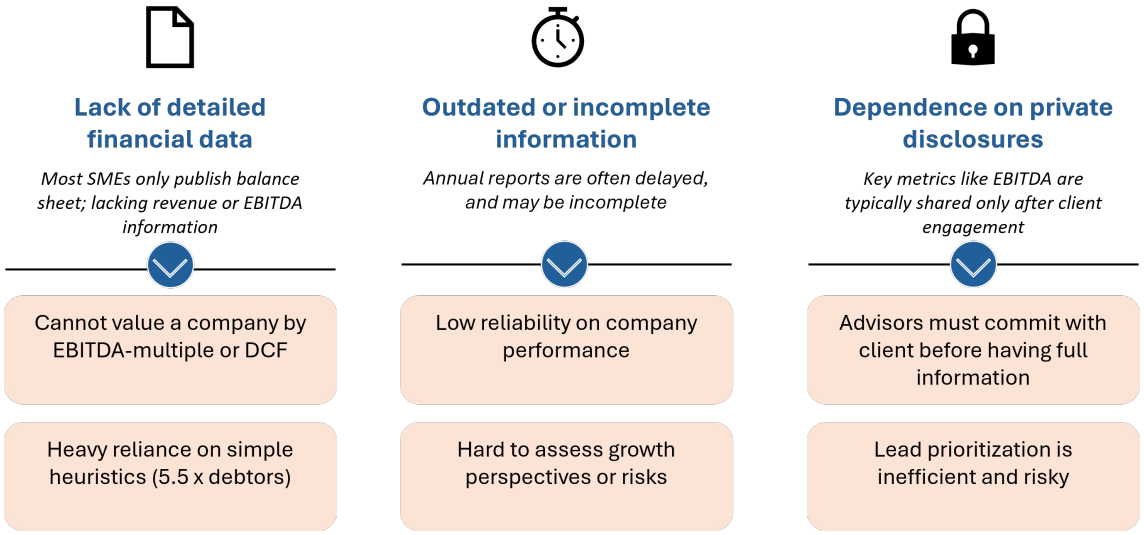


FIGURE 1.2: Key challenges in SME valuation using public financial data

Limited availability of financial data

Dutch SMEs are generally not required to publicly disclose detailed financial statements (Overheid.nl, 2025). As a result, essential financial information remains inaccessible during the initial stages of deal assessments. Critical figures, such as revenue, profitability, and cost structures, typically become available only after engagement with the seller and access to annual financial statements is granted. This lack of transparency poses significant challenges for M&A advisors, particularly in the identification and evaluation of potential acquisition targets. Furthermore, relevant financial data is often fragmented across various sources, including trade registries and industry reports, complicating the process of constructing a comprehensive financial profile of a company.

Outdated or incomplete financial information

SMEs are not required to update financial statements frequently, often resulting in reports that lag by approximately one year and may not accurately reflect the company's current performance (Kamer van Koophandel (KvK), 2025). This time lag in financial reporting can lead to misinterpretations of the company's financial health, particularly in rapidly growing or financially volatile businesses. Furthermore, incomplete disclosures often omit critical details such as outstanding liabilities or cash flow adjustments, further complicating the assessment of EBITDA (Overheid.nl, 2025).

Dependence on private disclosures and seller-provided data

Therefore, in many SME transactions, advisors must rely on voluntary financial disclosures from the client to execute accurate value propositions. The disclosure of private financial data often occurs after at least the introduction to the M&A advisor, according to Marktlink experts. The reliance on private disclosures confirms that currently, it is

indeed difficult to estimate a company value on publicly available financial data and that there is a need for a model that can execute this estimation.

1.3.4 The need for preliminary valuations

In the Dutch SME M&A market, it is crucial for M&A advisors such as Marktlink to evaluate a company's potential value before formally approaching the business owner. Conducting a preliminary valuation enables Marktlink to efficiently identify and prioritize high-potential acquisition targets. Without this early-stage assessment, the process becomes less effective, leading to inefficient resource allocation and the pursuit of businesses that may not align with the firm's criteria. To ensure strategic alignment, Marktlink generally considers only companies with a minimum EBITDA of €500K, with some exceptions, making this threshold a key determinant in identifying viable acquisition prospects. A schematic overview of the lead screening process at Marktlink, and the precise stage where the core valuation problem emerges, is shown in Figure 1.3:

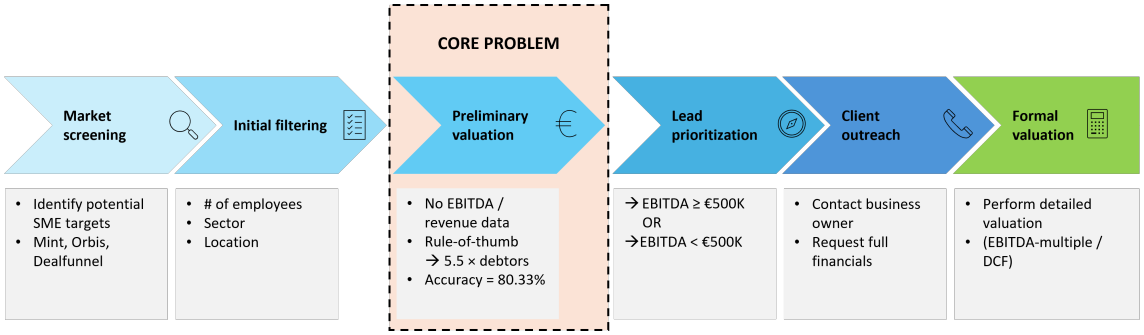


FIGURE 1.3: Lead screening process in SME M&A and location of core valuation challenge

1.4 Problem statement

Accurate SME valuation is a critical component of M&A transactions, allowing advisors to make well-informed acquisition decisions. However, as outlined in Section 1.3, publicly available financial data often lacks the depth and accuracy needed for reliable early-stage valuations resulting in high-potential leads. The problem mainly results from the lack of precision in estimating EBITDA based solely on balance sheet figures. The ultimate challenge then arises: M&A advisors struggle to assess the true EV of an SME before engaging with the company, leading to inefficiencies in the deal-sourcing process.

In this research, the Managerial Problem-Solving Method (MPSM) by Heerkens and van Winden (2017) is applied to systematically define the cause-and-effect relationships underlying the challenges in SME valuation. Using this approach, the problem is structured by identifying how various underlying issues, such as limited financial transparency, outdated data, and reliance on private disclosures, contribute to inaccurate valuation estimates for EBITDA, as shown in Figure 1.4.

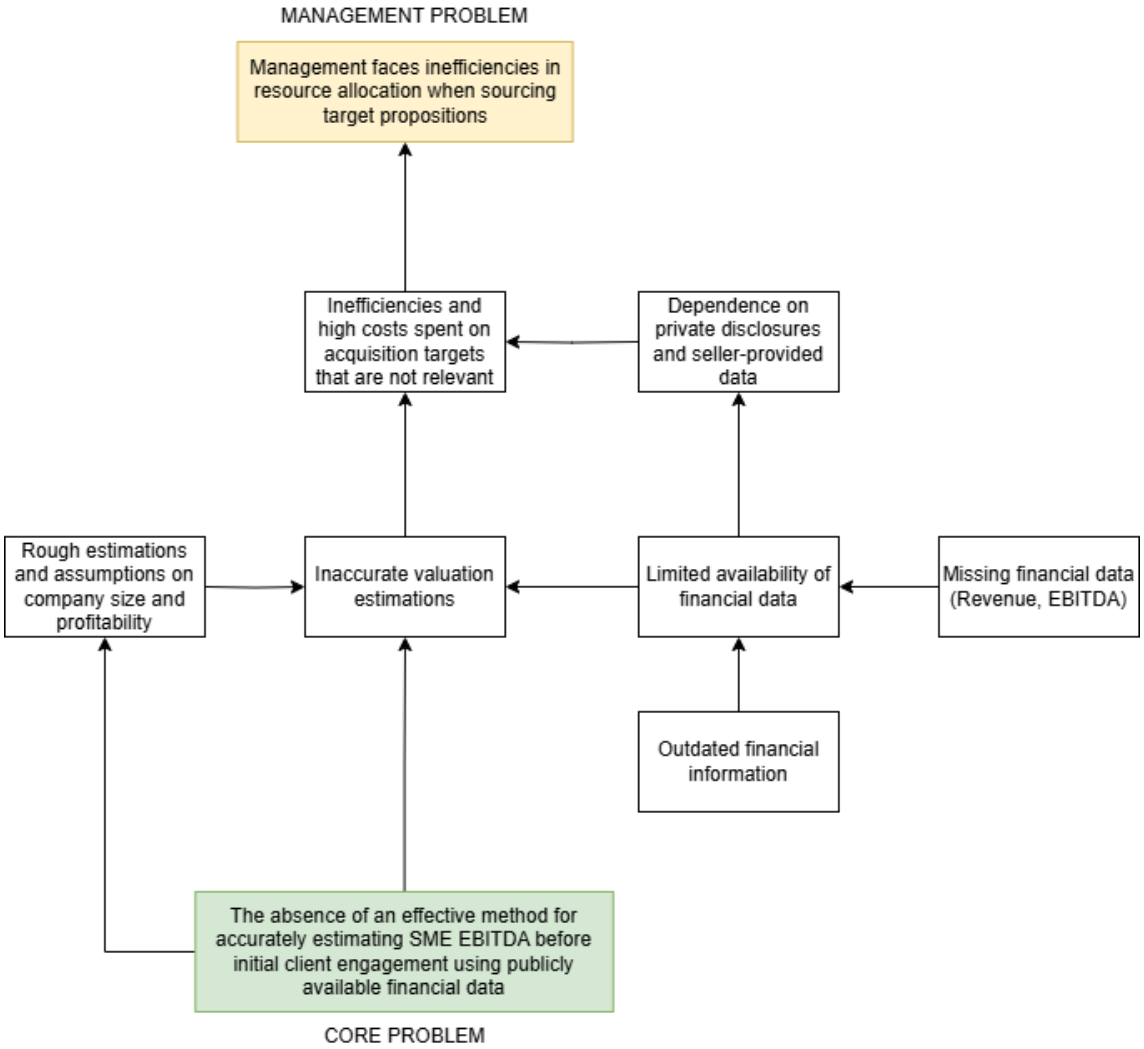


FIGURE 1.4: Problem cluster Marktlink SME valuation

By mapping these causal relationships, the core problem can be derived:

M&A advisors lack a reliable method for accurately estimating SME EBITDA before initial client engagement using publicly available financial data, resulting in inaccurately valued target propositions and inefficiencies in lead sourcing

1.5 Research design

To address the research objective and answer the core questions, a structured research design was developed combining both qualitative and quantitative methods. This approach ensures a comprehensive exploration of theoretical valuation concepts while testing different models using publicly available financial data. The design also supports practical validation through expert input and model benchmarking.

1.5.1 Research goal

The purpose of this research is to develop a data-driven method for estimating SME EBITDA using publicly available financial data, addressing the limitations of traditional valuation approaches that rely heavily on assumptions in early-stage assessments or private financial disclosures in later stages of the deal process. This study will evaluate and compare different modeling techniques to determine their accuracy in EBITDA estimation. The best-performing model will then be selected and implemented to improve early-stage target screening and valuation efficiency in SME M&A transactions. By enhancing the accuracy and accessibility of financial estimations, this research seeks to provide Marktlink with a scalable, reliable, and efficient tool for pre-assessing acquisition targets, reducing information asymmetry, and minimizing valuation errors. Hereby, Marktlink is given a competitive advantage over its competitors by being able to optimize their deal sourcing process.

1.5.2 Research questions

To address the core problem (Section 1.4) and achieve the research objective, research questions have been formulated to provide structure to this study. The main research question has been formulated as:

How can a method be developed to accurately estimate SME EBITDA based on publicly available financial data, enabling M&A advisors to improve early-stage valuation accuracy and lead sourcing efficiency?

To support the investigation of the main research question and to provide a clear structure for the study, the following sub-questions have been formulated:

- **RQ 1:** *What are the most commonly used valuation methods in M&A in the SME sector, and how do they compare in terms of applicability and reliability when using public data?*
- **RQ 2:** *What factors influence EBITDA of SMEs, and what are the key challenges in estimating EBITDA solely based on public financial data?*
- **RQ 3:** *How can publicly available financial data be utilized to estimate SME EBITDA?*

- **RQ 4:** *How can statistical or machine learning models be leveraged for EBITDA estimations by using publicly available financial data?*
- **RQ 5:** *How can the predictive model be evaluated, and what practical insights does it offer for EBITDA estimation?*

1.5.3 Research methods

To answer the research questions and achieve the research objective, this study applies a structured approach that combines qualitative and quantitative methods. Each research question is assigned a suitable method based on the type of data required. Qualitative methods are used for theoretical exploration, expert insights, and industry analysis, while quantitative methods focus on data collection, model development, and validation. Below each research question has been categorized by its corresponding research method.

- *Qualitative research*
  RQ1: A literature review will be conducted to examine established valuation methods, including EBITDA multiples, DCF, and asset-based approaches. Academic papers, industry reports, and M&A case studies will be analyzed to compare their effectiveness in SME transactions.
  RQ2: A qualitative literature study will identify factors that influence SME EBITDA and highlight the constraints of public financial data in estimating this metric . This will be supported by an analysis of industry papers and expert opinions on data limitations in SME M&A.

- *Quantitative research*
  RQ3: A quantitative data analysis will be performed by identifying and assessing publicly available datasets (e.g., trade registries such as Orbis). The quality, completeness, and reliability of this data source will be evaluated to determine its suitability for valuation modeling. After data processing, an exploratory data analysis will be executed to select features and tune parameters. Finally, back-testing techniques are applied to ensure validated feature selection.
  RQ4: A quantitative study will assess how statistical and machine learning models can be leveraged for EBITDA estimation using publicly available financial data (Alanazi, 2025; Fischer & Krauss, 2017). As a first step, EBITDA is estimated through classification, followed by regression for precise estimates on EBITDA. Estimation accuracy for regression based models is measured using performance metrics such as R^2, bias, Symmetric Mean Absolute Percentage Error (SMAPE), and scatterplots of the actual EBITDA versus the predicted EBITDA. Finally, a comparative analysis will evaluate differences in prediction accuracy between the proposed statistical and machine learning approaches, as well as the initial situation.
  RQ5: A quantitative evaluation framework is applied to assess model performance and validity. This includes cross-validation to test robustness, benchmarking against a baseline method currently used in practice, and an error analysis to identify blind spots. Additionally, a simplified DCF-based sanity check verifies whether predicted EBITDA values yield economically consistent outcomes.

1.6 Scope of the research

This research aims to develop an EBITDA estimation model for SMEs that do not publicly disclose income statements. However, a subset of Dutch SMEs (*controleplichtige MKB'ers*) publishes full financial statements, allowing them to serve as a validation dataset for benchmarking the model's accuracy. By testing the model on both non-disclosing SMEs and those with publicly available income statements, this study enhances its reliability. Prior research has demonstrated that data-driven modeling techniques can effectively estimate SME financials using publicly accessible sources (Stokkers, 2024). However, EBITDA estimation is highly sector-dependent, as financial performance, cost structures, and profit margins vary significantly across industries (Johnsen & McMahon, 2005). To improve accuracy, the model will incorporate sector-specific financial patterns, ensuring a more precise estimation approach.

This study is further limited to SMEs in the Netherlands and in Belgium, as defined in Appendix A.1, aligning with Marktlink Netherlands' target market. Large publicly traded companies are excluded due to their significantly different financial transparency and reporting requirements. The model development will be based on historical transaction data, financial databases, and industry benchmarks, with validation conducted using real-world SME data. This research does not seek to replace due diligence but rather to enhance early-stage lead screening for M&A advisors.

For clarity, the term "valuation model" in this study refers specifically to the estimation of EBITDA, a fundamental financial indicator in SME valuation. While EV is typically derived using the EBITDA-multiple method, this research primarily focuses on improving the accuracy of EBITDA estimations based on publicly available financial data. Given the sector-specific nature of financial performance, the final model may apply industry-specific EBITDA-multiples to approximate EV, ensuring that valuation estimates align with sectoral benchmarks.

By integrating sector-specific financial patterns and multiples, this research aims to provide a comprehensive and reliable approach to Dutch SME valuation, tailored to Marktlink's deal-sourcing and advisory processes.

1.7 Thesis outline

To systematically address the research objective and core problem, this study is structured into several key chapters. Each chapter builds upon the previous one to ensure a logical flow from theoretical foundations to practical implementation.

Chapter 2: Literature review
Chapter 2 establishes the theoretical foundation of the research by examining existing SME valuation methodologies, the role of publicly available financial data, and the financial determinants of EBITDA. It discusses the strengths and limitations of market-based, income-income, and asset-based valuation approaches, and analyzes key profitability drivers such as leverage, liquidity, firm size, and sector. The chapter also highlights the challenges posed by limited financial disclosures and regulatory differences, reinforcing the need for improved modeling techniques in data-constrained valuation environments.

Chapter 3: Data selection and preparation
Chapter 3 focuses on the identification, selection, and processing of publicly available financial data to develop a reliable SME EBITDA estimation model. It explores relevant data sources, evaluates the data, and applies data transformations to ensure the dataset is suitable for model training and validation. Key aspects include addressing missing data, standardizing financial variables, and selecting relevant features that influence SME EBITDA.

Chapter 4: Model application and evaluation
Chapter 4 develops and evaluates models to estimate SME EBITDA using supervised learning. It compares classification and regression methods, on normal and log-transformed data. Model accuracy is assessed using R^2, bias, and SMAPE, while SHapley Additive exPlanations (SHAP) analysis explains feature importance. The model is benchmarked against Marktlink's rule-of-thumb, validated through cross-validation, and further assessed via a critical error analysis and a stylized DCF check for economic validity.

Chapter 5: Conclusion
Chapter 5, the final chapter, presents the conclusion, discusses key limitations, and provides recommendations for practice and future research. It answers the main research question based on the empirical findings, followed by a reflection on data limitations and external validity, and suggests next steps such as model extension and external validation.

Chapter 2

Literature Review

Contents

This chapter provides a structured review of business valuation methodologies, focusing on their applicability to SMEs and the challenges of using publicly available financial data. Section 2.1 outlines the methodology used for the literature review, ensuring a systematic selection of relevant sources. Section 2.2 examines the three primary valuation approaches, market-based, income-based, and asset-based, highlighting their assumptions, advantages, and limitations. Section 2.3 explores the concept of SME EBITDA estimation, addressing financial disclosure regulations, key factors influencing EBITDA, and the limitations of public financial data. Finally, Section 2.4 summarizes the key insights from the literature, synthesizing the findings on valuation methodologies, financial determinants, and the limitations of publicly available data, highlighting the need for more precise financial modeling in SME valuation.

2.1 Methodology of the Literature Review

This literature review follows a structured methodology to ensure a comprehensive and relevant selection of sources. The research adheres to defined inclusion and exclusion criteria (Table A.3), prioritizing peer-reviewed articles, conference papers, and government reports published since 2000. The selected literature focuses on SME valuation methodologies, financial structure, and the usability of public financial data for EBITDA estimation.

Academic databases such as Scopus, Web of Science, and the University of Twente Library (Table A.5) were used to gather high-quality sources. Only studies relevant to privately held SMEs were considered, ensuring applicability to non-public firms. Key concepts, including EBITDA, financial structure, and valuation methods, were systematically mapped to structure the analysis (Table A.4). The full list of reviewed sources, along with their respective authors and keywords, is provided in Table A.6, offering an overview of the literature used in this study.

This methodological approach ensures a replicable literature review, forming a strong foundation for analyzing SME valuation methods and the role of publicly available financial data.

2.2 Business valuation methodologies of SMEs

Valuing a business is a critical step in the M&A process. Various valuation methods exist, each with distinct assumptions, advantages, and limitations (Bancel & Mittoo, 2014). The choice of valuation approach usually depends on factors such as industry characteristics, financial transparency, and the availability of market data. Additionally, the position that is taken in the acquisition process, either being the bidding or selling firm, influences how valuation is perceived (Damodaran, 2012).

This section examines the three primary valuation methods used in SME M&A: the market-based, income-based, and asset-based approach (Nenkov & Hristozov, 2022). The market approach relies on comparable company and transaction multiples to estimate value based on industry benchmarks (Żelazowski, 2015). The income approach determines value based on expected future cash flows, discounted to present value, making it a theoretically sound but assumption-dependent method (Beranová, 2013). The asset-based approach, on the other hand, values a company based on its net assets and is often used in liquidation scenarios or asset-heavy businesses (Jenkins & Kane, 2006).

2.2.1 Market-based approach

The market-based approach is a recognized valuation method in the finance industry, relying on comparative analysis to estimate a company's value based on previously traded similar businesses. This approach is classified as "market-based" because valuation is derived from actual market behavior, reflecting prices set by investors in real transactions (Bancel & Mittoo, 2014). The fundamental assumption underlying this method is that market participants collectively determine a company's value more accurately than intrinsic financial models, as their pricing decisions are informed by real-world transaction data and investor sentiment. The market-based approach can be further categorized into Comparable Company Analysis (CCA)) and a Comparable Transaction Analysis (CTA).

Comparable company analysis

The CCA is a relative valuation method that estimates a company's value by benchmarking its financial metrics against similar publicly traded firms. This approach operates under the assumption that companies with comparable business models, industry characteristics, and financial structures will trade at similar valuation multiples. The process involves selecting comparable peers, gathering their financial data, and deriving valuation multiples such as the Enterprise Value-to-EBITDA (EV/EBITDA) or Price-to-Earnings (P/E) ratio. These multiples are then applied to the financial figures of the subject company to estimate its market value. While EV/EBITDA is commonly used for valuing private companies due to

its focus on operational performance, the P/E ratio is primarily applied to publicly traded firms, where Earnings Per Share (EPS) is a key valuation metric (Meitner, 2006). Although CCA lacks theoretical fundament, it provides a market-conform valuation of businesses.

However, the applicability of this methodology in the SME sector is constrained by the limited availability of publicly listed comparable firms. The study by Bowman and Bush (2006) emphasizes that the accuracy of CCA relies on selecting firms with similar financial metrics. However, since most SMEs do not publicly disclose their financial data, identifying appropriate comparables becomes challenging. As a result, the scarcity of relevant market data significantly reduces the effectiveness of CCA for SME valuation, potentially making it impractical in many use cases (Bowman & Bush, 2006).

Comparable transaction analysis

The CTA, also known as precedent transaction analysis, is another commonly used valuation approach in M&A (Damodaran, 2006). This method derives a company's estimated value by analyzing historical transaction data of comparable firms that have been recently acquired. Unlike the CCA, which focuses on publicly traded firms, CTA examines actual deal prices, providing insights into realized market valuations rather than theoretical estimates. The key assumption behind CTA is that past M&A transactions serve as relevant benchmarks for determining a company's EV in the current market conditions (Bagna & Ramusino, 2017; Żelazowski, 2015).

The CTA methodology involves several steps. First, comparable transactions are selected based on industry, company size, and other publicly available variables. Next, valuation multiples, such as EV/EBITDA, EV/Revenue, or EV/EBIT , are derived from these transactions (Bagna & Ramusino, 2017). These multiples are commonly used methodologies, especially because they neutralize the impact of financed debt. Finally, these multiples are applied to the financial metrics of the target company to estimate its potential value. In formula form this results in the following equation:

$$EV = EBITDA * Multiple \tag{2.1}$$

Where:

- ***EV*** = Enterprise Value
- ***EBITDA*** = Earnings Before Interest, Taxes, Depreciation, and Amortization
- ***Multiple*** = Multiplication factor

To demonstrate the practical application of this valuation method, an illustrative example is presented in Table 2.1. In this case, the EV/EBITDA multiple method is applied, consistent with standard practices in real-world valuation scenarios. The example features three companies representing different SME classifications: micro, small, and medium-sized enterprises.

The advantage of CTA is that it reflects actual deal-making conditions, capturing market or industry behavior, negotiation effects, and synergies reached in completed transactions (Żelazowski, 2015). However, the method's effectiveness depends on the availability of relevant and recent transaction data, which can be limited in private markets, particularly for SMEs. Although this consideration is important, it does not appear to pose a significant issue, as multiple-based valuation methods derived from precedent transactions remain among the most widely applied approaches in company valuation in the SME sector, particularly the EV/EBITDA multiple (Żelazowski, 2015).

	Company 1	Company 2	Company 3
EBITDA (€m)	0.5	2	10
Multiple	8x	5x	6x
Enterprise Value (€m)	4	10	60

TABLE 2.1: An example situation where the EBITDA-multiple method is shown.

A key consideration in applying the EBITDA-multiple method is the subjectivity involved in defining EBITDA itself. While EBITDA is typically calculated on an annual basis, its application in M&A transactions often extends beyond a single year. Entrepreneurs and M&A advisors frequently adjust EBITDA by averaging it over multiple years, excluding outlier years, or normalizing earnings to reflect future performance expectations (Damodaran, 2016; Koller et al., 2012). These adjustments account for extraordinary expenses, one-time revenues, or projected synergies, making EBITDA an inherently flexible but somewhat subjective measure in valuation practices. As a result, different valuation practitioners may arrive at varying EBITDA figures for the same company, depending on their assumptions and financial adjustments. This EBITDA is often referred to as the Adjusted EBITDA, another type of EBITDA than is under investigation in this research (Palepu & Healy, 2012).

However, the EV/EBITDA multiple method also has notable limitations that can impact valuation accuracy. One key challenge is the selection of comparable enterprises, as differences in financial structures, industry positioning, and accounting practices can affect valuation outcomes if comparability is not carefully defined (Żelazowski, 2015). Additionally, the method does not account for capital expenditures, which may lead to an understatement of a company's financial health, especially in capital-intensive industries. Moreover, EV/EBITDA does not explicitly incorporate business risk, such as operating leverage, meaning two firms with similar EBITDA values but differing risk profiles may receive comparable valuations despite having fundamentally different financial stability (Mauboussin, 2018; Ribal et al., 2010). Lastly, the reliance on current market conditions introduces a static valuation approach, limiting its ability to reflect future growth potential or downturn risks (Żelazowski, 2015). Despite these drawbacks, EV/EBITDA remains a widely accepted valuation tool due to its practicality, efficiency, and ability to provide quick market-driven estimates (Bagna & Ramusino, 2017).

2.2.2 Income-based approach

The income-based approach is an intrinsic valuation method that determines a company's value based on its expected future earnings or cash flows, discounted to their present value (Beranová, 2013). This approach assumes that a company's worth is derived from its ability to generate future profits. The DCF method is the most commonly used income-based valuation technique, as it provides a theoretical framework for estimating EV based on projected financial performance (Nenkov & Hristozov, 2022).

Discounted Cash Flow method

The DCF method estimates a company's value by forecasting its FCF over a specific period and discounting them using an appropriate discount rate, typically the WACC, to the Present Value (PV) (Markus & Rideg, 2021). This, subsequently results in the EV of company by applying the following formula (Steiger, 2008):

$$EV = \sum_{t=1}^{n} \frac{FCF_t}{(1 + WACC)^t} + \frac{TV}{(1 + WACC)^n} \tag{2.2}$$

Where:

- $\boldsymbol{EV}$ = Enterprise Value
- $\boldsymbol{FCF}_t$ = Free Cash Flow in year t
- $\boldsymbol{WACC}$ = Weighted Average Cost of Capital
- $\boldsymbol{TV}$ = Terminal Value, representing the company's value beyond the forecast period

The DCF method is theoretically sound, as it focuses on a firm's intrinsic value rather than being influenced by market fluctuations. However, DCF valuation is highly assumption dependent, requiring precise estimates of FCF, growth rates, and the WACC (Steiger, 2008).

For SMEs, these assumptions pose significant challenges due to earnings volatility, uncertain growth trajectories, and difficulties in determining WACC. Additionally, Gama and Geraldes (2012) highlights the challenges of cash flow measurement in SMEs, pointing out that cash flow serves as a better performance indicator than earnings, given its direct link to financial health. SMEs often face constraints in capital structure decisions and limited access to financing, which can introduce further inaccuracies in DCF modeling.

Gama and Geraldes (2012) highlight the two main ways companies allocate their cash flow: toward past obligations, such as debt repayments and dividend distributions, or toward future growth, including CAPEX and innovation. Their findings indicate that firms investing more heavily in future-oriented initiatives, like CAPEX, tend to enhance their competitiveness more significantly than those prioritizing past-oriented cash flows. However, DCF valuation often struggles with CAPEX-heavy SMEs, as it assumes stable reinvestment rates over time. In reality, SMEs frequently experience irregular CAPEX cycles, with periods of heavy investment followed by reduced expenditures. This variability makes it difficult to project future free cash flows accurately, introducing volatility into DCF-based valuations. Moreover, forecasting cash flows for SMEs is inherently more challenging due to their higher business risk and greater uncertainty compared to large publicly traded companies. The reliance on assumptions in DCF modeling further complicates its application, as small changes in projected growth rates, or discount rates can significantly alter valuation outcomes. This assumption-driven nature also makes DCF susceptible to manipulation, as inputs can be adjusted to produce more favorable valuations. These challenges underscore the importance of understanding the dynamics of cash flow allocations in DCF modeling to ensure realistic and reliable valuation estimates.

Given these complexities, SME valuation practices often combine DCF with market-based approaches, particularly EV/EBITDA multiples, which rely on observable transaction data and reduce dependency on volatile cash flow projections.

2.2.3 Asset-based approach

The asset-based valuation approach determines a company's value based on the net worth of its assets, rather than its expected future earnings or market comparisons. This method calculates a firm's value by assessing the book value, fair market value, or liquidation value of its tangible and intangible assets, subtracting total liabilities to estimate the company's net asset value. It is particularly relevant for businesses with significant tangible assets, such as real estate firms, manufacturing companies, and capital-intensive industries.

While the asset-based approach provides a clear and objective valuation framework, its applicability to SMEs is limited in cases where intangible assets, goodwill, or earnings potential play a critical role. Many SMEs derive their value from brand reputation, customer relationships, and intellectual property, which are often undervalued or omitted in asset-based calculations. Additionally, this approach fails to capture future earnings potential, making it less suitable for businesses focused on growth. Consequently, while asset-based valuation can serve as a baseline estimate, it is often combined with market-based or income-based methods for a more comprehensive valuation (Jenkins & Kane, 2006).

2.2.4 Most commonly used valuation method in practice

Among the various valuation methodologies, the EV/EBITDA multiple has become the most widely used approach in SME M&A. Rooted in the CTA framework, this method is favored for its simplicity, market-driven nature, and ability to provide a quick estimation of EV without requiring extensive financial forecasting (Bagna & Ramusino, 2017; Żelazowski, 2015).

The EV/EBITDA multiple is particularly valuable in SME valuations as it reflects actual market conditions and investor sentiment, drawing on observed transaction data rather than theoretical financial projections (Beranová, 2013). Given the challenges of the DCF method, both M&A advisors and business owners seeking to sell their companies often prefer multiple-based approaches, either as a standalone method or in combination with DCF for validation purposes (Nenkov & Hristozov, 2022; Steiger, 2008).

Due to its widespread application in SME transactions and its practicality in real-world deal-making, this study adopts the EV/EBITDA multiple as the primary valuation method.

2.3 Challenges in using publicly available financial data for SME EBITDA estimation

2.3.1 Regulations on financial disclosure of SMEs

Financial disclosure regulations determine the extent to which SMEs are required to report their financial performance, directly influencing the accessibility of publicly available financial data. Unlike large publicly traded firms, SMEs often follow simplified reporting requirements that vary based on firm size and regulatory frameworks (Overheid.nl, 2025). These regulations are designed to reduce the administrative burden on smaller enterprises; however, they also lead to significant gaps in financial transparency, affecting valuation accuracy and comparability (Andreeva et al., 2016).

The degree of financial disclosure is typically determined by factors such as firm size, revenue thresholds, and legal structure. Many countries implement tiered reporting systems, where micro and small enterprises are excluded from extensive financial reporting,

while medium-sized firms are subject to more comprehensive standards (Overheid.nl, 2025; Schammo, 2018). In the European Union, the Accounting Directive (2013/34/EU) establishes reduced disclosure requirements for SMEs, allowing them to submit financial statements without detailed income or cash flow statements (Schammo, 2018). The Dutch national government aligns its regulations with this directive, requiring most SMEs to disclose only their balance sheet, without publishing income or cash flow statements (Overheid.nl, 2025). However, an exception applies to "controleplichtige MKB'ers", these firms are legally required to disclose full financial statements, including their income statement (Overheid.nl, 2025).

While these exceptions facilitate SME growth and minimize compliance costs, they present substantial challenges for financial analysis. The absence of standardized financial statements, particularly income statements, makes it difficult to assess key financial metrics such as EBITDA, further complicating valuation processes for Dutch SMEs (Johnsen & McMahon, 2005; Parliament & of the European Union, 2013).

2.3.2 Factors influencing EBITDA of SMEs

The profitability of SMEs is shaped by a combination of financial, operational, and external factors, with EBITDA serving as a key indicator of a firm's earnings potential. Understanding the variables that influence EBITDA is essential for accurate financial analysis and valuation (Malakauskas & Lakštutienė, 2021). Prior research has identified a wide range of determinants, including firm-specific characteristics such as size, age, financial structure, and industry sector, as well as broader market dynamics and external conditions (Tong & Serrasqueiro, 2020). These factors collectively impact a firm's ability to generate earnings and maintain financial stability.

Key attributes influencing EBITDA include financial leverage, liquidity, activity, and coverage ratios, all of which provide insights into a company's financial health (Gama & Geraldes, 2012). Additionally, structural characteristics such as firm size and industry classification play a crucial role in shaping profitability, as different sectors exhibit varying cost structures, revenue models, and capital requirements. By analyzing these determinants, this section aims to contribute to a more precise understanding of how different factors drive EBITDA in SMEs.

Firm's size and age

Firm size and age are key determinants of profitability, influencing cost efficiency, operational effectiveness, and financial stability. Larger firms benefit from economies of scale, reducing per-unit costs and improving profitability. However, as firms expand, they may face diseconomies of scale due to increasing bureaucratic complexity, coordination challenges, and rising administrative costs, which can erode profit margins (Bartlett & Bukvič, 2001; Lwango et al., 2017). Firm size, often measured in terms of the number of employees, therefore affects profitability, as structured operations can enhance efficiency but reduce adaptability (Strategic Direction Editorial Team, 2014). According to the passive learning theory, firms improve efficiency over time, yet sustaining high profitability becomes more challenging as complexity increases (Tong & Serrasqueiro, 2020).

Firm age similarly impacts profitability. Younger firms, while agile and innovative, often lack financial stability and experienced management, making them more vulnerable to any missteps (Dunne & Hughes, 1994). In contrast, mature firms benefit from established market positions and refined operations but may experience declining profitability due to rigid cost structures and reduced flexibility (Lotti et al., 2003; Lwango et al., 2017). Over

time, growth stagnation and increasing fixed costs can further constrain profit margins (Lwango et al., 2017; Tong & Serrasqueiro, 2020). This confirms that firm size and age significantly influence profitability, reinforcing their relevance in this study as key factors affecting EBITDA.

Sector of operation

The sector of operation of a firm plays a significant role in shaping its financial structure and profitability. Industry characteristics, such as capital intensity, risk exposure, and asset structure, influence financial decision-making, which in turn affects a company's profitability levels. In practice, a firm's sector classification is typically indicated by the "Standaard Bedrijfsindeling (SBI)" code, which helps categorize businesses based on their primary economic activity (Kamer van Koophandel (KVK), 2025). SMEs in capital-intensive industries, such as manufacturing for example, tend to rely more on debt financing, which can impact their EBITDA margins (Tong & Serrasqueiro, 2020). In contrast, service-based firms, which often have fewer tangible assets, tend to maintain lower debt levels and may exhibit higher EBITDA margins. Additionally, sector-specific growth patterns contribute to variations in profitability, as firms in high-growth industries may experience greater profit volatility, while those in stable industries generate more predictable but lower-margin earnings. These findings highlight the importance of incorporating industry-specific financial behaviors when estimating EBITDA, as sectoral differences significantly influence a firm's ability to generate profits (Johnsen & McMahon, 2005). This sector-specific variation is also the reason why this study adopts a sector-based approach to estimating EBITDA, ensuring that differences in cost structures, revenue models, and profitability dynamics are properly accounted for.

Financial structure

A firm's financial structure reflects its mix of debt and equity, influencing profitability. This can be further divided into leverage, liquidity, coverage, and activity, where each aspect covers another aspect of a firm's financial status. Leverage measures debt reliance, impacting financial stability. Liquidity, including working capital, indicates short-term financial health. Activity ratios assess operational efficiency in utilizing assets and managing receivables. These factors are crucial for understanding a firm's EBITDA and overall performance (Gama & Geraldes, 2012; Malakauskas & Lakštutienė, 2021).

- *Leverage*
  Leverage significantly influences SME profitability by affecting financial stability, growth potential, and risk exposure. While access to external financing can drive expansion, excessive debt increases financial risk and can lead to greater financial constraints, limiting reinvestment capacity and operational flexibility, which in turn may affect EBITDA generation (Tong & Serrasqueiro, 2020).

  Key leverage indicators include solvency ratios like the Debt-to-Assets ratio, which measures a firm's financial stability by assessing the proportion of total assets financed by debt. Other critical indicators include Total Liabilities and Long-term Liabilities, reflecting a firm's reliance on debt financing. The Debt-to-EBITDA ratio assesses debt sustainability, while Current Liabilities indicate short-term financial pressure. Return on Assets (ROA) and Return on Equity (ROE) are also essential indicators, as they measure how efficiently a firm utilizes its assets and equity to generate profits. High leverage can inflate ROE if returns exceed the cost of debt,

but excessive debt can also erode overall profitability by increasing financial risk (Afrifa & Padachi, 2016). Conversely, Owner's Equity strengthens financial stability by reducing dependency on external debt (Malakauskas & Lakštutienė, 2021). High short-term debt and trade credit can further strain liquidity, limiting reinvestment capacity and negatively impacting profitability (Tong & Serrasqueiro, 2020). Thus, maintaining an optimal debt-equity balance is crucial for sustaining profitability.

- *Liquidity*
  The ability of an SME to efficiently manage liquidity directly impacts its financial stability and profitability. Firms with strong liquidity positions can meet short-term obligations, invest in growth, and avoid financial distress without excessive reliance on external debt (Haron, 2015). Key liquidity indicators include Cash Equivalents and Marketable Securities, which provide immediate financial flexibility, and Current Assets, which reflect a firm's ability to cover short-term liabilities. Additionally, the composition of Total Assets influences liquidity, as firms with a larger asset base can leverage these resources for financing when needed. The Current Ratio, which measures a firm's ability to meet short-term liabilities with its current assets, serves as a fundamental indicator of liquidity strength, with higher ratios generally reflecting stronger financial health and lower insolvency risk.

  A crucial aspect of liquidity management is working capital. Research indicates that the relationship between working capital management and profitability follows a concave pattern, meaning that firms can maximize profitability at an optimal working capital level, beyond which profitability begins to decline. This is because firms with too little working capital may face disruptions in operations, loss of sales opportunities, and liquidity shortages, negatively impacting earnings. Conversely, excessive working capital ties up financial resources that could otherwise be reinvested in higher yielding opportunities, leading to inefficiencies and increased holding costs (Afrifa & Padachi, 2016).

  The trade-off between maintaining sufficient liquidity and minimizing idle resources is therefore a key financial decision for SMEs. Poor liquidity management, whether due to excessive short-term liabilities or overinvestment in working capital, can limit investment capacity and increase financial vulnerability, reducing EBITDA margins (Haron, 2015). To maintain optimal profitability, firms should aim to manage their working capital at an efficient level, ensuring that resources are neither excessively constrained nor inefficiently allocated.

- *Coverage*
  Coverage ratios, such as interest coverage ratio, indicate a firm's ability to meet its debt obligations from operating profits. The study emphasizes that firms with low interest coverage ratios are more likely to experience financial difficulties, as insufficient earnings to cover interest expenses increase the probability of default (Andreeva et al., 2016). Additionally, cash flow coverage is highlighted as an essential measure, reflecting a firm's liquidity position and ability to generate sufficient operating cash flow to sustain debt repayments (Markus & Rideg, 2021).

  The findings suggest that SMEs with strong coverage ratios, indicative of stable revenue streams and controlled debt levels, are less prone to financial problems. In contrast, firms with weak coverage ratios face higher financial risk, impacting their long-term profitability and survival (Andreeva et al., 2016). Therefore, maintaining adequate coverage ratios is crucial for SME financial health, as they directly influence

borrowing capacity, cost of capital, and overall financial stability.

- *Activity*
  The efficiency with which a firm utilizes its assets and resources to generate revenue, commonly referred to as "activity," plays a crucial role in determining profitability. Activity ratios measure how effectively a company manages its operational cycle, including inventory turnover, asset utilization, and receivables management. High activity levels generally indicate strong operational performance, as firms that can rapidly convert assets into revenue tend to achieve greater profitability (Malakauskas & Lakštutienė, 2021). Additionally, the level of shareholder funds influences a firm's ability to sustain high activity levels, as greater equity financing allows for operational flexibility and strategic investments without over-reliance on debt (Andreeva et al., 2016).

  Key indicators of activity include sales turnover, which reflects the speed at which a firm generates revenue, and working capital turnover, which assesses how efficiently a company utilizes its short-term assets and liabilities. Efficient firms maintain a balance between sales and resource allocation, preventing excessive capital lock-up in assets that do not directly contribute to profitability (Andreeva et al., 2016). Additionally, the asset turnover ratio provides insight into how effectively a company leverages its total assets to generate revenue, with higher ratios typically associated with stronger financial performance.

  Furthermore, the debt collection period influences a company's cash flow and profitability. Extended collection periods can create liquidity constraints, hereby increasing financial risk. SMEs with faster receivables turnover tend to maintain more stable cash flows, allowing for investment in growth opportunities, supporting profitability (Malakauskas & Lakštutienė, 2021). Moreover, the number of directors within a firm can impact decision-making and strategic oversight, affecting how effectively operational resources are allocated and utilized (Andreeva et al., 2016).

Management expertise

Managerial expertise within SMEs also plays a key role in profitability. Unlike larger firms with specialized managerial roles, SME owners must oversee all business functions, requiring adaptability and a strategic mindset. Research suggests that a higher level of education among SME owner-managers correlates positively with profitability, as it enhances problem-solving abilities and encourages external collaboration for business growth (Strategic Direction Editorial Team, 2014). To mitigate inefficiencies, SMEs can implement governance mechanisms, such as appointing specialized managers or diversifying decision-making, to enhance financial stability (Gama & Geraldes, 2012).

These findings underscore that asset and receivables management, along with financial governance, directly impact profitability. An overview of the discussed factors and their corresponding financial indicators is presented in Table A.2.

2.3.3 Financial ratio's as predictors in SME profitability

Research has demonstrated that while numerical balance sheet values provide useful financial insights, financial ratios exhibit particularly strong predictive power in estimating EBITDA (Batrancea et al., 2018). Ratios such as the current ratio, solvability ratio, and inventory ratio have been identified as significant explanatory variables in Multiple Linear Regression (MLR) models applied to SME profitability estimation (L. Li et al., 2023).

These ratios offer a normalized perspective on financial health by standardizing financial data relative to firm size, liquidity, and leverage, making them more effective than raw balance sheet values in capturing financial performance trends.

Empirical studies further reinforce the relevance of financial ratios in profitability modeling. The current assets ratio and equity-to-total liabilities ratio are consistently linked to higher financial performance, as they reflect a firm's liquidity position and financial stability (L. Li et al., 2023). Conversely, excessive reliance on trade receivables and a high debt-to-assets ratio negatively impact profitability due to increased credit risk and financial constraints. In sectors such as manufacturing, inventory management plays a particularly critical role, as firms with excessive stock levels experience reduced operational efficiency and tied-up capital (Batrancea et al., 2018).

Moreover, the predictive strength of financial ratios extends beyond individual indicators. Studies highlight that a combination of solvency, profitability, and liquidity ratios improves the accuracy of financial forecasting models (L. Li et al., 2023). By integrating multiple financial dimensions, these ratios provide a comprehensive view of an SMEs financial resilience, supporting their use as key input variables in financial modeling for EBITDA estimation.

2.3.4 Key limitations of public financial data

While financial structure indicators provide valuable insights into SME profitability, their accuracy depends heavily on the availability and reliability of public financial data. Incomplete financial disclosures, inconsistencies in reporting standards, and limited access to firm-specific data pose significant challenges in applying valuation models effectively. Understanding these limitations is crucial for refining financial estimations and improving the predictive power of valuation approaches.

The key limitations of public financial data, as discussed in the provided papers, primarily revolve around data availability, reliability, and completeness (Markus & Rideg, 2021). Andreeva et al. (2016) highlight that SMEs often do not provide full financial disclosures, leading to significant gaps in publicly available datasets (Haron, 2015). This issue is strengthened by differences in reporting standards across countries, making cross-country comparisons and valuations more challenging. Additionally, missing values are a common problem, and their treatment, whether through imputation or alternative statistical techniques, can introduce biases into financial analysis (Andreeva et al., 2016).

Another significant limitation is the inconsistency in the level of detail available for different firms. Andreeva et al. (2016) note that public sources often lack key financial indicators, especially for smaller enterprises, which can distort profitability assessments (Haron, 2015). Similarly, Malakauskas and Lakštutienė (2021) emphasize that simplified reporting requirements for SMEs result in less detailed financial statements compared to larger firms. This limitation restricts the ability to accurately apply valuation models, as essential inputs such as detailed revenue breakdowns, debt structures, and capital expenditures may be missing or inconsistently reported.

Furthermore, reliance on public data can lead to outdated or inaccurate financial assessments. Financial information in public databases is often updated with a delay, meaning that the most recent business performance indicators may not be reflected in the available records (Haron, 2015; Malakauskas & Lakštutienė, 2021). This issue is particularly relevant in fast-changing market environments where SMEs' financial health can fluctuate rapidly due to industry-specific developments, or operational challenges.

To address these limitations, practitioners often supplement public data with industry benchmarks, sector-specific averages, or proprietary datasets obtained through direct en-

gagement with firms (Andreeva et al., 2016; Haron, 2015). However, these workarounds introduce additional estimation errors and reduce comparability across different firms. As a result, the accuracy of SME valuations based solely on public financial data remains limited, highlighting the need for improved disclosure practices and enhanced data collection methodologies.

2.4 Conclusion on Literature Review

Understanding SME valuation requires a structured approach that considers various methodologies, financial determinants, and the constraints of publicly available data. This review explores the primary valuation methods, market-based, income-based, and asset-based approaches, assessing their applicability to SMEs and identifying key financial metrics used in valuation. Given EBITDA's role as a profitability indicator, this study examines factors influencing EBITDA, including firm size, age, industry sector, and financial structure, which encompasses leverage, liquidity, coverage, and activity ratios. The challenges of using public financial data are also analyzed, focusing on data limitations, inconsistencies in financial disclosures, and the reliability of reported figures. By highlighting these gaps, this review establishes the need for more precise financial modeling in SME valuation, ensuring a more accurate estimation of profitability based on publicly available financial data.

Chapter 3

Data Selection and Preparation

Contents

This chapter outlines the data selection and preparation process required for developing an accurate EBITDA estimation model. Section 3.1 covers data gathering and data cleaning, detailing the sources used, the inclusion of Dutch and Belgian SMEs, and the selection criteria applied to construct the dataset. It also addresses data cleaning procedures such as standardizing formats, handling missing values, and removing inconsistencies to ensure data integrity. Section 3.2 presents an exploratory data analysis, examining the correlation between financial variables and EBITDA across different sectors. Finally, Section 3.3 focuses on feature engineering and selection, discussing variable transformations and different feature selection strategies.

3.1 Data gathering and data cleaning

3.1.1 Data gathering

The dataset used in this study comprises 54,707 records and is derived from Moody's Orbis database, a financial information platform that includes data on both private and public companies worldwide (Moody's Analytics, 2025). Orbis is particularly relevant for this research as it provides standardized financial statements, industry classifications, and firm-level metrics across European markets, including Dutch and Belgian SMEs. Given the limited financial disclosure requirements for Dutch SMEs, Orbis represents a critical data source, offering one of the few accessible repositories of financial information for this segment.

The initial dataset was constructed by selecting only Dutch SMEs that met specific financial and structural criteria. To further expand the dataset, Belgian SMEs were also included. Unlike in the Netherlands, Belgian companies are subject to more extensive financial reporting requirements, which makes their financial statements publicly available (Amfico, 2025; Astro Tax, 2025). This increased transparency significantly contributed to the dataset's size, resulting in a total of 54,707 firms when both Dutch and Belgian SMEs are included.

The inclusion of Belgian SMEs is further justified by the structural similarities between the Dutch and Belgian SME markets. Both countries exhibit comparable economic environments and industry compositions, making Belgian firms a valuable addition for enhancing the reliability of Dutch SME valuation models (Rikkers & Thibeault, 2009; Steijvers, 2004). Furthermore, during the dataset construction process, additional selection criteria were applied, including company status, number of employees, and the range of reported EBITDA. Table 3.1 provides an overview of the criteria used to define the final dataset.

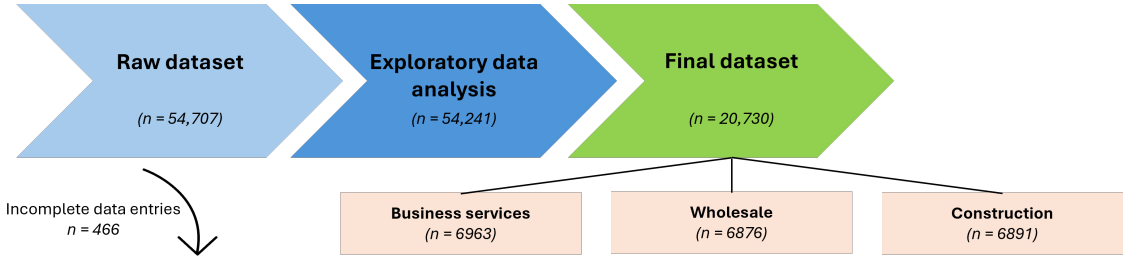


FIGURE 3.1: Dataset reduction process and final sector allocation

Figure 3.1 illustrates the refinement of the dataset from the raw Orbis extract to the final cleaned dataset. It also shows the division across three key sectors: business services, wholesale, and construction, which are further analyzed in Section 3.2.2.

Criterion	Description
Company status	Firms must be active, ensuring that only operational businesses are included.
Country	The company must be registered in the Netherlands or Belgium.
Employee count	The firm must have between 5 and 250 employees, excluding micro-enterprises while staying within the standard SME classification.
EBITDA	The company's reported EBITDA must be between €10K and €5M, including companies with a known EBITDA.

TABLE 3.1: Selection criteria for dataset extraction from Orbis

To ensure that the dataset accurately reflects the most recent financial performance of firms, the latest available fiscal year has been included. The majority of financial statements correspond to 2023, while for companies that have not yet reported their 2023 figures, the most recent available data from 2022 has been used.

To construct a dataset suitable for training the EBITDA estimation model, sixteen financial variables were selected from Moody's Orbis database. The selection process was guided by two key factors: the variables' presence on the balance sheet and their availability in Orbis. The resulting dataset incorporates company-specific attributes, key balance sheet components, and financial ratios, ensuring a comprehensive financial profile of SMEs.

A critical aspect of variable selection is the availability of data within Mint, the platform where Marktlink sources its deal-related financial information. Mint, as part of Orbis, provides financial data exclusively for Dutch SMEs. Given its role as a central repository for company data, integrating financial variables available in Mint ensures that the model aligns with the financial metrics typically used in Marktlink's deal-sourcing process. The available data primarily consists of balance sheet information and relevant financial ratios. However, certain balance sheet variables that are usually shown in Mint, such as financial fixed assets, interest-bearing liabilities, and cash and cash equivalents, were excluded due to their unavailability in Orbis.

This refined dataset serves as the foundation for a first data analysis, enabling the identification of relationships between financial indicators and EBITDA. A complete overview of the selected variables is provided in Table 3.2.

Company information	**Assets**
BvD sectors	Intangible fixed assets
Date of incorporation	Tangible fixed assets
Operating revenue	Current assets
EBITDA	Debtors
	Total assets
Liabilities and equity	**Other variables**
Shareholder funds (Equity)	Working capital
Non-current liabilities	
Current liabilities	
Total shareholder funds and liabilities	
Financial ratios	
Solvency ratio, Current ratio, Gearing, Liquidity ratio	

TABLE 3.2: Overview of selected variables and financial ratios

3.1.2 Data cleaning

Once the dataset was constructed, the next step involved data cleaning to ensure its quality and reliability for model training. Raw financial data often contains inconsistencies, missing values, and outliers, all of which can negatively impact the accuracy of predictive models. In this study, data cleaning included standardizing data formats, handling missing values, and removing duplicate entries.

Standardized data formats

For data standardization, only the date of incorporation required modification. This variable was standardized to represent only the year of establishment, enabling its transformation into the variable "age" by subtracting the year of incorporation from the current year.

Missing values

During the examination of the dataset, several features were found to contain missing values. In some cases, the proportion of missing data was minimal, while in others, a significantly larger percentage was missing. The following variables exhibited only a minimal number of missing entries. Given their negligible impact on the dataset, these specific data points were excluded. As a result, the total number of entries decreased from 54,707 to 54,241. This modest reduction reflects overlapping missing values across multiple variables, with the removed records typically containing three or more missing values. The variables include the following number of missing entries and percentages as part of the dataset:

- **Intangible assets:** 101 missing values (0.18%)
- **Tangible assets:** 97 missing values (0.18%)
- **Debtors:** 4 missing values (0.0073%)
- **Non-current liabilities:** 5 missing values (0.0091%)
- **Solvency ratio:** 247 missing values (0.45%)
- **Current ratio:** 27 missing values (0.049%)
- **Liquidity ratio:** 118 missing values (0.22%)

Furthermore, three other variables were identified as having significantly more missing values: Operating revenue, working capital, and gearing. The following values were observed:

- **Operating revenue:** 42715 missing values (78.08%)
- **Working capital:** 6535 missing values (11.61%)
- **Gearing ratio:** 9547 missing values (17.45%)

Since operating revenue is missing in more than half of the dataset and cannot be derived from the balance sheet according to a formula, it was decided to exclude this variable from the current dataset.

For working capital and gearing, further examination revealed that these variables can be reconstructed using the other selected features. Instead of removing these variables, they will be derived as follows:

- **Working capital:** Derived by: Working capital = Current assets−Current liabilities (Rabobank, 2025). While corporate tax is usually included in current liabilities rather than in working capital, for calculation purposes, it is assumed to be part of working capital in this case.
- **Gearing:** Derived by: Gearing $= \frac{\text{Total liabilities}}{\text{Total shareholder funds}}$ (Investopedia, 2025).

Following the application of data handling techniques, all missing values were successfully addressed, ensuring the completeness of the dataset. Additionally, duplicate entries were identified and removed to maintain data integrity. As a result, the final dataset consists of 54,241 unique entries, including sixteen features that serve as potential variables to estimate EBITDA.

3.2 Exploratory data analysis

This section presents an exploratory data analysis to examine the relationships between independent variables and EBITDA, as well as to assess the distribution of individual variables. Understanding these relationships is crucial for refining the dataset and selecting the most relevant features for model development.

To quantify the associations between the independent variables and EBITDA, both Pearson's and Spearman's rank correlation coefficients are employed. Pearson's correlation assesses the strength and direction of linear relationships under the assumption of normally distributed variables, while Spearman's correlation evaluates monotonic relationships by ranking values, making it more robust to outliers and better suited for detecting non-linear patterns (Morais et al., 2023; Rezaee et al., 2020; Tian et al., 2024). Both coefficients range from -1 to 1, where values close to 1 or -1 indicate strong positive or negative associations, and values near 0 suggest no correlation. Including both metrics provides a more comprehensive understanding of how the selected features relate to the dependent variable, EBITDA. The formula for Pearson's correlation coefficient is given by:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{3.1}$$

Where;

- r = Pearson correlation coefficient
- x_i = feature samples x $\bar{x}$ = mean of feature x
- y_i = feature samples y $\bar{y}$ = mean of feature y

And the formula for Spearman's correlation coefficient is then given by:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{3.2}$$

Where;

- ρ = Spearman's rank correlation coefficient
- d_i = difference between the two ranks of each observation i
- n = number of observations

3.2.1 Baseline scenario test including all sectors

As an initial assessment, the full dataset, encompassing all sectors, is utilized to examine the correlation between EBITDA and sixteen different features (Table 3.3). This broad analysis establishes a general first understanding of the relationships between financial indicators and EBITDA before delving into sector-specific variations.

Variable	Pearson	Spearman
Number of employees	0.509	0.536
Intangible assets	0.153	0.230
Tangible fixed assets	0.183	0.547
Current assets	0.151	0.759
Debtors	0.142	0.648
Total assets	0.137	0.806
Shareholders' funds	0.084	0.725
Non-current liabilities	0.102	0.352
Current liabilities	0.113	0.718
Total shareholders' funds	0.137	0.806
Working capital	0.050	0.471
Solvency ratio	0.138	0.169
Current ratio	0.053	0.138
Liquidity ratio	0.018	0.057
Gearing	-0.132	-0.139
Age	0.160	0.177

TABLE 3.3: Correlation across all sectors between EBITDA and all features

The correlation analysis shown Table 3.3 confirms that the number of employees remains the strongest predictor of EBITDA, considering both its Pearson correlation of 0.509 and Spearman correlation of 0.536. This finding reinforces firm size as a key driver of profitability in the SME sector, aligning with existing literature (Strategic Direction Editorial Team, 2014). Additionally, total assets, current assets, and shareholders' funds show strong Spearman correlations (0.806, 0.759, and 0.725), suggesting that firms with larger asset bases and stronger equity positions generally achieve higher EBITDA.

While tangible fixed assets, debtors, and shareholders' funds display relatively weak Pearson correlations (0.183, 0.142, and 0.084), their substantially higher Spearman values (0.547, 0.648, and 0.725) indicate the presence of non-linear relationships with EBITDA. Similarly, working capital shows a weak Pearson correlation (0.050) but a moderate Spearman correlation (0.471), implying that it has a non-linear impact on profitability.

Financial ratios such as solvency, liquidity, and gearing exhibit, contrary to literature, relatively weak correlations with EBITDA (Batrancea et al., 2018; L. Li et al., 2023). Notably, gearing (Pearson: -0.132, Spearman: -0.139) shows a negative relationship, suggesting that higher leverage is generally associated with lower earnings. Firm age also demonstrates only a modest influence (Pearson: 0.160, Spearman: 0.177), indicating that while maturity may confer some advantages, it is not a strong standalone predictor of profitability.

These findings highlight that firm size, asset-related variables, and financial structure are the primary drivers of EBITDA, whereas financial ratios contribute less directly. The consistently stronger Spearman correlations underscore the non-linear nature of these relationships, reinforcing the importance of using modeling techniques capable of handling and capturing such complexities.

3.2.2 Sector-specific correlation test

To capture sector-specific financial dynamics, correlation tests were also conducted for three of the five largest sectors by deal volume for Marktlink: Business services, Wholesale, and Construction (abbrevation of "Construction and maintenance"). These sectors were selected due to their substantial deal volume for Marktlink and the significant financial differences observed across industries, making them particularly relevant for evaluating sector-specific EBITDA estimation. This selection is supported by prior research highlighting sectoral variations in financial structures and performance, as well as the descriptive statistics in Appendix A.7 (Table A.7) (Johnsen & McMahon, 2005; Tong & Serrasqueiro, 2020). By analyzing these sectors individually, the study provides a more precise assessment of how different financial variables influence EBITDA across different industries, ensuring that sector-specific financial patterns are accurately identified and accounted for.

Variable	**Business services**		**Wholesale**		**Construction**	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
Number of Employees	0.519	0.604	0.481	0.569	0.563	0.523
Intangible Assets	0.199	0.288	0.116	0.174	0.095	0.175
Tangible Fixed Assets	0.408	0.604	0.322	0.426	0.290	0.565
Current Assets	0.138	0.761	0.438	0.756	0.588	0.750
Debtors	0.118	0.694	0.377	0.665	0.530	0.613
Total Assets	0.151	0.788	0.079	0.780	0.399	0.796
Shareholders' Funds	0.112	0.698	0.048	0.729	0.358	0.746
Non-Current Liabilities	0.117	0.440	0.117	0.166	0.200	0.365
Current Liabilities	0.105	0.712	0.308	0.669	0.472	0.674
Total Shareholders' Funds	0.151	0.788	0.079	0.780	0.399	0.796
Working Capital	0.282	0.493	0.264	0.576	0.387	0.510
Solvency Ratio	0.185	0.190	0.133	0.159	0.091	0.139
Current Ratio	0.055	0.120	0.049	0.127	0.041	0.114
Liquidity Ratio	-0.015	-0.085	0.030	0.088	0.036	0.098
Gearing	-0.220	-0.337	-0.154	-0.196	-0.113	-0.110
Age	0.071	0.064	0.159	0.167	0.291	0.212

TABLE 3.4: Pearson and Spearman correlation test between EBITDA and all features for three sectors: Business services, wholesale, and construction

The sector-specific correlation analysis, presented in Table 3.4, highlights distinct financial drivers of EBITDA across business services, wholesale, and construction. Across all sectors, the number of employees remains the strongest linear predictor of EBITDA, with the highest Pearson correlation in construction, emphasizing firm size as a key driver of profitability in that sector. Asset-related variables, such as total assets, current assets, and shareholders' funds, exhibit strong Spearman correlations across all sectors, particularly in construction, indicating the capital-intensive nature of the industry (Syed & Elwakil,

2019).

Tangible fixed assets play a more significant role in business services than in construction, suggesting that infrastructure investments are crucial even in service-based industries. Shareholder funds and non-current liabilities show the highest correlations in construction, likely due to the need for long-term financing (Bal, 2020). Meanwhile, gearing exhibits a consistently negative correlation across all sectors, suggesting that higher leverage does not necessarily translate into increased profitability.

Working capital management also proves relevant across all sectors, underscoring the importance of efficient cash flow management in driving profitability for SMEs. However, consistent with the baseline scenario, financial ratios such as solvency, liquidity, and gearing exhibit weak correlations with EBITDA, reinforcing their limited value as predictive indicators.

Firm age shows the strongest relationship with EBITDA in construction, indicating that industry experience and reputation contribute to profitability, whereas its impact is more limited in business services and wholesale.

Furthermore, it can be observed that total assets and total shareholders' funds have identical correlation values within each sector across all sectors, similarly as in Table 3.3. This outcome can be derived from the fact that these variables represent equivalent values, as each forms the concluding total of opposing sides of the balance sheet. Consequently, total assets are selected as the representative input variable for further use in this research. This results in a new total of fifteen features that can serve as input parameters.

These findings confirm that firm size, assets, and financial structure are the primary determinants of EBITDA, while financial ratios have a more limited direct effect, aligning with the results of the baseline correlation analysis. The strong Spearman correlations suggest non-linear dependencies, emphasizing the need for more advanced data handling and modeling techniques to capture these complex relationships more effectively. This is further supported by the distribution analysis in Figures A.1, A.2, and A.3, which illustrate a right-skewed distribution, indicating deviations from normality.

3.2.3 Backtesting results with 2019 scenario

To assess the stability and representativeness of the correlations identified in the baseline scenario (Table 3.4), a backtesting procedure was performed using a 2019 dataset comprising 48,737 records of Dutch and Belgian SMEs. This dataset includes the same features and underwent identical data cleaning procedures to ensure comparability. By selecting data from 2019, the analysis leverages historical financial information while explicitly excluding the economic distortions introduced by the COVID-19 pandemic in 2020 (Fissler & Hoga, 2024). This provides a valuable measure for assessing the consistency of financial relationships under normal economic conditions.

The results, presented in Table A.8, demonstrate that the correlations between financial variables and EBITDA have remained largely stable over time, with only minor deviations. This consistency suggests that the identified relationships in Table 3.4 are valid and not merely the result of short-term economic fluctuations. By validating these correlations across multiple years, the findings further support the reliability of the selected financial indicators.

3.3 Feature engineering and selection

Following the initial data analysis, this subsection focuses on refining the dataset to enhance future model performance. This involves identifying and addressing outliers that could distort predictive accuracy, as well as applying transformation techniques to change variable distributions (Osborne, 2002). Log-transformations, for example, are used to reduce skewness in heavily skewed variables, making relationships more linear and thereby increasing the predictive power of the model (R. M. West, 2022). Additionally, feature selection will be applied to test different sets of variables, ensuring that the final model captures the most meaningful financial patterns without unnecessary complexity.

3.3.1 Feature engineering

Handling outliers

Several methods exist for handling outliers, with the 1.5*IQR method and the Z-score method being among the most commonly applied approaches. The Z-score method is appropriate when data follows a normal distribution (Chikodili et al., 2021; Kannan et al., 2015). However, the variables in this study show right-skewed distributions with long tails, indicating the presence of extreme values and deviations from normality, as shown in Figures A.1, A.2, and A.3. Consequently, the Z-score method is unsuitable for this dataset.

Instead, the IQR method was considered, as it is more appropriate for non-normally distributed data. This method identifies outliers by first determining the first (Q1) and third (Q3) quartiles and then calculating the interquartile range (IQR = Q3 - Q1). The lower and upper bounds are defined as Q1 - 1.5IQR and Q3 + 1.5IQR, respectively. Any values falling outside these thresholds are classified as outliers and removed from the dataset. Although the IQR method is well-suited for handling non-normal data, it can be sensitive to extreme values (Alabrah, 2023; Barbato et al., 2011).

Applying this method resulted in the removal of more than half of the dataset, indicating a substantial number of extreme observations. However, eliminating such a large proportion of the data risks discarding meaningful financial information. Therefore, an alternative approach was sought to handle extreme values more conservatively while retaining the integrity of the dataset.

To address this, Winsorization at the 1st and 99th percentiles was applied. This technique mitigates the influence of extreme values by capping them at predefined thresholds rather than removing them entirely. It maintains the overall distribution of the data and helps reduce the distortion that extreme values can introduce into correlation structures and regression estimates. Compared to direct outlier removal, Winsorization avoids excessive data loss and is a widely adopted approach in empirical financial research (Mekelburg & Strauss, 2024; Sullivan et al., 2021). While Winsorization may reduce financial interpretability in extreme value cases, it was considered a necessary trade-off to preserve sample size and stabilize linear relationships critical for model development.

An illustrative example of this process is shown in Figure 3.2, which demonstrates how extreme values are capped at the 1st and 99th percentiles while preserving the core shape of the distribution.

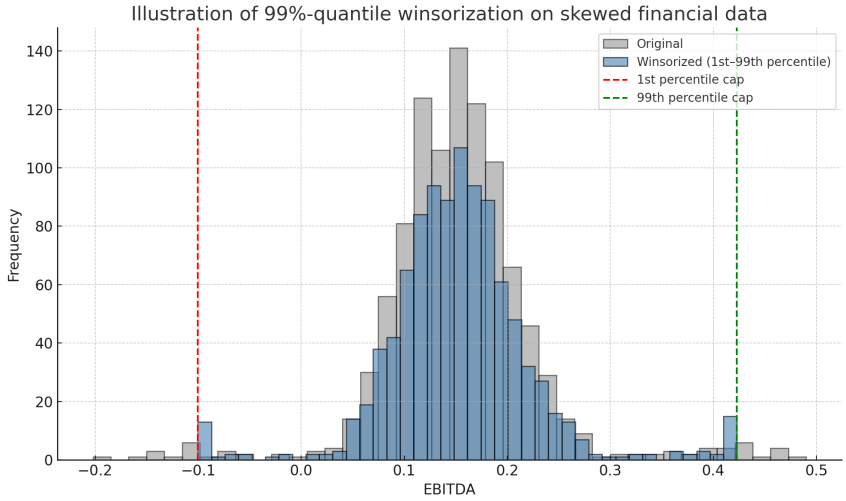


FIGURE 3.2: Illustration of 99%-quantile winsorization on financial data (EBITDA)

The Winsorized dataset preserves the original structure of the dataset while limiting the influence of extreme values. As shown in Table A.9, Winsorization improves the Pearson correlations between explanatory variables and EBITDA, indicating enhanced linear alignment.

Variable transformations

To mitigate the right-skewed distribution typical of financial data, transformation techniques are commonly employed to enhance linearity and stabilize variance. Among the most widely used approaches for addressing skewed data are log transformations and square root transformations (Manikandan, 2010; Osborne, 2002; R. M. West, 2022).

In this study, both the dependent and independent variables were transformed using log and square-root methods as they both show skewness to the right. Given the presence of zero values among several independent variables, the log transformation was implemented using the adjusted formula $\ln(x+1)$. The transformations applied are summarized in Table 3.5.

Transformation type	**Formula**
Log-transformation	$\ln(x+1)$
Square-root transformation	$\sqrt{x}$

TABLE 3.5: Overview of applied variable transformations

Applying these transformations reduced the skewness of most variables and shifted their distributions toward a more normal shape. Consequently, Pearson correlation coefficients with EBITDA increased for the majority of variables, as shown in Table A.10 and Table A.11.

The results indicate that the log transformation consistently led to greater improvements in correlation strength compared to the square-root transformation, particularly for asset-related variables such as total assets, current assets, and shareholders' funds. This suggests that log transformation more effectively aligns financial variables with EBITDA in a linear manner, especially in asset-intensive industries like construction. However,

for some variables, such as non-current liabilities and tangible fixed assets in the wholesale and construction sectors, the correlation decreased after log transformation relative to square-root transformation. This may indicate over-transformation effects, where the applied adjustment distorts rather than improves the underlying relationship.

Given these findings, the log transformation was selected instead of square-root transformations for further modeling steps. The observed improvements in linearity and correlation make it especially appropriate for regression-based models such as MLR, which assume linear relationships, and for boosting algorithms like XGBoost, which benefit from smoothed feature distributions and reduced variance (Mahesh, 2020; Uyanık & Güler, 2013).

Untransformed data versus log-transformed data

Understanding the distinction between log-transformed and raw-data models is essential for interpreting their outputs and selecting the right approach for different use cases. Each method serves a different goal and comes with trade-offs in terms of statistical behavior and practical application.

Log-transformed models operate on a proportional scale, making them particularly effective for assessing relative performance, such as how high a firm's EBITDA is compared to its peers. This makes them well suited for early-stage screening, ranking, or target prioritization. Metrics such as SMAPE, bias, and R^2 in the log domain therefore reflect percentage-based or relative deviations rather than absolute monetary errors. Nonetheless, log transformations can pose interpretability challenges, particularly when presenting results to stakeholders without a technical background. As such, their use should be balanced against the need for transparency and straightforward communication, depending on the context in which the model is applied (Osborne, 2002).

On the other hand, models trained on raw (non-transformed) financial data predict absolute EBITDA values, expressed directly in euros. These outputs are easier to interpret and more directly applicable to financial decision-making processes such as valuation. While these models may yield lower statistical performance, often due to variance and outliers, they are more transparent and actionable in practice. Despite their different scales, log-transformed model outputs can still be used to estimate values in euros through inverse transformation (e.g., exponentiation).

3.3.2 Feature selection

To optimize predictive performance of the EBITDA estimation model in the next chapter, a structured feature selection process is implemented. Since financial characteristics and relationships vary across industries, the analysis is conducted on a sector-specific basis. This sector-based approach ensures that the model captures industry-specific financial dynamics, improving its overall reliability and accuracy.

The feature selection process involves testing two different types of features, each designed to examine how data transformations affect model performance:

- **Non-transformed variables:** Representing the raw financial data without transformations applied.
- **Log-transformed variables:** Applied to reduce skewness and improve linearity in relationships with EBITDA.

For each type of feature, two selection strategies are applied to evaluate the impact of dimensionality reduction:

- **Full feature set:** Including all available variables to assess their collective predictive power.
- **Top-5 correlating features:** Selecting only the five variables that show the strongest Pearson correlation with EBITDA within each sector, thereby focusing on the most relevant predictors.

This approach results in the following four test configurations, each applied separately per sector:

# Test	Variable type	Feature set
1	Non-transformed	Full set
2	Non-transformed	Top-5 features
3	Log-transformed	Full set
4	Log-transformed	Top-5 features

TABLE 3.6: Overview of test configurations

By evaluating these combinations across different sectors, the analysis aims to identify the most effective feature set for EBITDA estimation. This structured approach allows for a comparison of non-transformed versus transformed data and full versus reduced feature sets.

Chapter 4

Model Application and Evaluation

Contents

This chapter presents the development and evaluation of models for estimating EBITDA in Dutch and Belgian SMEs. Section 4.1 introduces the supervised learning approach, comparing classification and regression techniques, and outlines the data partitioning strategy using a train/test split and k-fold cross-validation. Section 4.2 presents the classification results, assessing whether firms fall above or below the €500,000 EBITDA threshold. Section 4.3 covers the regression-based estimates, evaluated with R^2, bias, and SMAPE. Section 4.4 applies machine learning to estimate EBITDA, and SHAP to interpret feature importance in the best-performing models. Section 4.5 conducts further model evaluation, including cross-validation checks, critical error analysis, and a economic validity test using a stylized DCF framework.

4.1 Modeling approach and data partitioning

The modeling process for EBITDA estimation follows a supervised learning approach, where historical SME data is used to train models that predict future outcomes (Cunningham et al., 2024). Within supervised learning, two primary modeling techniques are considered: classification and regression (IBM, 2025). These methods serve different purposes in the world of modeling, where classification is used to sort data into categories and regression to understand relationships between dependent and independent variables resulting in a continuous output variable. In this research classification acts as a first step to gain insight into the EBITDA range a company might fall in and regression provides then precise estimates of EBITDA.

4.1.1 Supervised learning: Classification vs. regression

Supervised learning involves training a model using labeled data, where input features (independent variables) are mapped to known outcomes (dependent variable) (Hastie et al., 2009; Jensen et al., 2010). In this study, two approaches are used:

- **Classification:** The first step is to grasp a first indication of the profitability of a SME by classifying in which EBITDA range it falls. This classification model serves as a preliminary screening before applying regression techniques, which try to predict the continuous value of EBITDA (Bergen et al., 2023).
- **Regression:** Once classified, the next step involves precisely estimating EBITDA as a continuous variable. The regression models are trained and tested to predict exact values based on the feature selections as specified in Section 3.3.2 (Jensen et al., 2010).

4.1.2 Data partitioning

To effectively improve the model's performance, a two-step data partitioning strategy is applied. Firstly, the dataset is split into a training set and a test set using an 80/20 ratio. Second, k-fold cross-validation is implemented within the training set to further enhance model reliability by validating performance across multiple iterations. This approach reduces the risk of overfitting and yields a more robust and generalizable estimate of the model's predictive performance.

Train/test split

The initial split separates 80% of the data for training and 20% for testing, allowing the model to learn from historical data while being evaluated on unseen instances (Jensen et al., 2010). This approach prevents the model from being assessed on the same data it was trained on, thereby reducing overfitting and offering a realistic measure of generalization performance (Joseph, 2022; Pawluszek-Filipiak & Borkowski, 2020).

K-Fold Cross-validation

To further improve model reliability, 5-fold cross-validation is performed on the training set (Nti et al., 2021; Varoquaux & Colliot, 2023). The training data is divided into five equal subsets. In each iteration, the model is trained on four folds and validated on the remaining one. This process is repeated five times, ensuring each fold serves as the validation set once. The resulting performance metrics are averaged across all iterations,

producing a more stable and generalizable evaluation than a single train-test split. The final results are reported as mean and standard deviation across the folds.

4.1.3 Tree-based model introduction: Random Forest and XGBoost

Prior research identified that tree-based modeling approaches have a great potential to show promising results in SME EBITDA estimation (Stokkers, 2024). Therefore are two tree-based machine learning models used for the classification task (Section 4.2) as well as the regression problem (Section 4.4): Random Forest (RF) and XGBoost. Both approaches are widely recognized for their ability to handle high-dimensional data and capture complex relationships among features (Frank et al., 1998). These model characteristics make them very suitable for application to the problem under investigation in this research.

Random Forest

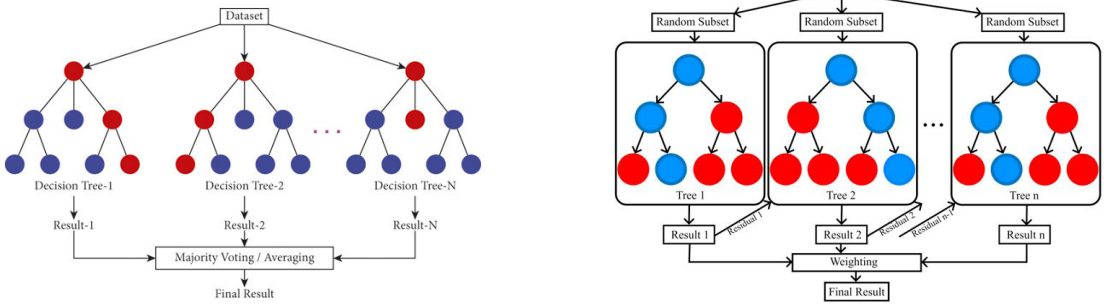
RF is an ensemble learning technique that builds upon the decision tree algorithm and can be applied to both classification and regression tasks. Rather than constructing a single decision tree, RF creates a large number of trees using different random subsets of the data and input features. Specifically, at each split in a tree, only a randomly selected subset of the available variables is considered for decision-making. This randomness introduces diversity among trees and helps reduce overfitting. Each tree is trained independently on a bootstrapped sample of the training data. For classification tasks, the final prediction is made by aggregating the individual tree votes (majority rule), while in regression, the model averages the outputs across all trees. This approach combines high predictive power with improved robustness and generalizability (Scornet, 2016).

XGBoost

XGBoost (Extreme Gradient Boosting) is a scalable tree-based ensemble algorithm built on the principle of gradient boosting. Unlike RF, which builds trees independently, XGBoost constructs trees sequentially. Each new tree is trained to correct the prediction errors made by the ensemble of previous trees. This is done by optimizing a differentiable loss function using gradient descent techniques, allowing the model to iteratively focus on the most difficult observations. XGBoost includes regularization terms in its objective function to penalize overly complex models, making it both accurate and resistant to overfitting. Therefore, one would expect predictive performance to be stronger than classic RF (Chen & Guestrin, 2016; Zhu et al., 2024).

Model comparison: Random Forest versus XGBoost

While both RF and XGBoost are tree-based ensemble methods, their learning strategies differ: RF builds many independent trees in parallel, aggregating their outputs to reduce variance, whereas XGBoost constructs trees sequentially, each correcting the previous model's residuals. This makes RF generally easier to tune and robust by design, while XGBoost, though more complex, often yields higher accuracy due to its optimization and regularization capabilities. An overview of how these models work in practice is shown in Figure 4.1.



(A) Simplified structure of a Random Forest

(B) Simplified structure of a XGBoost model

FIGURE 4.1: Simplified overview of how Random Forest and XGBoost models work in practice

4.1.4 Model assumptions

In this subsection the key assumptions underlying the modeling process are acknowledged. These assumptions are categorized into three parts: general assumptions applicable to all predictive modeling methods conducted in this study, assumptions specific to MLR, and assumptions specific to RF and XGBoost.

General assumptions

The following assumptions are relevant to all predictive models applied in this research, regardless of type:

- **Independence of observations:** Each SME in the dataset is treated as an independent observation. There is no assumed structural dependence between firms.
- **Consistent accounting logic:** It is assumed that the financial statement variables (e.g., current assets, liabilities, solvency ratios) follow comparable accounting rules and are reliably reported across all firms within each country.
- **Stationarity of relationships:** It is assumed that the relationships between financial features and EBITDA are stable across time, as partially validated through backtesting on 2019 data (Section 3.2.3).
- **Sectoral homogeneity:** Models are developed per sector (business services, wholesale, construction) under the assumption that firms within the same sector share comparable financial structures and operating characteristics.
- **Explanatory power of input features:** The models rely on the assumption that the set of available balance sheet variables provides sufficient explanatory power to predict EBITDA in the absence of full income statement data.

Model-specific assumptions: Multiple Linear Regression

MLR relies on several assumptions that ensure valid coefficient estimation and reliable inference (Osborne, 2002; Williams et al., 2013):

- **Linearity:** It is assumed that a linear relationship exists between the independent variables and the dependent variable, whether EBITDA is modeled in its original or log-transformed form.
- **Constant variance of errors:** The residuals are assumed to have constant variance across all levels of predicted EBITDA values.
- **Normality of residuals:** For purposes such as interpreting coefficient significance, it is assumed that the model residuals are approximately normally distributed.

Model-specific assumptions: Machine learning (RF and XGBoost)

Tree-based machine learning models are less reliant on classical statistical assumptions, but they do come with modeling assumptions relevant to this study (Frank et al., 1998; Scornet, 2016):

- **Representativeness of training data:** It is assumed that the training set is representative for the broader SME population. Any bias in the training data may affect generalizability.
- **Independence of features:** While not strictly required, models such as RF and XGBoost can suffer from overfitting if highly correlated features dominate the splits.
- **Stability across folds:** The models are assumed to generalize well to unseen data.

4.2 Classification based modeling

As a preliminary step in the modeling process, a classification task was conducted to determine whether firms fall above or below an EBITDA threshold of €500,000, a benchmark used by the company to distinguish relevant leads from less promising ones:

- *Class 0:* EBITDA $<$ €500,000
- *Class 1:* EBITDA $\geq$ €500,000

Model performance is assessed using a confusion matrix and accuracy derived from the matrix, both of which are derived from the model's correct and incorrect predictions (Davis & Goadrich, 2006; Rainio et al., 2024; Varoquaux & Colliot, 2023):

- **True Positives (TP):** Firms correctly predicted to have an EBITDA of **€500K or more** (i.e., predicted $\geq 500K$ and actually $\geq 500K$).
- **False Positives (FP):** Firms incorrectly predicted to have an EBITDA of **€500K or more**, while their actual EBITDA is **below €500K** (predicted $\geq 500K$, actual $< 500K$).
- **True Negatives (TN):** Firms correctly predicted to have an EBITDA of **less than €500K** (i.e., predicted $< 500K$ and actually $< 500K$).
- **False Negatives (FN):** Firms incorrectly predicted to have an EBITDA of **less than €500K**, while their actual EBITDA is **€500K or more** (predicted $< 500K$, actual $\geq 500K$).

Then, the confusion matrix and accuracy follow after defining the type of predictions:

- **Confusion matrix:** Provides a breakdown of actual vs. predicted classifications across the three EBITDA categories:

Actual / Predicted	**Class 0 (<€500K)**	**Class 1 (≥€500K)**
Class 0 (<€500K)	TN	FP
Class 1 (≥€500K)	FN	TP

TABLE 4.1: Confusion matrix for EBITDA Classification

- **Accuracy:** The proportion of all correct predictions (both positive and negative) out of the total number of predictions made:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.1}$$

In the classification matrix, the misclassified entries FN and FP can be interpreted in two distinct ways. For the purposes of this research, they are defined and categorized as follows:

- **Critical % errors:** The observations denoted as FN are the most critical. These correspond to firms that are actually highly profitable (EBITDA ≥ €500K), but are incorrectly classified as less profitable. Minimizing this type of misclassification is essential, particularly given the research focus on deal sourcing and lead generation. In model evaluation this type of error is referred to as the critical percentage.
- **Imprecision %:** In contrast, the inverse error, where firms classified as highly profitable but ultimately having lower profitability, is less critical. This type of error is referred to as *imprecision*. While such cases (FP) may lead to temporary misallocation of attention, they do not risk missing out on valuable acquisition opportunities.

4.2.1 Results of the classification

To evaluate the model's classification performance, accuracy and critical error rates were assessed across the three sectors that are under investigation in this research, using both RF and XGBoost classifiers. Models were tested following the test configurations as mentioned in Section 3.3.2 (Table 3.6). Finally, the confusion matrices are shown of the best performing tests per sector in Appendix A.15.

Sector	Test #	ML Classifier	Variable type	Feature set	Accuracy	Critical %	Imprecision %
Business services	1	RF	Normal	All	0.8597	6.90%	7.12%
Business services	2	RF	Normal	Top-5	0.8468	7.44%	7.88%
Business services	5	XGBoost	Normal	All	0.8657	6.74%	6.69%
Business services	6	XGBoost	Normal	Top-5	0.8544	7.39%	7.17%
Business services	3	RF	Log	All	0.8737	6.46%	6.17%
Business services	4	RF	Log	Top-5	0.8564	6.96%	7.40%
Business services	7	XGBoost	Log	All	0.8751	6.25%	6.24%
Business services	8	XGBoost	Log	Top-5	0.8679	6.25%	6.97%
Wholesale	1	RF	Normal	All	0.8342	8.38%	8.19%
Wholesale	2	RF	Normal	Top-5	0.8207	8.96%	8.97%
Wholesale	5	XGBoost	Normal	All	0.8335	8.32%	8.33%
Wholesale	6	XGBoost	Normal	Top-5	0.8239	8.45%	9.16%
Wholesale	3	RF	Log	All	0.8416	7.12%	8.72%
Wholesale	4	RF	Log	Top-5	0.8212	8.14%	9.74%
Wholesale	7	XGBoost	Log	All	0.8353	7.04%	9.43%
Wholesale	8	XGBoost	Log	Top-5	0.8318	6.54%	10.28%
Construction	1	RF	Normal	All	0.8795	6.75%	5.14%
Construction	2	RF	Normal	Top-5	0.8719	7.68%	5.13%
Construction	5	XGBoost	Normal	All	0.8824	5.82%	5.94%
Construction	6	XGBoost	Normal	Top-5	0.8718	6.88%	5.94%
Construction	3	RF	Log	All	0.8678	7.40%	5.66%
Construction	4	RF	Log	Top-5	0.8489	8.56%	6.55%
Construction	7	XGBoost	Log	All	0.8653	6.67%	6.80%
Construction	8	XGBoost	Log	Top-5	0.8617	7.39%	6.44%

TABLE 4.2: Accuracy, critical percentage, and imprecision results across all sectors by ML classifier, variable type, and feature set

The classification results shown in Table 4.2 indicate that, overall, the models are capable of accurately classifying SMEs into either *class 0* or *class 1* based on their results on accuracy. More specifically, XGBoost outperformed RF in all three sectors considering all performance metrics. In all cases, using the full feature set yielded better results than restricting the model to the top-5 variables, suggesting that a broader range of financial indicators contributes positively to accuracy. Notably, the business services and construction sectors achieved higher accuracy scores and lower critical percentages compared to wholesale, indicating more consistent model performance in those sectors. Overall, the combination of XGBoost, log-transformed variables, and the full feature set delivered the most reliable results in this research on EBITDA classification. The results of the best performing tests are further visualized through confusion matrices shown in Appendix A.15 (Figures A.4, A.5, and A.6).

4.3 Multiple Linear Regression as prediction method for EBITDA

This section presents the results of the MLR approach for estimating EBITDA. The statistical MLR model serves as baseline estimator, providing interpretable benchmarks for performance. MLR is one of the most widely used techniques in both statistics and machine learning for modeling the linear relationship between a continuous dependent variable, EBITDA in this research, and multiple independent variables. Within a supervised learning context, MLR attempts to learn the best-fitting linear combination of features that minimizes the residual error between predicted and actual values (Alabrah, 2023;

Uyanık & Güler, 2013). The formula in this study for a MLR model is defined as follows:

$$\hat{y}_{\text{EBITDA}} = \beta_0 + \beta_n \cdot x_n + \varepsilon \tag{4.2}$$

Where:

- $\hat{y}_{\text{EBITDA}}$ is the predicted value of EBITDA (dependent variable),
- β_0 is the intercept,
- $\beta_1, \beta_2, \ldots, \beta_n$ are the model coefficients,
- x_n are the independent variables n
- ε is the error term.

Model performance is then evaluated using the critical error and imprecision percentages, next to other key metrics that provide insights into accuracy and error characteristics: the coefficient of determination (R^2), bias, and SMAPE (Botchkarev, 2018; Tatachar, 2021).

- **Coefficient of determination (R^2):** This metric indicates the proportion of variance in the dependent variable that is explained by the independent variables. An R^2 value of 1 denotes perfect prediction, while a value of 0 implies that the model explains none of the variability. Higher R^2 scores suggest that the model captures more of the underlying data structure (Chicco et al., 2021):

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{4.3}$$

- **Adjusted coefficient of determination (Adjusted R^2):** Adjusted R^2 refines the standard R^2 metric by correcting for the number of predictors relative to the number of observations. It penalizes model complexity, providing a more reliable measure of goodness-of-fit when multiple features are included. Unlike R^2, which can only increase with additional predictors, Adjusted R^2 can decrease if new variables fail to add explanatory power. Values typically range from just below 0 to 1, with higher values indicating better fit (Leach & Henson, 2007):

$$\text{Adjusted } R^2 = 1 - \left(1 - R^2\right) \times \frac{n-1}{n-p-1} \tag{4.4}$$

  Where n is the number of observations and p is the number of predictors.

- **Bias:** Bias measures the average difference between predicted and actual values, capturing whether a model tends to overestimate or underestimate the target variable. A positive bias indicates consistent overprediction, while a negative bias suggests underprediction, and a value of zero suggests near perfect predictions (Brighton & Gigerenzer, 2015):

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \tag{4.5}$$

- **Symmetric Mean Absolute Percentage Error (SMAPE):** SMAPE expresses the prediction error as a percentage of the average between actual and predicted values, making it scale-independent and more interpretable across datasets with different magnitudes. (Chicco et al., 2021):

$$\text{SMAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|\hat{y}_i - y_i|}{(|y_i| + |\hat{y}_i|)/2} \tag{4.6}$$

- **Scatter plot (Predicted vs. Actual):** In addition to numerical metrics, a visual representation is provided by plotting predicted EBITDA values against their actual counterparts. Ideally, accurate predictions will cluster closely around the diagonal, indicating near-perfect relationship between observed and predicted outcomes.

4.3.1 Performance of multiple linear regression

The following subsection presents the results of the MLR model applied to estimate EBITDA across the three sectors. Each test configuration evaluates the impact of feature transformation and feature selection as defined in Section 3.3.2 (Table 3.6).

Sector	Test #	Classifier	Variable type	Feature set	R^2	Adj. R^2	Bias	SMAPE	Accuracy	Critical %	Imprecision %
Business services	1	MLR	Normal	All	0.56	0.56	1220.12	71.91%	0.7384	4.75%	21.41%
Business services	2	MLR	Normal	Top-5	0.51	0.51	-7503.44	72.97%	0.7627	4.53%	19.20%
Business services	3	MLR	Log	All	0.65	0.65	-0.014	56.51%	0.8485	8.83%	6.32%
Business services	4	MLR	Log	Top-5	0.60	0.60	-0.0045	59.13%	0.8478	8.26%	6.96%
Wholesale	1	MLR	Normal	All	0.54	0.53	1911.91	64.83%	0.7587	2.56%	21.57%
Wholesale	2	MLR	Normal	Top-5	0.44	0.44	-1973.44	70.25%	0.6697	0.32%	32.71%
Wholesale	3	MLR	Log	All	0.65	0.64	-0.0022	53.24%	0.8322	9.08%	7.70%
Wholesale	4	MLR	Log	Top-5	0.63	0.63	0.0028	54.32%	0.8321	8.07%	8.72%
Construction	1	MLR	Normal	All	0.58	0.57	-24880.99	60.51%	0.8575	7.43%	6.82%
Construction	2	MLR	Normal	Top-5	0.50	0.50	-26877.45	67.06%	0.8383	9.60%	6.57%
Construction	3	MLR	Log	All	0.66	0.65	0.017	51.22%	0.8535	9.50%	5.15%
Construction	4	MLR	Log	Top-5	0.62	0.62	0.0035	54.34%	0.8448	10.30%	5.22%

TABLE 4.3: MLR performance across all sectors with error diagnostics

As shown in Table 4.3, the application of log-transformed variables consistently enhanced the performance of the MLR models across all sectors. This improvement is evidenced by increases in R^2 ranging from 0.07 to 0.12, and absolute reductions in SMAPE of 9.29% to 15.50%. In all cases, models utilizing the full feature set outperformed those restricted to the top five variables, suggesting that a broader inclusion of financial indicators contributes positively to predictive accuracy, regardless of transformation method. The construction sector yielded the highest R^2 value at 0.66, indicating relatively strong explanatory power for MLR within that domain.

These findings are supported visually in Appendix A.16 (Figures A.7, A.8, and A.9), which present scatter plots for the tests with the highest R^2 and accuracy per sector. Additionally, the corresponding regression coefficients and formula examples for both the full and top-five feature models are included in Appendix A.17 and A.18, respectively.

To further assess predictive quality, bias was calculated to capture over- or underestimation. In the normal space, bias values generally remained small in two of the three sectors, 1,220 in business services and 1,911 in wholesale. This indicates that MLR can produce reasonably centered predictions even without transformation. The construction sector stands out as an exception, with a larger bias of over 25,000, suggesting a larger

deviation between predicted and actual EBITDA in that case. In contrast, models trained on log-transformed variables consistently yielded bias values close to zero across all sectors, reflecting more symmetrical error distributions. However, it is important to recognize that low bias may not necessarily imply high accuracy, as these low biases can be caused by offsetting over- and underpredictions. This is reflected in the still high SMAPE values, with even the best performing log-transformed model (Construction) exhibiting a SMAPE of 51.22%, indicating that substantial relative deviations remain present.

The error-type diagnostics reveal further distinctions between transformation strategies. In both the business services and wholesale sectors, log-transformed models achieved a more balanced trade-off between critical and imprecision error rates. This is especially important from a practical perspective: while normal-space models include lower critical error rates, reducing the risk of overlooking high-potential targets, they suffer from high imprecision rate, flagging many low-value firms as promising. This inefficiency is reduced in the log-transformed models, which maintain acceptable critical rates while significantly lowering imprecision, thereby improving lead quality. In contrast, the construction sector shows relatively stable critical and imprecision percentages across all configurations, suggesting that prediction performance in this sector is less sensitive to transformation choice or feature selection, likely due to more homogeneous distribution of financial profiles among firms.

In summary, while MLR demonstrates reasonably strong accuracy and low critical error rates, its effectiveness is limited by high SMAPE values, indicating substantial absolute deviation between predictions and actual values. This underlines a core limitation of linear models: their inability to capture complex, non-linear relationships. The consistently high relative errors across sectors, despite acceptable R^2 values, highlight the need for more flexible modeling approaches. This motivates the transition to advanced machine learning models, explored in Section 4.4, with the goal of capturing deeper patterns within financial data.

4.3.2 Sensitivity analysis on input parameters for MLR

This subsection explores additional model configurations beyond those discussed earlier, with the goal of enhancing the predictive performance of MLR. All configurations in this section are applied to normal (i.e., untransformed) data. Two sensitivity analyses are conducted: one using a simple heuristic that adds features based on improvements in adjusted R^2, and another focusing on multicollinearity among features, assessed using the Variance Inflation Factors (VIF) metric.

Simple heuristic based on adjusted R^2

To further improve the performance of the MLR models across sectors, a simple heuristic approach based on the evolution of the adjusted R^2 value was tested. The principle behind this heuristic is that variables are only added to the model if their inclusion results in an increase in the adjusted R^2, thus ensuring that each added variable meaningfully contributes to explaining the variance in EBITDA (Karch, 2020; Leach & Henson, 2007).

The process was as follows:

1. Variables were ranked per sector based on their Pearson correlation with EBITDA, starting from the highest correlation.

2. Variables were added one by one to the model in order of correlation strength.

3. After each addition, the adjusted R^2 was recalculated.

4. A variable was retained if the adjusted R^2 increased, otherwise, it was excluded from the model.

This method helps prevent overfitting, as adjusted R^2 penalizes the addition of non-informative predictors (Leach & Henson, 2007). Figure 4.2 illustrates how the adjusted R^2 evolved as more features were considered per sector.

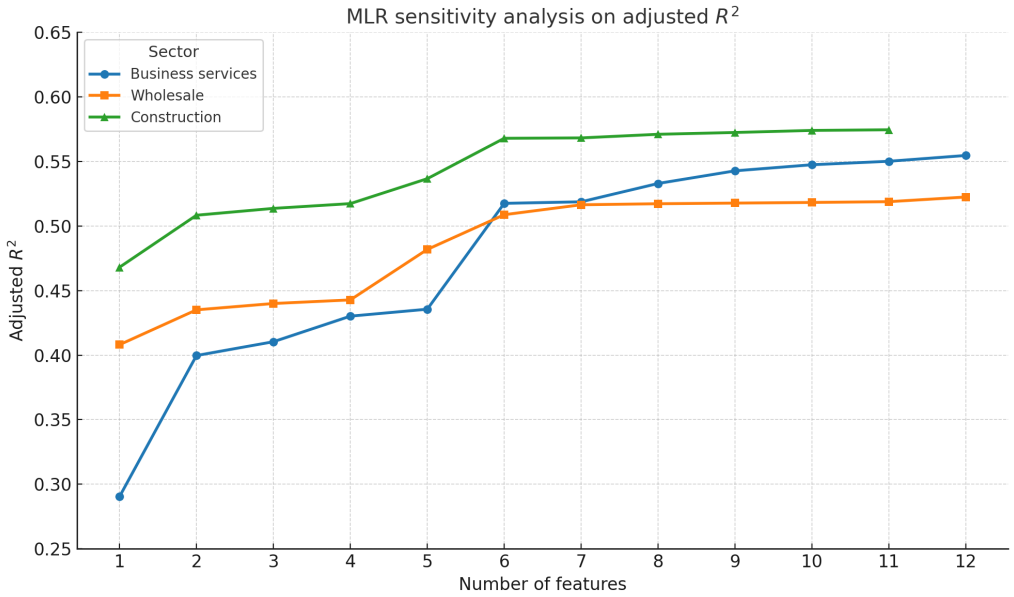


FIGURE 4.2: Evolution of adjusted R^2 with increasing number of features across sectors

The features that were ultimately excluded by this heuristic per sector were:

- **Business services:** Debtors, age, gearing
- **Wholesale:** Total assets, current liabilities, age
- **Construction:** Debtors, age, current liabilities, non-current liabilities

The final models resulting from this feature selection process delivered the following performance metrics:

Sector	Variable type	R^2	Adj. R^2	Bias	SMAPE	Accuracy	Critical error %	Imprecision %
Business services	Normal	0.56	0.55	-717.35	71.01%	0.7330	4.58%	20.12%
Wholesale	Normal	0.53	0.52	2742.04	65.66%	0.7449	2.75%	22.86%
Construction	Normal	0.58	0.57	-25,801.51	61.54%	0.8532	7.74%	6.94%

TABLE 4.4: Performance metrics per sector after simple heuristic application

When comparing the initial results in Table 4.3 with the sensitivity analysis results in Table 4.4, it becomes clear that applying the simple heuristic slightly improved overall performance of MLR on normal data. For business services and wholesale, R^2 and adjusted R^2 remained largely stable, but a small reduction in SMAPE and imprecision

errors was achieved, suggesting slightly better predictive accuracy and fewer false positives. In construction, performance remained unchanged, with similar R^2 values but a small improvement in SMAPE and imprecision. These findings suggest that applying a basic feature selection heuristic helped to improve MLR performance on normal data just marginally.

Multicollinearity assessment by using Variance Inflation Factors

One key consideration in refining MLR models is multicollinearity among independent variables, which can negatively affect the model's stability and predictive accuracy (Dertli et al., 2024). To assess the degree of multicollinearity, the VIF is employed. VIF quantifies the extent to which a regression coefficient's variance is increased due to multicollinearity with other predictors. The formula for VIF is as follows (Akhtar et al., 2024; Dawoud & Eledum, 2025):

$$\text{VIF} = \frac{1}{1 - R^2} \tag{4.7}$$

Here, R^2 represents the coefficient of determination from regressing the independent variable on all other predictors. Higher R^2 values indicate stronger linear relationships between predictors and thus greater multicollinearity.

VIF values are commonly interpreted as follows (Akhtar et al., 2024):

- **VIF $= 1$:** The variable is not correlated with other predictors, no multicollinearity is present.
- **$1 <$ VIF < 5:** Moderate correlation exists, multicollinearity is present but typically not problematic.
- **$5 \leq$ VIF < 10:** High correlation is present, multicollinearity may affect the model and should be examined further.
- **VIF ≥ 10:** Severe multicollinearity is likely, corrective action is recommended to improve model reliability.

As the VIF increases, the risk of multicollinearity rises, potentially leading to increased standard errors and unstable coefficient estimates. If high VIF values are identified, strategies such as variable elimination, or feature transformation may be required to mitigate multicollinearity (Akhtar et al., 2024; Ortiz et al., 2023). Table 4.5 presents the VIF for the selected features in the dataset per sector:

Feature	Business services	Wholesale	Construction
	VIF	VIF	VIF
Total assets	34.021	33.233	46.600
Current ratio	31.120	17.001	33.126
Liquidity ratio	25.560	12.248	27.949
Current liabilities	14.780	19.524	17.753
Current assets	14.527	22.993	21.019
Shareholders' funds	12.578	13.950	15.254
Solvency ratio	6.161	7.737	7.467
Debtors	4.769	5.187	5.347
Non-current liabilities	4.565	2.734	5.656
Working capital	3.638	4.352	4.041
Tangible fixed assets	2.716	2.489	3.853
Number of employees	2.705	2.767	3.776
Age	2.271	3.339	3.235
Intangible assets	1.650	1.287	1.186
Gearing	1.107	1.349	1.401

TABLE 4.5: Variance Inflation Factors (VIF) per feature across all sectors

Table 4.5 demonstrates that total assets shows the highest VIF, indicating severe multicollinearity. This is empirically explainable given its role in several financial ratios. It forms the denominator in solvency (shareholders' funds / total assets), gearing, and asset-related ratios, and is also directly related to working capital (current assets – current liabilities). Similarly, the current ratio and liquidity ratio both have very high VIFs. These are directly derived from current assets and current liabilities, both of which also show high VIFs. This is expected, as solvency can be calculated as shareholders' funds divided by total assets. These results are underlined by the results of the correlation heatmaps shown in Appendix A.19 (Figures A.10, A.11, and A.3). It can be concluded that the inclusion of both base variables and their corresponding ratios introduces redundancy, making multicollinearity an expected outcome. This pattern is structurally embedded in accounting relationships and, while not necessarily problematic in predictive modeling, it should be addressed with care to avoid instability in coefficient estimation.

Other features such as debtors, non-current liabilities, and working capital are within acceptable ranges, indicating manageable multicollinearity. These variables are likely to provide independent explanatory value. Moreover, tangible fixed assets, number of employees, age, intangible assets, and gearing all have VIFs below 3, suggesting minimal multicollinearity and strong justification for inclusion in the model from a statistical standpoint.

To account for the results from Table 4.5 additional test configurations are configured on untransformed data in Table 4.6:

# Test	Test type	Feature set
1	MLR	Features with VIF < 5
2	MLR	Features with VIF > 5
3	MLR	Features excluding ratio's

TABLE 4.6: Test configurations for sensitivity analysis on input parameters for MLR

The results on these new configurations are presented in Table 4.7:

Sector	# Test	Classifier	Variable type	Feature set	R^2	Adj. R^2	Bias	SMAPE	Accuracy	Critical %	Imprecision %
Business services	1	MLR	Normal	VIF < 5	0.52	0.51	-2393.11	76.39%	0.6882	4.80%	26.38%
Business services	2	MLR	Normal	VIF > 5	0.38	0.38	13885.88	80.65%	0.5863	1.67%	39.70%
Business services	3	MLR	Normal	No ratio's	0.55	0.54	-979.82	71.34%	0.7794	4.48%	17.58%
Wholesale	1	MLR	Normal	VIF < 5	0.48	0.48	19268.44	71.35%	0.6703	2.30%	30.67%
Wholesale	2	MLR	Normal	VIF > 5	0.46	0.45	-5151.73	69.31%	0.6761	1.98%	30.41%
Wholesale	3	MLR	Normal	No ratio's	0.53	0.53	2916.93	64.71%	0.7868	2.69%	18.63%
Construction	1	MLR	Normal	VIF < 5	0.52	0.52	-28868.41	68.82%	0.8290	8.80%	8.30%
Construction	2	MLR	Normal	VIF > 5	0.51	0.50	-27247.32	66.51%	0.8383	9.29%	6.88%
Construction	3	MLR	Normal	No ratio's	0.57	0.57	-23892.33	61.78%	0.8457	8.49%	6.94%

TABLE 4.7: Sensitivity analysis: Model performance per sector across new test configurations

When comparing the sensitivity analysis results in Table 4.7 to the model performances reported in Tables 4.3 several conclusions can be drawn. Firstly, the explanatory power, as measured by R^2, did not improve under the new configurations, suggesting that addressing multicollinearity through feature exclusion based on VIF did not enhance model fit. Secondly, a reduction in bias was observed, SMAPE values remained higher, indicating lower predictive accuracy. Thirdly, critical percentage thresholds in the business services and wholesale sectors remained relatively low, which is positive. However, they were not well-balanced with imprecision. These findings confirm that despite the presence of multicollinearity, full-featured log-transformed models consistently yield stronger predictive accuracy. Furthermore, it can be concluded that the results are likely caused by the fact that removing collinear variables also eliminates financially meaningful information, as many accounting features are interrelated and collectively contribute to predictive power (IBM, 2025).

4.3.3 Summary of best-performing MLR configurations per sector

Before transitioning to the evaluation of machine learning models, this section summarizes the best-performing configurations for MLR across sectors. A clear distinction is made between models trained on untransformed (normal) data and those on log-transformed data. While log-transformed models consistently deliver the strongest overall predictive performance, the sensitivity analyses show that on normal data, excluding ratio-based variables leads to better outcomes for business services and wholesale. The resulting best-performing configurations per sector are summarized below in Figure 4.3 and in Appendix A.20 (Table A.14):

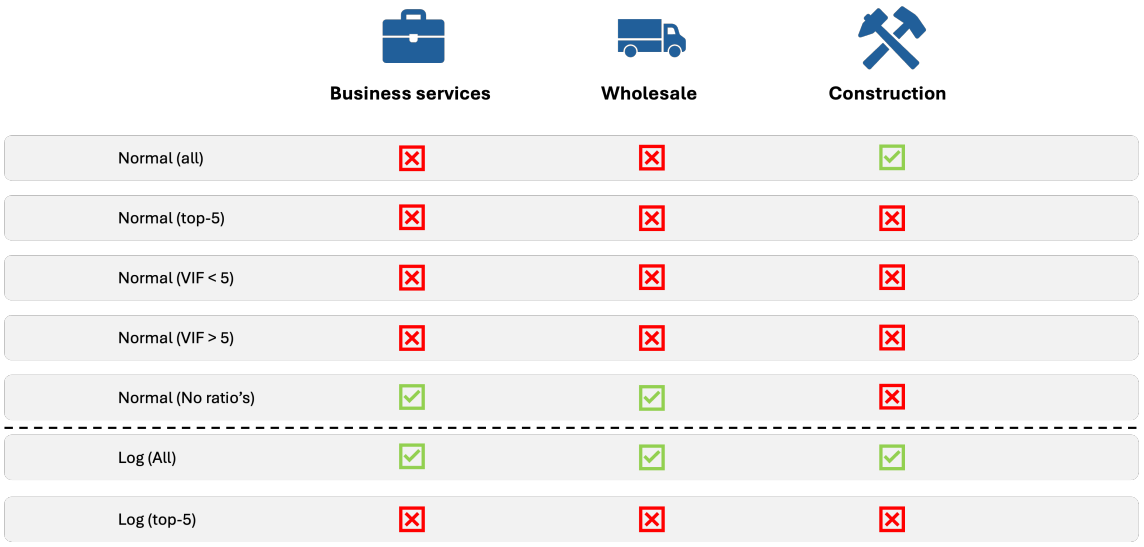


FIGURE 4.3: Overview of MLR configurations and the best performing configurations indicated by a check mark

4.4 Performance of machine learning models on predicting EBITDA

This subsection presents the performance of tree-based machine learning models, RF and XGBoost, for estimating EBITDA across the three sectors. These models are particularly effective in estimating target values based on known input-output relationships. This aligns with the structure of the dataset used, financial features linked to observed EBITDA values. The tree-based models used, RF and XGBoost, are the same as those applied in the classification tests (Ampomah et al., 2020; Mahesh, 2020).

All models are trained using normal and log-transformed input variables and the full feature set, based on an updated configuration. This revised setup differs from the test configuration introduced in Section 3.3.2 (Table 3.6), which also included the top five variables. As shown in Sections 4.2.1 and 4.3.1, including the top five features consistently resulted in reduced predictive performance. Therefore, the decision was made to include only all available variables in the next test setups to maximize the models' accuracy.

The performance of the machine learning models is evaluated using the same metrics as applied in the regression analysis (Section 4.3): R^2, bias, SMAPE, and classification-based measures such as accuracy, critical and imprecision error percentages. However, Adjusted R^2 is not reported for machine learning models. Unlike regression models, tree-based algorithms such as RF and XGBoost do not rely on a fixed number of predictors or assume linear relationships (Scornet, 2016; Zhu et al., 2024). Their internal structure allows dynamic feature selection and weighting during model training, making Adjusted R^2 an inappropriate metric for evaluation. In this context, R^2 is used purely as an overall indicator of predictive power, independent of assumptions regarding degrees of freedom or explicit feature inclusion.

Sector	Test #	ML Classifier	Variable type	Feature Set	R^2	Bias	SMAPE	Accuracy	Critical %	Imprecision %
Business services	1	RF	Normal	All	0.63	8569.28	53.25%	0.8447	4.80%	10.73%
Business services	3	XGBoost	Normal	All	0.63	-6870.39	54.97%	0.8587	5.88%	10.25%
Business services	2	RF	Log	All	0.70	-0.029	4.56%	0.8567	8.26%	6.17%
Business services	4	XGBoost	Log	All	0.71	-0.010	4.50%	0.8593	7.90%	6.17%
Wholesale	1	RF	Normal	All	0.56	19658.06	52.53%	0.8398	5.31%	11.91%
Wholesale	3	XGBoost	Normal	All	0.57	-1779.30	51.87%	0.8335	5.38%	11.27%
Wholesale	2	RF	Log	All	0.66	-0.023	4.44%	0.8387	7.85%	8.28%
Wholesale	4	XGBoost	Log	All	0.66	0.0030	4.35%	0.8378	7.92%	8.50%
Construction	1	RF	Normal	All	0.64	-548.05	51.00%	0.8540	6.32%	8.30%
Construction	3	XGBoost	Normal	All	0.65	-10726.49	50.28%	0.8561	7.06%	7.13%
Construction	2	RF	Log	All	0.65	-0.0069	4.60%	0.8524	9.64%	5.22%
Construction	4	XGBoost	Log	All	0.66	0.0171	4.50%	0.8507	8.63%	5.87%

TABLE 4.8: Performance of RF and XGBoost across all sectors

The results shown in Table 4.8 demonstrate that tree-based machine learning models, particularly XGBoost, achieve high accuracy in estimating EBITDA across all three sectors. Configurations using log-transformed variables consistently yield better predictive performance than the normal variable set, reflected in higher R^2 values and lower error metrics. The best results are obtained using XGBoost with log-transformed features, achieving R^2 scores of 0.71 in the business services sector and 0.66 in both wholesale and construction. These outcomes are positive, since they indicate a strong proportion of variance explained by the models.

Moreover, the bias results highlight the improved performance of the machine learning models. While models using untransformed variables show substantial positive or negative bias, indicating over- or underestimation, log-transformed configurations consistently produce values close to zero. The lowest observed bias is in the wholesale sector, with a value of 0.0030, suggesting that the predicted EBITDA values are more symmetrically distributed around the actual outcomes. However, these values should be interpreted with caution, as positive and negative deviations may offset each other, potentially understating the true magnitude of prediction errors.

In addition to bias, the SMAPE provides a valuable measure of relative accuracy. In the top-performing log-transformed machine learning models, SMAPE values range from 4.35% to 4.60%, suggesting that predictions deviate only slightly from actual values on average in relative (log-transformed) terms. In contrast, models trained and evaluated on untransformed (normal) data include higher SMAPE values, ranging from 50.28% to 54.97%, which reflect much larger deviations when interpreted in absolute euro terms. While these SMAPE values appear different, it is important to recognize that SMAPE in log space captures proportional differences on a compressed scale, and should not be directly compared to SMAPE in the normal space. From a practical standpoint, these differences imply that log-based models are especially suitable when the goal is to rank or compare firms based on relative performance, such as in early-stage deal sourcing or lead prioritization. Their strong performance in SMAPE and R^2 suggests high consistency in identifying which firms are likely to perform better. On the other hand, normal-space models, though showing higher error percentages, remain more appropriate for scenarios requiring precise monetary estimation, such as forecasting expected deal size, pricing, or cash flow modeling. This distinction helps guide the choice of modeling approach based on whether relative ranking or absolute value prediction is the primary objective.

Interestingly, while log-transformed models consistently outperform normal configurations in terms of overall predictive accuracy, reflected in R^2 and SMAPE values, another picture emerges when examining critical and imprecision types of error. Specifically, mod-

els using normal variables tend to achieve lower critical error percentages across all sectors. This outcome is favorable from a lead-generation perspective, as it reduces the likelihood of overlooking highly profitable firms. For example, in the wholesale sector, XGBoost with normal variables yields a critical error of just 5.38%, compared to 7.92% for its log-transformed counterpart. However, this benefit comes at a cost. Normal variable configurations consistently exhibit substantially higher imprecision rates, often exceeding 10%, meaning a greater number of low-value firms are mistakenly classified as promising. In contrast, log-transformed models offer a more balanced trade-off between these two error types. While they slightly increase the risk of missing a profitable firm, they substantially reduce false positives, as seen in the business services sector, where imprecision drops from 10.73% (normal) to 6.17% (log) under XGBoost. This more even distribution of errors may be preferable in contexts where both types of misclassification carry equal implications.

As a final validation, the scatter plots of the best performing models (highlighted in green) presented in Figures A.13, A.14, and A.15 display predicted values that are closely clustered around the diagonal line, indicating strong alignment between predicted and actual EBITDA outcomes. This visual evidence further supports the conclusion that XGBoost combined with log-transformed variables offers the most effective configuration for EBITDA estimation compared to RF as shown in Figure 4.4 as well.

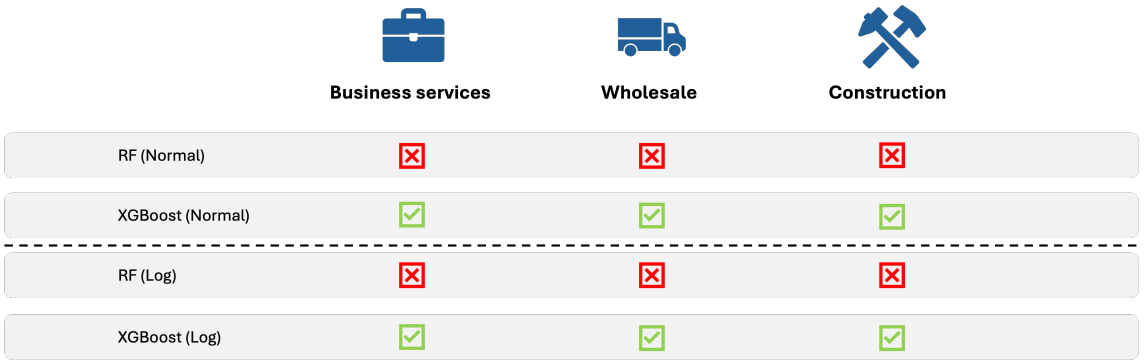


FIGURE 4.4: Machine learning configurations

4.4.1 XGBoost feature explanation using SHAP

To complement the performance evaluation of the XGBoost models in the previous section, it is essential to understand which financial variables most significantly influence the predicted EBITDA values. Identifying these key features provides insight into the primary drivers of SME profitability and supports more informed decision-making in practice for the future. To assess feature importance, this section applies SHAP values on the performances of XGBoost on log data within each sector.

SHAP offers a framework for interpreting the impact of each input variable on the model's output (Z. Li, 2022). Originated in cooperative game theory, SHAP builds on the Shapley value method by attributing a prediction to individual features based on their marginal contribution across different combinations (Lundberg & Lee, 2017; Wang et al., 2022). This allows for a consistent and theoretically grounded interpretation of feature importance in models such as XGBoost. The formula calculating Shapley values for a player i in a cooperative game as it is originally done, is shown in equation 4.6.

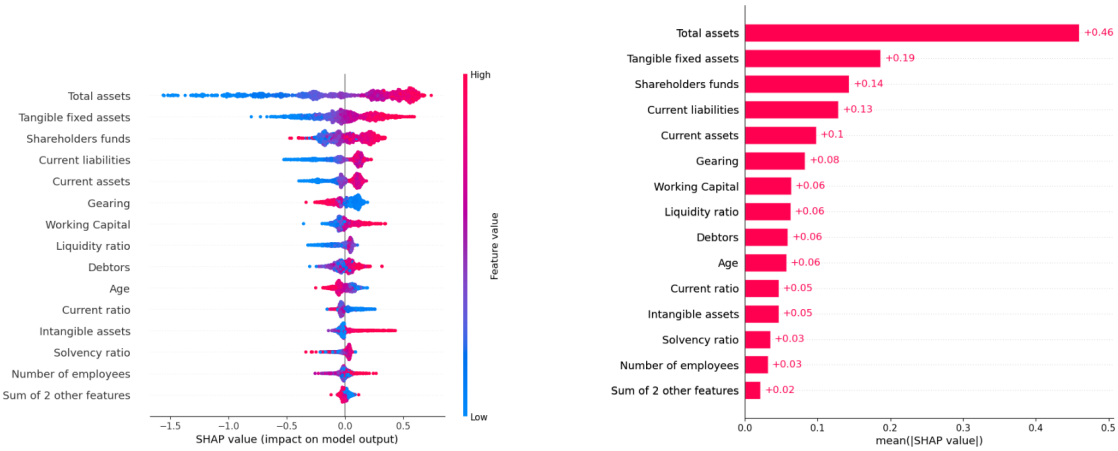
$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \tag{4.8}$$

Where:

- $\phi_i(v)$ = the Shapley value for player i
- N = the set of all players N in the game
- S = a coalition of players that does not include player i
- $v(S)$ = the value of coalition S
- $|S|$ = the number of elements in set S

Table 4.8 showed that XGBoost applied to log-transformed data achieved the highest model explainability, as reflected by the highest R^2 values across sectors. In addition, this configuration produced minimal prediction bias and strong performance on other evaluation metrics such as SMAPE and classification accuracy. Based on these results, SHAP analysis was applied to the log-transformed XGBoost models. For visualization, SHAP values are presented using both beeswarm plots and bar charts. These plots indicate the relative importance of each feature, where higher mean SHAP values reflect greater influence on the model's EBITDA predictions (Hamad et al., 2025). The results for each sector are shown in Figures 4.5, 4.6, and 4.7.

Business services

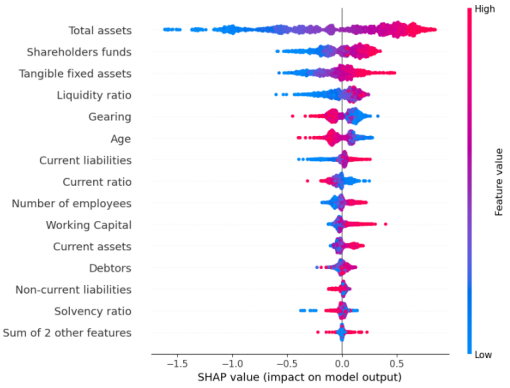


(A) SHAP beeswarm plot business services

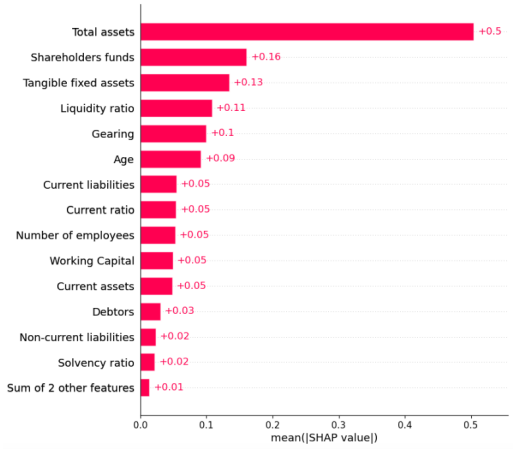
(B) SHAP bar chart business services

FIGURE 4.5: Feature explanation in XGBoost used for EBITDA estimation in business services sector

Wholesale



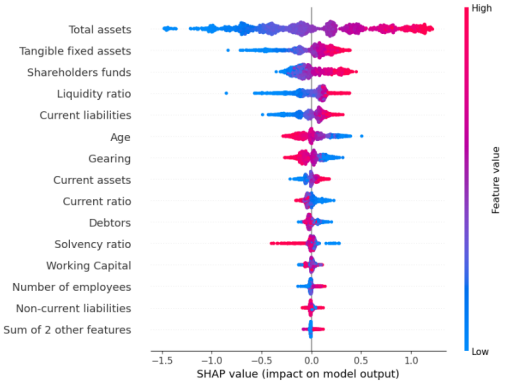
(A) SHAP beeswarm plot wholesale



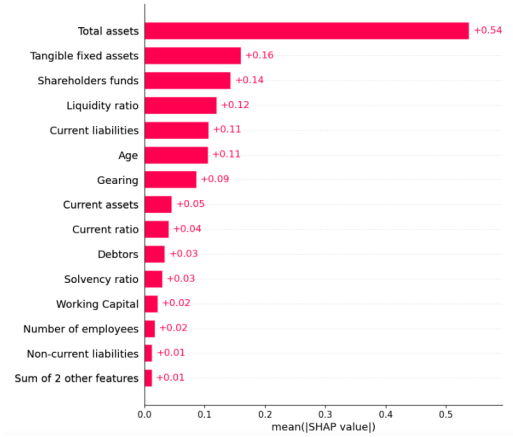
(B) SHAP bar chart wholesale

FIGURE 4.6: Feature explanation in XGBoost used for EBITDA estimation in wholesale sector

Construction



(A) SHAP beeswarm plot construction



(B) SHAP bar chart construction

FIGURE 4.7: Feature explanation in XGBoost used for EBITDA estimation in construction sector

The SHAP bar charts reveal that total assets consistently hold the highest mean SHAP value across all three sectors, ranging from 0.46 to 0.54. While not a direct measure of firm size, total assets reflect operational scale and resource availability, which influence both a firm's liquidity and leverage position (Haron, 2015). As total assets form a major component of liquidity-related metrics, such as working capital and current ratios, and simultaneously act as the denominator in key leverage ratios (e.g., solvency and debt-to-assets), their dominant SHAP values suggest that liquidity and leverage are the most influential financial drivers of SME EBITDA (Afrifa & Padachi, 2016; Haron, 2015). In practice, strong liquidity can reduce reliance on external financing, improving a firm's solvency and

capital structure, thereby enhancing earnings potential. The consistent importance of tangible fixed assets, shareholders' funds, and current liabilities further reinforces the relevance of these dimensions, as they represent long-term investment, internal financing capacity, and short-term obligations, respectively (Tong & Serrasqueiro, 2020). These findings align closely with established profitability factors outlined in Chapter 2, the literature review (Table A.2).

This interpretation is also conceptually supported by the DuPont analysis (shown in Figure A.16), which decomposes firm profitability (ROE) into three key drivers: operational efficiency (net profit margin), asset efficiency (asset turnover), and financial leverage (equity multiplier). While this study does not directly estimate ROE, the SHAP-identified importance of assets, liabilities, and equity aligns with the DuPont structure, highlighting how a firm's profitability emerges from its ability to efficiently manage operations, utilize assets, and leverage financial structure (Saus-Sala et al., 2020; Shabani et al., 2021). Integrating this view further elaborates on the explanatory power of the model and connects machine learning output to classic financial theory.

Although the top predictors are broadly consistent, sector-specific patterns provide more specific economic context (Johnsen & McMahon, 2005). In the wholesale sector, the liquidity ratio and gearing are relatively more prominent, reflecting the sector's reliance on working capital and external credit to support high inventory turnover and sales volume (Chauhan & Rameshbhai, 2024). The construction sector places greater emphasis on current liabilities and liquidity, likely due to cash flow volatility from project-based billing cycles and deferred payments (Bal, 2020; Tong & Serrasqueiro, 2020). In contrast, the business services sector, being less asset-intensive, shows relatively more importance for current assets, highlighting the role of liquid resources in sustaining operations (Syed & Elwakil, 2019). These differences suggest that while liquidity and leverage are fundamental across all sectors, their composition and economic interpretation vary by industry, shaped by the underlying operating model of each sector.

Feature	Business services			Wholesale			Construction		
	SHAP	Corr.	Match	SHAP	Corr.	Match	SHAP	Corr.	Match
Total assets	1	1	✓	1	1	✓	1	1	✓
Tangible fixed assets	2	8	✗	3	8	✗	2	3	≈
Shareholders funds	3	3	✓	2	2	✓	3	6	✗
Current liabilities	4	4	✓	7	6	≈	5	4	≈
Current assets	5	2	✗	11	3	✗	8	2	✗
Gearing	6	15	✗	5	14	✗	7	15	✗
Working capital	7	6	≈	10	7	✗	12	7	✗
Liquidity ratio	8	14	✗	4	12	✗	4	13	✗
Debtors	9	5	✗	12	5	✗	10	5	✗
Age	10	13	✗	6	9	✗	6	10	✗
Current ratio	11	12	≈	8	11	✗	9	14	✗
Intangible fixed assets	12	10	✗	15	10	✗	15	12	✗
Solvency ratio	13	11	✗	14	13	≈	11	12	≈
Number of employees	14	7	✗	9	4	✗	13	8	✗
Non-current liabilities	15	9	✗	13	15	✗	14	11	✗
Total match types			✓ 3			✓ 2			✓ 1
			≈ 2			≈ 2			≈ 3
			✗ 10			✗ 11			✗ 11

TABLE 4.9: Comparison of SHAP and Pearson correlation feature rankings across sectors, including match type counts per sector. Match types: ✓= exact, ≈ = 1-rank deviation, ✗= otherwise.

Lastly, the SHAP analysis provides a more nuanced understanding of feature relevance compared to earlier correlation-based assessments. Table 4.9 shows the importance ranking of each feature in the SHAP analysis versus prior correlation analyses (Tables A.9, A.10, and A.11). The comparison reveals only a few exact matches: three in business services, two in wholesale, and one in construction. When allowing for one-rank deviations, the number of matches marginally improves. However, in most cases, ten to eleven features per sector, there is no match between the SHAP and correlation rankings.

This discrepancy underlines SHAP's added value in identifying features with true explanatory power beyond linear correlations. SHAP, being model-based, captures complex interactions and non-linear relationships that are typically overlooked in traditional correlation metrics (Hamad et al., 2025; Wang et al., 2022). Notably, SHAP consistently identifies total assets, tangible fixed assets, and shareholders' funds among the top features across all sectors, aligning with core valuation principles (Haron, 2015; Malakauskas & Lakštutienė, 2021). Overall, SHAP enhances the interpretability of the XGBoost model by revealing the sector-specific financial logic underlying EBITDA estimation and confirming the central role of liquidity and leverage in SME EBITDA (Gama & Geraldes, 2012; Malakauskas & Lakštutienė, 2021).

4.5 Model evaluation

This section evaluates the performance and validity of the developed models, focusing on the best-performing configuration: XGBoost applied to log-transformed data. The analysis includes three components: a cross-validation check to confirm the reliability of model results, a benchmark comparison against the baseline estimation method currently used in practice, and a deeper investigation into critical classification errors. Lastly, a stylized DCF-based consistency assessment is conducted to verify whether the predicted outputs yield economically coherent valuation outcomes.

4.5.1 Cross-validation of best performing models

As an initial step in the model evaluation, the reliability of the results is assessed through 5-fold cross-validation, as outlined in Section 4.1.2. The cross-validated performance metrics for each model (MLR, RF, and XGBoost) are reported for the tests that yielded the highest model explainability per sector, as indicated by the R^2 values. These metrics are summarized in Appendix A.24 (Table A.18). The results show low standard deviations and mean values that are closely aligned with those obtained from the test sets. This consistency supports the conclusion that the models developed in this study offer stable and reliable performance in estimating EBITDA.

4.5.2 Benchmarking model output to baseline situation

To evaluate the added value of the developed models, their performance is benchmarked against the baseline estimation method currently used in practice for both the untransformed and log-transformed variables. The baseline method approximates EBITDA by multiplying a fixed multiple of 5.5 to the value of debtors resulting in revenue, where the last step yields EBITDA by multiplying the revenue with the EBITDA margin per sector. These margins, retrieved from Orbis, are on average for business services 8.79%, wholesale 9.14%, and construction 5.75% (Moody's Analytics, 2025). The performance of the current baseline method is presented in Tables 4.10.

Sector	Variable type	R^2	Bias	SMAPE	Accuracy	Critical %	Imprecision %
Business services	Normal	-0.53	29,757	82.96%	0.7868	13.15%	6.97%
Wholesale	Normal	-1.60	295,523	81.16%	0.7624	12.96%	10.80%
Construction	Normal	0.29	-191,103	86.28%	0.8141	14.98%	3.61%
Business services	Log	-0.24	0.9806	9.37%	0.7446	3.48%	22.06%
Wholesale	Log	-0.79	0.9774	10.22%	0.6967	3.66%	26.63%
Construction	Log	-0.27	0.8049	8.82%	0.6932	3.69%	26.99%

TABLE 4.10: Baseline model performance per sector and variable type: R^2, Bias, SMAPE, Accuracy, critical errors, and imprecision errors

Table 4.10 illustrates that in both the untransformed and log-transformed spaces, the baseline model yields very low or even negative R^2 values. A negative R^2 indicates that the model performs worse than a naive mean-based predictor, meaning it fails to capture meaningful variance in the target variable (Chicco et al., 2021). While negative values are mathematically valid, they fall outside the conventional interpretational range of $R^2 \in [0, 1]$, where values closer to 1 indicate higher explanatory power. As such, negative R^2 values are not interpretable as proportions of explained variance. For this reason,

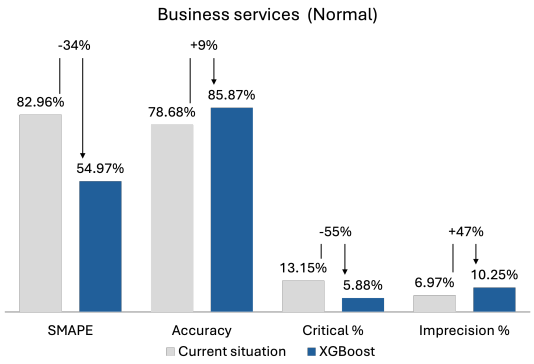
model performance is primarily compared using SMAPE, bias, and classification-based error metrics, which remain interpretable regardless of R^2.

Bias values are also extreme, ranging from 29,757 to over 295,000, highlighting large over- and underestimations. The corresponding SMAPE values exceed 80% in every case, confirming poor relative accuracy. Even when evaluated in the log-transformed space, performance remains weak. Although the bias becomes numerically smaller (due to compression in log space), R^2 scores are still negative, and SMAPE values remain around 9–10%, substantially higher than the best log-based machine learning models, which achieve SMAPE values between 4.35% and 4.60%. This shows that the baseline, even in a log-based comparison, fails to provide competitive accuracy.

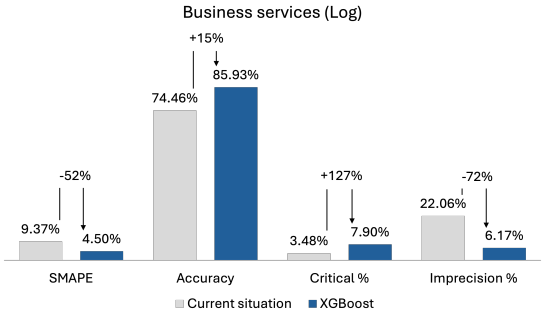
Furthermore, error-type analysis reinforces this conclusion. While the baseline sometimes maintains moderate critical error rates (e.g., 3.48% for business services in log space), it does so at the cost of significantly higher imprecision rates, reaching over 26% in wholesale and construction. In practice, this would mean that many low potential firms are incorrectly flagged as valuable, making the approach highly inefficient for deal sourcing.

In contrast, the best-performing models, particularly XGBoost with log-transformed features, achieve a much more balanced trade-off, combining low SMAPE, near-zero bias, and improved handling of both error types. This confirms that the developed models offer an improvement over traditional, rule-based estimation methods. A comparative analysis between XGBoost, achieving best results on normal and log data, and the current situation on the performance metrics is shown in Figures 4.8, 4.9, and 4.10, for each sector:

Business services



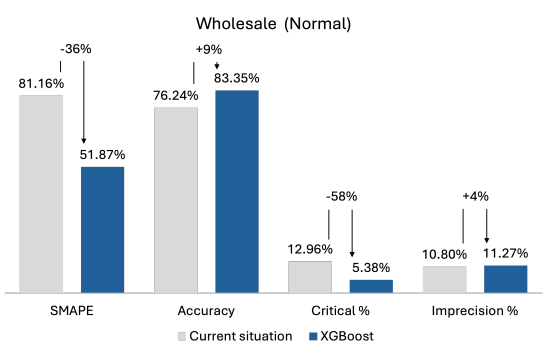
(A) Normal data



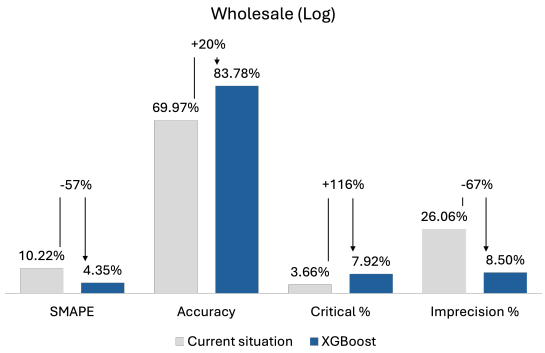
(B) Log data

FIGURE 4.8: Business services: Comparison of performance between benchmark and XGBoost model (best performance across models)

Wholesale



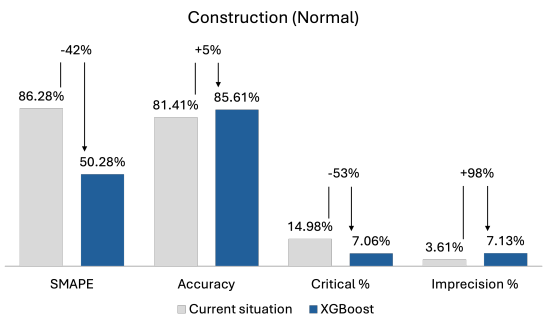
(A) Normal data



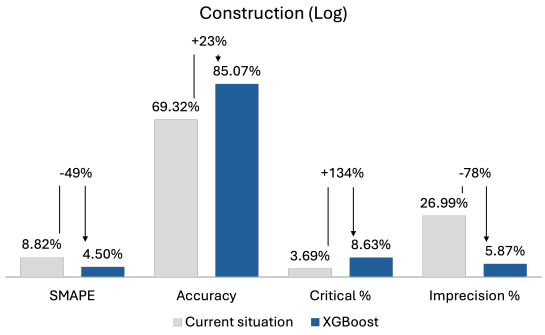
(B) Log data

FIGURE 4.9: Wholesale: Comparison of performance between benchmark and XGBoost model (best performance across models)

Construction



(A) Normal data



(B) Log data

FIGURE 4.10: Construction: Comparison of performance between benchmark and XGBoost model (best performance across models)

4.5.3 Analysis of critical classification errors

As part of the model evaluation, particular attention is directed toward the group of critical errors, cases in which the model predicted an EBITDA below €500,000 while the actual value exceeded this threshold. Comparing this group to correctly predicted instances (i.e., true positives) helps identify which features may not be adequately captured by the model, thereby contributing to misclassification. To maintain the sector-specific structure of this study, the analysis is conducted separately for each sector.

The data used in this section is derived from the model with the highest score on model explainability (R^2), XGBoost, applied to the log-transformed dataset. This configuration consistently yielded the highest predictive accuracy considering all metrics and is therefore considered as the most reliable basis for analyzing errors in classification behavior.

Descriptive statistics, including means and standard deviations, are first computed for both the critical error group and the correctly predicted group. These statistics, presented in Appendix A.23, offer an initial understanding of structural differences between the two populations. A notable finding is that the average EBITDA for the critical group

is remarkably consistent across sectors and hovers around €730,000 (reversed engineering applied to the log space variable). This suggests that, in most cases, the model fails to identify firms with EBITDA values moderately above the €500,000 threshold, pointing to a systematic underestimation near the decision boundary.

To assess whether the means of specific features differ significantly between the two groups, Welch's two-sided t-test is applied to the descriptive statistics (Appendix A.23) (Curtis, 2024; B. T. West, 2021). This test evaluates the null hypothesis (H_0) that the group means are equal against the alternative hypothesis (H_1) that they are not. Welch's t-test is preferred over the conventional Student's t-test because it does not assume equal variances between groups, an assumption that is violated in this context, as shown in Tables A.15, A.16, and A.17. The test statistic is calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{4.9}$$

The degrees of freedom are approximated using:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \tag{4.10}$$

And the associated p-value is calculated as:

$$p\text{-value} = 2 \cdot P(T > |t|), \quad T \sim t(df) \tag{4.11}$$

This procedure enables the identification of features for which the group differences are unlikely to be due to random variation, highlighting potential gaps in model sensitivity or reflecting complex financial patterns among SMEs for specific features. In this research a difference in mean is considered on a 99% confidence level, therefore only features with p-values below 0.01 are considered statistically significant. The results are shown in Table 4.11:

Feature	Business services		Wholesale		Construction	
	$\|t\text{-stat}\|$	p-value	$\|t\text{-stat}\|$	p-value	$\|t\text{-stat}\|$	p-value
Number of employees	7.437	0.000	7.401	0.000	0.871	0.385
Working capital	7.098	0.000	4.565	0.000	4.242	0.000
Current assets	6.184	0.000	5.738	0.000	5.384	0.000
Debtors	5.997	0.000	3.057	0.003	2.025	0.045
Total assets	5.704	0.000	7.253	0.000	5.227	0.000
Shareholders funds	5.384	0.000	6.293	0.000	4.072	0.000
Tangible fixed assets	5.084	0.000	3.591	0.000	0.160	0.873
Gearing	4.014	0.000	1.018	0.311	1.881	0.062
Current liabilities	3.851	0.000	4.220	0.000	4.136	0.000
Solvency ratio	3.549	0.001	1.396	0.165	0.070	0.944
Non-current liabilities	2.182	0.031	1.449	0.150	1.516	0.132
Age	1.502	0.135	0.672	0.503	1.899	0.060
Current ratio	1.352	0.179	0.074	0.941	0.324	0.747
Intangible assets	1.117	0.266	0.496	0.621	0.857	0.393
Liquidity ratio	0.865	0.388	1.704	0.091	0.067	0.947

TABLE 4.11: Welch's two-sided t-test results comparing critical and correctly predicted groups across sectors

The statistical differences observed between the critical error group and the correctly predicted firms, as shown in Table 4.11 and highlighted in yellow, suggest that certain financial features, such as total assets, working capital, current assets, current liabilities, and shareholders' funds, are not adequately captured for firms with critical error profiles across all sectors. Although these features demonstrate strong overall predictive importance, as evidenced by the SHAP values in Figures 4.5, 4.6, and 4.7, they may behave differently for SMEs with EBITDA values at €730,000. For these borderline cases, alternative handling of these variables may be necessary to mitigate the risk of critical misclassifications.

This indicates a blind spot in the current model configuration where XGBoost is used on log-transformed data. Firms that deviate from the dominant financial patterns the model has learned, are more prone to severe underestimation. To address this, further refinement is needed, such as hybrid modeling including post-model correction mechanisms tailored to this subgroup located near €500,000 EBITDA. Such adjustments may help capture these specific financial patterns of these firms more effectively and enhance the model's robustness and practical applicability.

4.5.4 Mapping model output on DCF model as validity check

As a final validation step, a sanity check is conducted on the log-transformed data to assess whether the predicted EBITDA values produce reasonable WACC estimates when translated into a stylized DCF framework. This step ensures consistency between model

outputs and widely accepted financial valuation principles. Figure 4.11 below presents a visual summary of the stepwise mapping from predicted EBITDA to implied WACC, as applied in the stylized DCF framework.

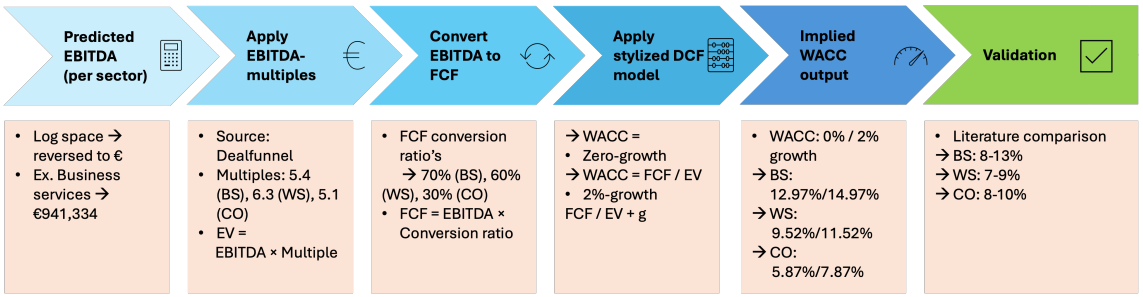


FIGURE 4.11: Stepwise flow from predicted EBITDA to implied WACC in stylized DCF validation

To begin, the predicted mean EBITDA per sector is retrieved from the log-space and translated to monetary values by reversed engineering (€941,334 for business services, €847,315 for wholesale, and €473,563 for construction). The EBITDAs are then multiplied by the corresponding sector-specific average EBITDA multiples of 5.4, 6.3, and 5.1, as obtained from Dealfunnel, Marktlink's workflow management system. These multiples are derived from real-world M&A transaction data. Applying the EBITDA-multiple valuation method yields estimated EVs of approximately €5.08 million, €5.34 million, and €2.42 million for the respective sectors.

To evaluate whether these implied values are economically reasonable, they are mapped onto a stylized constant-growth DCF model. Two cases are considered: one assuming zero growth ($g = 0$), and one assuming a constant 2% growth rate ($g = 2\%$). The zero-growth case assumes perpetual FCF with no reinvestment, reducing the DCF to a simple perpetuity where WACC $= r$ (Kagan, 2024; O'Brien, 2022):

$$\text{EV} = \frac{\text{FCF}}{r - g} = \frac{\text{FCF}}{\text{WACC}} \tag{4.12}$$

Where:

$$\text{WACC} = \frac{\text{FCF}}{\text{EV}} \tag{4.13}$$

When incorporating a non-zero growth rate, the stylized DCF formulation is adjusted as:

$$\text{WACC} = g + \frac{\text{FCF}}{\text{EV}} \tag{4.14}$$

In the absence of detailed financial data to calculate FCF directly, such as capital expenditures, changes in working capital, or taxes, sector-specific EBITDA-to-FCF conversion ratios are applied. These ratio's are derived from literature and industry reports:

- **Business services:** 70% ⇒ asset-light, low capex and working capital (Capital, 2023; S&P Global Ratings, 2024)
- **Wholesale:** 60% ⇒ moderate working capital requirements (GlobeNewswire, 2024)
- **Construction:** 30% ⇒ high working capital and capex intensity (Amwal Al Ghad, 2024; Investing.com, 2024)

These ratios approximate the proportion of EBITDA that translates into FCF:

$$\text{FCF} = \text{EBITDA} \times \text{Conversion ratio} \tag{4.15}$$

These assumptions reflect sector-level capital intensity and working capital demands, which fundamentally affect cash flow generation. For example, construction firms often face negative cash conversion cycles and heavy investment in equipment and materials, leading to lower FCF realization (Bal, 2020; Syed & Elwakil, 2019). In contrast, business services firms, being asset-light, can convert a large share of EBITDA into cash.

The final step involves estimating the sector-specific WACC by applying the DCF formulas under both growth scenarios. In the zero-growth scenario ($g = 0$), WACC equals FCF divided by EV. In the growth scenario, a uniform 2% terminal growth rate ($g = 2\%$) is applied across all sectors. This assumption is based on the observation that over time, mature firms tend to converge toward growth rates near long-term inflation (Damodaran, 2006). Specifically, the European Central Bank targets a 2% inflation rate, which is frequently used in valuation practice as a proxy for steady-state nominal growth (Big 4 Confidential, 2024; De Nederlandsche Bank, 2024). Therefore, a terminal growth assumption of 2% is considered valid, consistent, and theoretically grounded for all sectors under analysis.

The resulting implied WACC values are shown below in Table 4.12:

Sector	**WACC** ($g = 0\%$)	**WACC** ($g = 2\%$)
Business services	12.97%	14.97%
Wholesale	9.52%	11.52%
Construction	5.87%	7.87%

TABLE 4.12: WACC across sectors under zero and 2% growth assumptions

These implied WACC values align reasonably well with sector averages reported in literature and industry analyses, which estimate SME WACC levels to range between approximately 6% and 10%, depending on sector. For example, WACC in business services is typically estimated between 8% and 13%, wholesale between 7% and 9%, and construction between 8% and 10% (Damodaran, 2024; KPMG AG, 2024). The fact that the calculated values fall within or close to these ranges strengthens the internal validity of the model's EBITDA predictions. It also confirms that mapping predicted EBITDA to a simplified DCF framework delivers valuation outputs that are justifiable from both an economic and theoretical perspective. This exercise thus serves as a sanity check, verifying that model-based outputs are not only statistically sound but also aligned with financial logic and market practice.

4.6 Overview of the modeling process

To conclude this chapter, Figure 4.12 provides a structured overview of the modeling process applied in this study. It outlines the sequential steps from data preparation through model development to final validation, highlighting the techniques and transformations used at each stage. This diagram serves as a visual summary of the methodological flow and helps contextualize the evaluation results presented throughout the chapter.

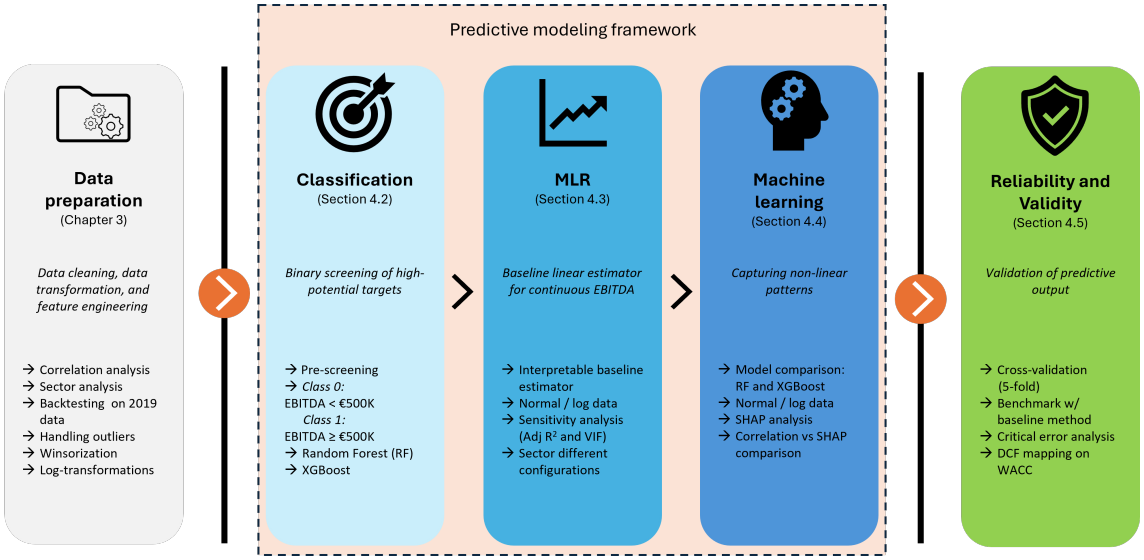


FIGURE 4.12: Overview of the modeling process

Chapter 5

Conclusion

Contents

This chapter concludes the research by answering the main research question and summarizing the key findings of the study. Section 5.1 presents the main conclusion, reflecting on how the developed model addresses the valuation challenge for SMEs using publicly available data. Section 5.2 outlines the limitations of the study, including data availability, generalizability, and the predictive nature of the models. Section 5.3 provides practical recommendations for Marktlink, focusing on model adoption, expansion across sectors, and integration into deal sourcing tools. Section 5.4 discusses the contribution to both academic literature and practice. Finally, Section 5.5 outlines suggestions for future research, including external validation, causal analysis, and broader application within other sectors.

5.1 Conclusion

This section addresses the main research question formulated in Section 1.5.2:

How can a method be developed to accurately estimate SME EBITDA based on publicly available financial data, enabling M&A advisors to improve early-stage valuation accuracy and lead sourcing efficiency?

This study demonstrates that accurate EBITDA estimation for Dutch SMEs is feasible using publicly available financial statement data. A structured approach was followed, consisting of a theoretical foundation, data preparation, and a modeling framework using MLR, RF, and XGBoost, applied across the business services, wholesale, and construction sectors. Key methodological components included sector-specific modeling, log-transformations, backtesting on historical data to account for year-specific economic effects, and a sensitivity analysis on MLR performance.

A clear distinction was made between models trained on normal (untransformed) and log-transformed data, as they serve different purposes:

- **In the normal space**, where the objective is to estimate precise euro-level EBITDA values, both XGBoost and MLR delivered promising results, though in different ways. XGBoost achieved solid R^2 values of up to 0.65 (construction), but its SMAPE remained high across all sectors, ranging from 50.28% to 54.97%. Notably, MLR performed surprisingly well in this setting, yielding low bias in both the business services (1,220) and wholesale (1,911) sectors, demonstrating its capacity to produce well-centered predictions. However, SMAPE values for MLR remained similarly high (60–72%), underscoring the challenges of absolute precision in the untransformed space. A practical advantage of normal-space models, including both MLR and XGBoost, is their tendency to yield lower critical error percentages, reducing the risk of overlooking high-potential acquisition targets.
- **In the log-transformed space**, where the emphasis shifts toward proportional accuracy and relative ranking, performance improved across all models compared to performance in normal space and to the current situation. XGBoost was the top-performing method, reaching R^2 scores of 0.71 (business services) and 0.66 (wholesale and construction), with near-zero bias and outstanding SMAPE values between 4.35% and 4.60%. These figures indicate highly reliable proportional predictions. Importantly, MLR also performed competitively in this setting, achieving R^2 values up to 0.66 (construction) and bias values close to zero, while offering clear interpretability of model coefficients. Although SMAPE for MLR remained higher than for XGBoost (ranging from 51.22% to 56.51%), it still marked a significant improvement over normal-space results. These findings emphasize that MLR remains a strong and relevant prediction method, particularly in use cases where transparency and interpretability are valued as much as performance.

To assess the practical relevance of the developed models, their performance was benchmarked against the existing rule-of-thumb approach currently used by Marktlink. The baseline estimation method currently used in practice, multiplying debtors by a fixed revenue multiplier and applying an average EBITDA margin, performed significantly worse than both statistical and machine learning models in both untransformed and log-transformed spaces. Baseline SMAPE values exceeded 80% in the normal space and ranged between 9–10% in the log space, with negative R^2 values across most sectors, confirming its limited explanatory power. In contrast, the developed models, particularly MLR and XGBoost applied on log-transformed data, demonstrated substantial improvements across nearly all performance metrics. Depending on the sector, the best configurations achieved reductions in SMAPE of 57%, and improvements in classification metrics such as accuracy (+23%), critical error rate (-58%), and imprecision (-78%).

SHAP analysis revealed that asset-related indicators, particularly total assets, tangible fixed assets, and shareholders' funds, were the most influential predictors of EBITDA when applying XGBoost on log-transformed data. These findings are economically sound and reflect sector-specific dynamics, as further validated through financial theory such as the DuPont analysis. Furthermore, mapping predicted EBITDA into a stylized DCF framework, produced implied WACC values that aligned with industry benchmarks. This confirms that the developed models not only perform statistically, but also yield outputs that are financially credible.

This research demonstrates that accurate, sector-specific EBITDA estimations can be achieved using a data-driven approach based on balance sheet items. The findings offer a

practical tool for M&A advisors, such as those at Marktlink, to better identify promising acquisition targets at an early stage. Importantly, MLR should not be overlooked: while more advanced models like XGBoost provide top performance, MLR offers a strong baseline alongside transparency. Depending on the use case, whether ranking and comparing leads or forecasting precise deal sizes, a suitable model configuration (log or normal space) can be chosen.

To conclude, this thesis presents a replicable framework for data-driven SME valuation that combines predictive accuracy with financial interpretability, offering concrete improvements over the traditional rule-of-thumb method and supporting more informed investment decisions.

5.2 Limitations

While this research demonstrates the potential of statistical and machine learning models for estimating SME EBITDA using public financial data, several limitations should be acknowledged:

- Firstly, one area where the model shows room for improvement is in predicting firms with EBITDA values clustered at €730,000. As discussed in Section 4.5.3, these borderline cases are occasionally underestimated, likely due to their deviation from dominant financial patterns. While this does not materially impact the overall model performance, it highlights a specific scenario where further refinement, such as hybrid modeling, could enhance classification precision.

- Secondly, Winsorization was applied to mitigate the influence of extreme outliers in in the dataset. While this technique reduces the influence of extreme values in the data and results in a more balanced input distribution, it introduces a trade-off by capping values at predefined percentiles. As a result, the model may underrepresent firms with highly atypical financials, which could be relevant in the context of deal sourcing. Therefore, although Winsorization enhances generalizability, it also limits the model's sensitivity to extreme cases that may hold strategic value.

- Thirdly, the models were trained using complete balance sheet data from SMEs. This does not fully reflect real-world data availability, as many SMEs disclose only partial financial information. To illustrate, this study included 54,721 SMEs with full balance sheet data, while a total of 200,574 SMEs are registered in the Netherlands and Belgium. Although additional tests showed that predictive models can still operate using a reduced set of variables, such as the top five most correlated features, this resulted in a significant decrease in predictive accuracy.

- Finally, while the modeling framework is technically transferable and could be retrained on datasets from other countries, its effectiveness depends heavily on the availability and consistency of financial disclosures. Differences in national reporting standards, variable definitions, and data quality may affect performance when applied outside the Dutch and Belgian SME context.

5.3 Recommendations

Based on the findings of this research, it is recommended that Marktlink operationalizes the developed XGBoost model, particularly the sector-specific variant using log-transformed

balance sheet data, as an integral part of its early-stage lead sourcing process. This model demonstrably outperforms the existing rule-of-thumb method, increasing predictive accuracy and reducing SMAPE by up to 57%, while offering a significant improvement in explained variance. These gains suggest substantial potential for more precise financial estimation during the early stages of acquisition screening.

To extract maximum value from this predictive framework, the model should be embedded within Marktlink's lead qualification workflow in Dealfunnel. This would enable consultants to estimate EBITDA earlier and more reliably, thereby improving prioritization of acquisition targets, and optimizing internal lead screening efficiency. These improvements align directly with Marktlink's operational objectives of scaling deal flow quality without proportionally increasing resource input.

Furthermore, it is advised that the model's predictive performance is monitored continuously using new deal flow data to ensure model robustness over time and to identify any changes or sectoral shifts. This ongoing evaluation should feed into regular model updates or training efforts. Simultaneously, development should be extended to additional sectors beyond business services, wholesale, and construction to increase the tool's coverage across Marktlink's full client base. Finally, extending the model's scope to international markets, particularly those in which Marktlink is expanding, could further improve cross-border lead generation capabilities.

5.4 Contribution to theory and practice

This research contributes to theory by addressing a clear gap in the academic literature on SME valuation. To the best of the author's knowledge, no prior studies have developed sector-specific models to estimate EBITDA using only publicly available balance sheet data. Furthermore, this study demonstrates that high predictive accuracy is achievable under such constraints, achieving a SMAPE of just 4.35% in the best case scenario, a level of precision not previously reported in the SME valuation literature.

The practical relevance of this contribution is equally significant. The model developed in this study substantially outperforms the current rule-of-thumb used in practice. By bridging the gap between data-driven forecasting and real-world deal sourcing, this research equips Marktlink with a scalable and objective tool for early-stage EBITDA estimation. This enables the firm to reduce time spent on non-viable leads and to focus resources more effectively on the most promising acquisition targets early in the M&A deal-funnel.

5.5 Future research

While this study provides a solid foundation for data-driven SME valuation in the context of early-stage lead sourcing, several opportunities exist for future research. A first next step would be to assess the generalizability of the developed model by validating its performance across other European markets. As Marktlink expands internationally, testing the model on firms outside the Dutch and Belgian landscape would provide insights into its cross-border applicability and reveal whether country-specific financial patterns influence model accuracy.

Moreover, future research could simulate real-world data constraints more realistically by training models using partial or limited financial disclosures. This would reflect the information environment typically available in the very earliest stages of lead generation, and help improve model resilience under information scarcity.

Further enhancement of model performance may also be achieved by integrating external variables beyond company-level financials. Macroeconomic indicators, sector growth, or even firm ownership characteristics could enrich the feature set and improve explanatory power, especially for capturing economic context or behavioral effects that are not included in standard financial metrics. It may also be worthwhile for future research to incorporate balance sheet normalizations, enabling the comparison between reported and adjusted EBITDA figures.

From a methodological point of view, future studies might also consider different feature transformations beyond the ones explored in this research, such as feature multiplications. This may uncover hidden relationships and improve the model's sensitivity to specific firm profiles.

Another practical direction for future research would be the critical error zone identified in Section 4.5.3, where companies with EBITDA near the €500,000 threshold were frequently misclassified. Future studies might explore targeted sub-models or other methods to reduce misclassification in this region.

Finally, future work should explore how predictive valuation tools like the one developed here can be seamlessly integrated into operational deal sourcing systems. Embedding such models into platforms like Dealfunnel, combined with interactive dashboards or risk scoring mechanisms, could enhance consultant workflows and increase adoption of data-driven decision-making in corporate finance practice.

Bibliography

Afrifa, G. A., & Padachi, K. (2016). Working capital level influence on sme profitability [Accessed: 2025-02-11]. *Journal of Small Business and Enterprise Development*, *23*(1), 44–63. https://doi.org/10.1108/JSBED-01-2014-0014

Akhtar, N., Alharthi, M. F., & Khan, M. S. (2024). Mitigating multicollinearity in regression: A study on improved ridge estimators. *Mathematics*, *12*(19), 3027. https://doi.org/10.3390/math12193027

Alabrah, A. (2023). An improved ccf detector to handle the problem of class imbalance with outlier normalization using iqr method. *Sensors*, *23*(9), 4406. https://doi.org/10.3390/s23094406

Alanazi, B. S. (2025). A comparative study of traditional statistical methods and machine learning techniques for improved predictive models. *Int. J. Anal. Appl.*, *23*(18). https://doi.org/10.28924/2291-8639-23-2025-18

Amfico. (2025). *De jaarrekening: Wat is dat nu eigenlijk? en wie moet die publiceren?* [Accessed: March 4, 2025]. https://www.amfico.be/Nieuws/Artikel/Id/97/De-jaarrekening-wat-is-dat-nu-eigenlijk-En-wie-moet-die-publiceren

Ampomah, E. K., Qin, Z., & Nyame, G. (2020). Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement. *Information*, *11*(6), 332. https://doi.org/10.3390/info11060332

Amwal Al Ghad. (2024). *Fitch projects stable outlook for european construction industry* [Accessed April 2025]. https://en.amwalalghad.com/fitch-projects-stable-outlook-for-european-construction-industry/

Andreeva, G., Calabrese, R., & Osmetti, S. A. (2016). A comparative analysis of the uk and italian small businesses using generalised extreme value models [Accessed: 2025-02-11]. *European Journal of Operational Research*, *249*, 506–516. https://doi.org/10.1016/j.ejor.2015.07.062

Astro Tax. (2025). *Zo vind je de statuten en jaarrekeningen van belgische ondernemingen* [Accessed: March 4, 2025]. https://www.astro.tax/blog/zo-vind-je-de-statuten-en-jaarrekeningen-van-belgische-ondernemingen

Bagna, E., & Ramusino, E. C. (2017). Market multiples and the valuation of cyclical companies [Accessed: 2025-02-11]. *International Business Research*, *10*(12), 246–264. https://doi.org/10.5539/ibr.v10n12p246

Bal, H. (2020). Determinants of capital structure in the construction companies across europe and central asia region. *International Conference on Eurasian Economies*, 23–28.

Bancel, F., & Mittoo, U. R. (2014). The gap between the theory and practice of corporate valuation: Survey of european experts [Accessed: 2025-02-11]. *Journal of Applied Corporate Finance*, *26*(4), 106–117. https://doi.org/10.1111/jacf.12095

Barbato, G., Barini, E. M., Genta, G., & Levi, R. (2011). Features and performance of some outlier detection methods. *Journal of Applied Statistics*, *38*(10), 2133–2149. https://doi.org/10.1080/02664763.2010.545119

Bartlett, W., & Bukvič, V. (2001). Barriers to sme growth in slovenia [Accessed: 2025-02-11]. *MOCT-MOST: Economic Policy in Transitional Economies*, *11*, 177–195. https://doi.org/10.1023/A:1012240419885

Batrancea, I., Morar, I.-D., Masca, E., Catalin, S., & Bechis, L. (2018). Econometric modeling of sme performance. case of romania. *Sustainability*, *10*(192), 1–15. https://doi.org/10.3390/su10010192

Beranová, M. (2013). The problem of accounting methods in company valuation [Accessed: 2025-02-11]. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, *61*(4), 867–872. https://doi.org/10.11118/actaun201361040867

Bergen, S., Huso, M. M., Duerr, A. E., Braham, M. A., Schmuecker, S., Miller, T. A., & Katzner, T. E. (2023). A review of supervised learning methods for classifying animal behavioural states from environmental features. *Methods in Ecology and Evolution*, *14*, 189–202. https://doi.org/10.1111/2041-210X.14019

Big 4 Confidential. (2024). How to estimate the long-term growth rate (g) in valuations [Accessed: April 10, 2025]. https://www.big4confidential.com/post/estimate-g#:~:text=In%20general%2C%20you%20would%20expect,inflation%20target

Blom, J. (2024, July). *Post-loi deal value adjustments in dutch sme m&a transactions* [Master's Thesis]. University of Twente [Accessed: 2025-02-11]. Retrieved February 11, 2025, from https://essay.utwente.nl/100544/1/Blom_MA_BMS%20%282%29.pdf

Botchkarev, A. (2018). *Evaluating performance of regression machine learning models using multiple error metrics in azure machine learning studio* (tech. rep. No. 3177507) (Electronic copy available at: https://ssrn.com/abstract=3177507). Social Science Research Network (SSRN).

Bowman, R. G., & Bush, S. R. (2006). Using comparable companies to estimate the betas of private companies [Accessed: 2025-02-11]. *Journal of Applied Finance*. Retrieved February 11, 2025, from https://ssrn.com/abstract=956443

Brighton, H., & Gigerenzer, G. (2015). The bias bias. *Journal of Business Research*, *68*(8), 1772–1784. https://doi.org/10.1016/j.jbusres.2015.01.061

Brookz. (2025). *Bedrijfswaarde en ebitda-multiples: Hoe bepaal je de waarde van een bedrijf?* [Accessed: 2025-02-11]. Retrieved February 11, 2025, from https://www.brookz.nl/kennisbank/waardebepaling/bedrijfswaarde-ebitda-multiples

Capital, A. (2023). *Comparing free cash flow metrics in tech services* [Accessed April 2025]. https://alten.capital/blog/comparing-free-cash-flow-metrics-in-tech-services

Chauhan, R., & Rameshbhai, B. Y. (2024). Working capital management in different sector. *International Journal of Innovative Science and Research Technology (IJISRT)*. https://api.semanticscholar.org/CorpusID:269961483

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623. https://doi.org/10.7717/peerj-cs.623

Chikodili, N. B., Abdulmalik, M. D., Abisoye, O. A., & Bashir, S. A. (2021). Outlier detection in multivariate time series data using a fusion of k-medoid, standardized euclidean distance and z-score. *ICTA 2020, Communications in Computer and In-*

formation Science (CCIS), *1350*, 259–271. https://doi.org/10.1007/978-3-030-69143-1_21

Corporate Finance Institute (CFI). (2025). *Ebitda multiple: Definition, formula, and examples* [Accessed: 2025-02-11]. Retrieved February 11, 2025, from https://corporatefinanceinstitute.com/resources/capital_markets/ebitda-multiple/

Cunningham, P., Cord, M., & Delany, S. J. (2024). Supervised learning. In *Machine learning for multimedia content analysis* (pp. 21–48). Springer. https://doi.org/10.1007/978-3-030-72395-0_2

Curtis, D. (2024). Welch's t test is more sensitive to real world violations of distributional assumptions than student's t test but logistic regression is more robust than either [Open access under CC BY 4.0]. *Statistical Papers*, *65*, 3981–3989. https://doi.org/10.1007/s00362-024-01531-7

Dagblad, F. (2025). *Minder mkb-bedrijven onder verplichte controle accountant* [Accessed: March 3, 2025]. https://fd.nl/bedrijfsleven/1493892/minder-mkb-bedrijven-onder-verplichte-controle-accountant

Damodaran, A. (2006, November). *Valuation approaches and metrics: A survey of the theory and evidence* (tech. rep.) (Accessed: 2025-02-11). Stern School of Business, New York University. Retrieved February 11, 2025, from https://pages.stern.nyu.edu/~adamodar/pdfiles/papers/valuesurvey.pdf

Damodaran, A. (2012). *Investment valuation: Tools and techniques for determining the value of any asset* (3rd) [Accessed: 2025-02-11]. John Wiley & Sons. Retrieved February 11, 2025, from https://www.wiley.com/en-us/Investment+Valuation%3A+Tools+and+Techniques+for+Determining+the+Value+of+Any+Asset%2C+3rd+Edition-p-9781118130735

Damodaran, A. (2016). *Investment valuation: Tools and techniques for determining the value of any asset* (3rd). John Wiley Sons.

Damodaran, A. (2024). *Cost of capital by sector (us)* [Accessed April 2025]. https://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/wacc.html

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and roc curves. *Proceedings of the 23rd international conference on Machine learning*, 233–240.

Dawoud, I., & Eledum, H. (2025). Detection of influential observations for the regression model in the presence of multicollinearity: Theory and methods [Published online: 09 Jan 2025]. *Communications in Statistics - Theory and Methods*. https://doi.org/10.1080/03610926.2024.2449107

De Nederlandsche Bank. (2024). The ecb's monetary policy [Accessed: April 10, 2025]. https://www.dnb.nl/en/the-euro-and-europe/the-ecb-s-monetary-policy/#:~:text=The%20price%20stability%20sought%20by,best%20when%20prices%20are%20stable

Dertli, H. I., Hayes, D. B., & Zorn, T. G. (2024). Effects of multicollinearity and data granularity on regression models of stream temperature. *Journal of Hydrology*, *639*, 131572. https://doi.org/10.1016/j.jhydrol.2024.131572

Dunne, P., & Hughes, A. (1994). Age, size, growth and survival: Uk companies in the 1980s [Accessed: 2025-02-11]. *The Journal of Industrial Economics*, *42*(2), 115–140. https://doi.org/10.2307/2950612

Fischer, T., & Krauss, C. (2017). *Deep learning with long short-term memory networks for financial market predictions* (tech. rep. No. 11/2017). Friedrich-Alexander University Erlangen-Nürnberg, Institute for Economics. https://hdl.handle.net/10419/157808

Fissler, T., & Hoga, Y. (2024). Backtesting systemic risk forecasts using multi-objective elicitability. *Journal of Business & Economic Statistics*, *42*(2), 485–498. https://doi.org/10.1080/07350015.2023.2200514

for Consumers, A., & (ACM), M. (2015, June). *Competition on the dutch sme loan market* (Accessed: 2025-02-11). Authority for Consumers and Markets (ACM). Retrieved February 11, 2025, from https://www.acm.nl/sites/default/files/old_publication/publicaties/14681_report-competition-on-the-dutch-sme-loan-market-june-2015.pdf

Frank, E., Wang, Y., Inglis, S., Holmes, G., & Witten, I. H. (1998). Using model trees for classification [Technical Note]. *Machine Learning*, *32*, 63–76.

Gama, A. P. M., & Geraldes, H. S. A. (2012). Credit risk assessment and the impact of the new basel capital accord on small and medium-sized enterprises: An empirical analysis [Accessed: 2025-02-11]. *Management Research Review*, *35*(8), 727–749. https://doi.org/10.1108/01409171211247712

GlobeNewswire. (2024). *Rexel q3 2024 sales* [Accessed April 2025]. https://www.globenewswire.com/news-release/2024/10/15/2963505/0/en/Rexel-Q3-2024-Sales.html

Hamad, K., Alotaibi, E., Zeiada, W., Al-Khateeb, G., Abu Dabous, S., Omar, M., Mantha, B. R., Arab, M. G., & Merabtene, T. (2025). Explainable artificial intelligence visions on incident duration using extreme gradient boosting and shapley additive explanations. *Multimodal Transportation*, *4*, 100209. https://doi.org/10.1016/j.multra.2025.100209

Haron, R. (2015). Modelling debt financing behaviour of malaysia smes [Accessed: 2025-02-11]. *2015 International Symposium on Technology Management and Emerging Technologies (ISTMET)*, 297–302. https://doi.org/10.1109/ISTMET.2015.00057

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd). Springer. https://doi.org/10.1007/b94608_2

Heerkens, H., & van Winden, A. (2017). *Solving managerial problems systematically*. Noordhoff Uitgevers B.V.

IBM. (2025). *Supervised learning* [Accessed: 2025-03-20]. https://www.ibm.com/think/topics/supervised-learning

Investing.com. (2024). *Earnings call: Construction partners sees robust growth in fiscal 2024* [Accessed April 2025]. https://www.investing.com/news/transcripts/earnings-call-construction-partners-sees-robust-growth-in-fiscal-2024-93CH-3735403

Investopedia. (2025). *Gearing ratio definition* [Accessed: March 4, 2025]. https://www.investopedia.com/terms/g/gearingratio.asp

Jenkins, D. S., & Kane, G. D. (2006). A contextual analysis of income and asset-based approaches to private equity valuation [Accessed: 2025-02-11]. *Accounting Horizons*, *20*(1), 19–35. https://doi.org/10.2308/acch.2006.20.1.19

Jensen, C. S., Pedersen, T. B., & Thomsen, C. (2010). *Multidimensional databases and data warehousing* (Vol. 9). Morgan & Claypool Publishers. https://doi.org/10.2200/S00299ED1V01Y201009DTM009

Johnsen, P. C., & McMahon, R. G. (2005). Cross-industry differences in sme financing behaviour: An australian perspective [Accessed: 2025-02-11]. *Journal of Small Business and Enterprise Development*, *12*(2), 160–177. https://doi.org/10.1108/14626000510594584

Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *15*(4), 531–538. https://doi.org/10.1002/sam.11583

Kagan, J. (2024). Perpetuity: Financial definition, formula, and examples [Updated April 07, 2024. Reviewed by Melody Bell. Fact checked by Michael Rosenston. Accessed: 2025-04-10].

Kamer van Koophandel (KVK). (2025). Overzicht standaard bedrijfsindeling (sbi)-codes voor activiteiten [Accessed: 2025-02-11]. Retrieved February 11, 2025, from https://www.kvk.nl/over-het-handelsregister/overzicht-standaard-bedrijfsindeling-sbi-codes-voor-activiteiten/

Kamer van Koophandel (KvK). (2025). *Uiterste termijn deponeren jaarrekening* [Accessed: 2025-02-11]. Retrieved February 11, 2025, from https://www.kvk.nl/deponeren/uiterste-termijn-deponeren-jaarrekening/

Kannan, K. S., K., M., & Arumugam, S. (2015). Labeling methods for identifying outliers. *International Journal of Statistics and Systems*, *10*(2), 231–238. https://www.researchgate.net/publication/283755180_Labeling_Methods_for_Identifying_Outliers

Karch, J. D. (2020). Improving on adjusted r-squared. *Collabra: Psychology*, *6*(1), 45. https://doi.org/10.1525/collabra.343

Koller, T., Goedhart, M., & Wessels, D. (2012). *Valuation: Measuring and managing the value of companies* (4th). John Wiley Sons.

KPMG AG. (2024). *Cost of capital study* [Accessed April 2025]. https://kpmg.com/de/en/home/insights/overview/cost-of-capital.study.html

Leach, L. F., & Henson, R. K. (2007). The use and impact of adjusted r^2 effects in published regression research. *Multiple Linear Regression Viewpoints*, *33*(1), 1–11.

Li, L., Yousif, M., & El-Kanj, N. (2023). Prediction of corporate financial distress based on digital signal processing and multiple regression analysis. *Applied Mathematics and Nonlinear Sciences*, *8*(1), 2209–2220. https://doi.org/10.2478/amns.2022.2.0140

Li, Z. (2022). Extracting spatial effects from machine learning model using local interpretation method: An example of shap and xgboost. *Computers, Environment and Urban Systems*, *96*, 101845. https://doi.org/10.1016/j.compenvurbsys.2022.101845

Lotti, F., Santarelli, E., & Vivarelli, M. (2003). Does gibrat's law hold among young, small firms? [Accessed: 2025-02-11]. *Journal of Evolutionary Economics*, *13*(3), 213–235. https://doi.org/10.1007/s00191-003-0153-0

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 4765–4774.

Lwango, A., Coeurderoy, R., & Roche, G. A. G. (2017). Family influence and sme performance under conditions of firm size and age [Accessed: 2025-02-11]. *Journal of Small Business and Enterprise Development*. https://doi.org/10.1108/JSBED-11-2016-0174

M&A Community. (2024). *League tables 2024: Deze adviseurs deden de meeste en meest waardevolle deals* [Accessed: 2025-02-11]. Retrieved February 11, 2025, from https://mena.nl/artikel/league-tables-2024-deze-adviseurs-deden-de-meeste-en-meest-waardevolle-deals/

Magnimetrics. (2025). *Ebitda multiple for business valuation* [Accessed: 2025-02-11]. Retrieved February 11, 2025, from https://magnimetrics.com/ebitda-multiple-for-business-valuation/

Mahesh, B. (2020). Machine learning algorithms - a review. *International Journal of Science and Research (IJSR)*, *9*(1), 381–386. https://doi.org/10.21275/ART20203995

Malakauskas, A., & Lakštutienė, A. (2021). The application of artificial intelligence tools in creditworthiness modelling for sme entities [Accessed: 2025-02-11]. *2021 IEEE*

International Conference on Technology and Entrepreneurship (ICTE), 1–7. https://doi.org/10.1109/ICTE51655.2021.9584528

Manikandan, S. (2010). Data transformation. *Journal of Pharmacology & Pharmacotherapeutics*, *1*(2), 126–127. https://doi.org/10.4103/0976-500X.72373

Markus, G., & Rideg, A. (2021). Understanding the connection between smes' competitiveness and cash flow generation: An empirical analysis from hungary [Accessed: 2025-02-11]. *Competitiveness Review: An International Business Journal*, *31*(3), 397–419. https://doi.org/10.1108/CR-01-2020-0019

Mauboussin, M. J. (2018, September). *What does an ev/ebitda multiple mean?* (Tech. rep.) (Accessed: 2025-02-11). BlueMountain Investment Research. Retrieved February 11, 2025, from https://www.bluemountaincapital.com

Meitner, M. (2006). *The market approach to comparable company valuation* (Vol. 35) [Accessed: 2025-02-11]. Physica-Verlag. Retrieved February 11, 2025, from https://www.amazon.com/Approach-Comparable-Company-Valuation-Economic/dp/3790817228

Mekelburg, E., & Strauss, J. (2024). Pooling and winsorizing machine learning forecasts to predict stock returns with high-dimensional data. *Journal of Empirical Finance*, *79*, 101538. https://doi.org/10.1016/j.jempfin.2024.101538

Moody's Analytics. (2025). *Orbis: Global company reference data* [Accessed: March 4, 2025]. https://www.moodys.com/web/en/us/capabilities/company-reference-data/orbis.html

Morais, É. T., Barberes, G. A., Souza, I. V. A. F., Leal, F. G., Guzzo, J. V. P., & Spigolon, A. L. D. (2023). Pearson correlation coefficient applied to petroleum system characterization: The case study of potiguar and reconcavo basins, brazil. *Geosciences*, *13*(9), 282. https://doi.org/10.3390/geosciences13090282

Nenkov, D., & Hristozov, Y. (2022). Dcf valuation of companies: Exploring the interrelation between revenue and operating expenditures [Accessed: 2025-02-11]. *Economic Alternatives*, *Issue 4*, 626–646. https://doi.org/10.37075/EA.2022.4.04

Nissim, D. (2024, June). *Ebitda, ebita or ebit?* (Tech. rep.) (Accessed: 2025-02-11). Columbia Business School. Retrieved February 11, 2025, from https://papers.ssrn.com/abstract_id=2999675

Nti, I. K., Nyarko-Boateng, O., & Aning, J. (2021). Performance of machine learning algorithms with different k values in k-fold cross-validation. *International Journal of Information Technology and Computer Science*, *13*(6), 61–71. https://doi.org/10.5815/ijitcs.2021.06.05

O'Brien, T. J. (2022). Cross-border valuation using the international capm and the constant perpetual growth model. *Journal of Economics and Business*, *119*, 106042. https://doi.org/10.1016/j.jeconbus.2021.106042

Ortiz, R., Contreras, M., & Mellado, C. (2023). Regression, multicollinearity and markowitz. *Finance Research Letters*, *58*, 104550. https://doi.org/10.1016/j.frl.2023.104550

Osborne, J. W. (2002). Notes on the use of data transformation. *Practical Assessment, Research Evaluation*, *8*(6). http://pareonline.net/getvn.asp?v=8&n=6

Overheid.nl. (2025). *Burgerlijk wetboek boek 2 titel 9 - jaarrekening en verslaggeving* [Accessed: 2025-02-11]. Retrieved February 11, 2025, from https://wetten.overheid.nl/BWBR0003045/2025-01-01#Boek2_Titeldeel9

Palepu, K. G., & Healy, P. M. (2012). *Business analysis and valuation: Using financial statements* (5th). Cengage Learning.

Parliament, E., & of the European Union, C. (2013). Directive 2013/34/eu of the european parliament and of the council of 26 june 2013 on the annual financial state-

ments, consolidated financial statements and related reports of certain types of undertakings, amending directive 2006/43/ec of the european parliament and of the council and repealing council directives 78/660/eec and 83/349/eec [Accessed: 2025-02-11]. Retrieved February 11, 2025, from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02013L0034-20240528

Pawluszek-Filipiak, K., & Borkowski, A. (2020). On the importance of train–test split ratio of datasets in automatic landslide detection by supervised classification. *Remote Sensing*, *12*(18), 3054. https://doi.org/10.3390/rs12183054

Rabobank. (2025). *Werkkapitaal bepalen: Hoe bereken je je werkkapitaal?* [Accessed: March 4, 2025]. https://www.rabobank.nl/bedrijven/groei/financien/werkkapitaal-bepalen

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1), 6086. https://doi.org/10.1038/s41598-024-56706-x

Rezaee, Z., Aliabadi, S., Dorestani, A., & Rezaee, N. J. (2020). Application of time series models in business research: Correlation, association, causation. *Sustainability*, *12*(4833), 1–17. https://doi.org/10.3390/su12124833

Ribal, J., Blasco, A., & Segura, B. (2010). Estimation of valuation multiples of spanish unlisted food companies [Accessed: 2025-02-11]. *Spanish Journal of Agricultural Research*, *8*(3), 547–558. https://doi.org/10.5424/sjar/2010083-1250

Rikkers, F., & Thibeault, A. E. (2009). A structural form default prediction model for smes, evidence from the dutch market. *Multinational Finance Journal*, *13*(3/4), 229–264. https://ssrn.com/abstract=2622995

Saus-Sala, E., Farreras-Noguer, M., Arimany-Serrat, N., & Coenders, G. (2020). Compositional dupont analysis: A visual tool for strategic financial performance assessment. *Preprint*. https://doi.org/10.13140/RG.2.2.14168.85760

Schammo, P. (2018). The eu securities law framework for smes: Can firms and investors meet? [Accessed: 2025-02-11]. In D. Busch, E. Avgouleas, & G. Ferrarini (Eds.), *Capital markets union in europe* (pp. 355–380). Oxford University Press. https://doi.org/10.1093/oso/9780198815815.003.0014

Scornet, E. (2016). On the asymptotics of random forests. *Journal of Multivariate Analysis*, *146*, 72–83. https://doi.org/10.1016/j.jmva.2015.06.009

Shabani, H., Morina, F., & Berisha, A. (2021). Financial performance of the smes sector in kosovo: An empirical analysis using the dupont model. *International Journal of Sustainable Development and Planning*, *16*(5), 819–831. https://doi.org/10.18280/ijsdp.160503

S&P Global Ratings. (2024). *European software business services: How rising rates are impacting credit quality* (tech. rep.) (Accessed April 2025). S&P Global. https://www.spglobal.com/_assets/documents/ratings/research/101612488.pdf

Steiger, F. (2008). The validity of company valuation using discounted cash flow methods [Accessed: 2025-02-11]. *Seminar Paper*.

Steijvers, T. (2004). Existence of credit rationing for smes in the belgian corporate bank loan market. *SSRN Electronic Journal*. https://ssrn.com/abstract=495162

Stokkers, D. (2024). *Estimating ebitda for smes* [Master's Thesis]. University of Twente [Accessed: 2025-02-11]. Retrieved February 11, 2025, from https://essay.utwente.nl/97964/1/Stokkers_MA_BMS.pdf

Strategic Direction Editorial Team. (2014). Ten top tips for small to medium enterprise (sme) success [Accessed: 2025-02-11]. *Strategic Direction*, *30*(2), 14–17. https://doi.org/10.1108/SD-02-2014-0005

Sullivan, J. H., Warkentin, M., & Wallace, L. (2021). So many ways for assessing outliers: What really works and does it matter? *Journal of Business Research*, *132*, 530–543. https://doi.org/10.1016/j.jbusres.2021.03.066

Syed, S., & Elwakil, E. (2019). Project performance index for capital intensive construction projects. In D. Ozevin, H. Ataei, M. Modares, A. Gurgun, S. Yazdani, & A. Singh (Eds.), *Interdependence between structural engineering and construction management*. ISEC Press.

Tatachar, A. V. (2021). Comparative assessment of regression models based on model evaluation metrics. *International Research Journal of Engineering and Technology (IRJET)*, *8*(9), 853–860. https://www.irjet.net/archives/V8/i9/IRJET-V8I9164.pdf

Tian, Y., Tian, J., Dong, M., Ai, L., Bu, L., & Sun, R. (2024). Correlation analysis of highway asphalt pavement distress based on spearman correlation coefficient. *E3S Web of Conferences*, *512*, 02027. https://doi.org/10.1051/e3sconf/202451202027

Tong, Y., & Serrasqueiro, Z. (2020). A study on the influence of financial factors on the growth of small and medium-sized enterprises in portuguese high technology and medium-high technology sectors [Accessed: 2025-02-11]. *WSEAS Transactions on Business and Economics*, *17*, 703–716. https://doi.org/10.37394/23207.2020.17.68

Uyanık, G. K., & Güler, N. (2013). A study on multiple linear regression analysis. *Procedia - Social and Behavioral Sciences*, *106*, 234–240. https://doi.org/10.1016/j.sbspro.2013.12.027

Varoquaux, G., & Colliot, O. (2023). Evaluating machine learning models and their diagnostic value. In O. Colliot (Ed.), *Machine learning for brain disorders* (pp. 601–630, Vol. 197). Springer. https://doi.org/10.1007/978-1-0716-3195-9_20

Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., & Tysklind, M. (2022). Towards better process management in wastewater treatment plants: Process analytics based on shap values for tree-based machine learning methods. *Journal of Environmental Management*, *301*, 113941. https://doi.org/10.1016/j.jenvman.2021.113941

West, B. T. (2021). Best practice in statistics: Use the welch t-test when testing the difference between two groups. *Frontiers in Psychology*, *12*, 3028. https://doi.org/10.3389/fpsyg.2021.653473

West, R. M. (2022). Best practice in statistics: The use of log transformation. *Annals of Clinical Biochemistry*, *59*(3), 162–165. https://doi.org/10.1177/00045632211050531

Wikipedia contributors. (2024). Dupont analysis — wikipedia, the free encyclopedia [Accessed: 2024-04-01].

Williams, M. N., Grajales, C. A. G., & Kurkiewicz, D. (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research & Evaluation*, *18*(11). https://doi.org/10.7275/55hn-wk47

Żelazowski, K. (2015). Application of multiple-based methods in valuation of real estate development companies [Accessed: 2025-02-11]. *Real Estate Management and Valuation*, *23*(3), 26–35. https://doi.org/10.1515/remav-2015-0022

Zhu, M., Zhang, Y., Gong, Y., Xing, K., Yan, X., & Song, J. (2024). Ensemble methodology: Innovations in credit default prediction using lightgbm, xgboost, and localensemble. *2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI)*, 421–426. https://doi.org/10.1109/ICETCI61221.2024.10594630

Appendix A

Appendix

A.1 SME definition

Type of enterprise	Employees	Annual turnover	Balance sheet total
Medium-sized enterprise	50-250	€10–€50 million	€10–€43 million
Small enterprise	10-49	€2–€10 million	€2–€10 million
Micro enterprise (including self-employed persons)	< 10	€2 million	€2 million

TABLE A.1: SME classification based on employees, turnover, and balance sheet total (for Consumers & (ACM), 2015)

A.2 Factors influencing SME profitability

Factor	Variable	Unit
1) Firm's size	Number of employees	Number
2) Firm's age	Number of years active	Number
3) Sector of operation	SBI code (Industry classification)	Non-numerical value
4) Leverage	Solvency ratio	Ratio
	Debt-to-assets ratio	Ratio
	Debt-to-equity ratio	Ratio
	Debt-to-EBITDA ratio	Ratio
	Total liabilities	Value
	Long-term liabilities	Value
	Short-term liabilities	Value
	Current liabilities	Value
	Owner's equity	Number
	Return on assets (ROA)	Ratio
	Return on equity (ROE)	Ratio
5) Liquidity	Cash and marketable securities	Value
	Current assets	Value
	Total assets	Value
	Current ratio	Ratio
	Working capital	Value
	Change in working capital	Percentage
6) Coverage	Interest coverage ratio	Ratio
7) Activity	Sales turnover	Value
	Working capital turnover	Ratio
	Asset turnover ratio	Ratio
	Debtor collection period	Days
	Number of directors	Value
	Shareholder funds	Value
8) Management expertise	Managerial experience	Number of years

TABLE A.2: Overview of factors affecting SME profitability

A.3 Literature Review: Inclusion and exclusion criteria

Criterion	Inclusion criteria	Exclusion criteria	Explanation
Time frame	Publications from 2000 with more than 5 citations	Studies published before 2000	Ensures relevance to modern SME valuation and financial modeling.
Language	English and Dutch articles	Non-English, and Non-Dutch articles	Focuses on accessible and relevant research.
Publication type	Peer-reviewed articles, conference papers, government/financial reports	Blogs, opinion pieces, unverified online sources	Ensures credibility and academic level.
Research population	Research on SMEs, privately held firms, and mid-sized enterprises	Research on publicly listed firms	Ensures a focus on SMEs rather than large corporations.
Methodology	Empirical studies, systematic reviews, and theoretical models. Studies including Qualitative, Quantitative, and Mixed-Method approaches	Studies lacking a clear methodology, non-financial research	Ensures that selected studies use structured, replicable research methods.
Applicability	Research on SME valuation methods, and public financial data	Research focused on non-financial related topics for SMEs	Ensures the research aligns with publicly available financial data sources and estimation models.

Table A.3: Inclusion and exclusion criteria for literature review

A.4 Literature review: Key concepts

Key concepts	Related terms	Narrower terms	Broader terms
SME valuation methods	Business valuation, company valuation, firm valuation	Market-based valuation, Income-based valuation, Asset-based valuation	Corporate finance, investment valuation
EBITDA	Earnings, operating profit, cash flow	Adjusted EBITDA, normalized EBITDA, sector-specific EBITDA	Profitability, financial performance
Public financial data	Financial disclosure, financial transparency, reporting standards	SME financial statements, open financial databases, regulatory filings	Corporate reporting, financial regulation
Financial structure	Financial indicators, performance metrics, key financial ratios	Leverage ratios, liquidity ratios, activity ratios	Financial health, financial analysis
Influence	Impact, effect, causation, correlation	Market trends, business cycles, financial policies	Economic behavior, industry dynamics
Factors	Determinants, drivers, variables	Financial factors, operational factors, external factors	Business conditions, economic influences

Table A.4: Key concepts with related, narrower, and broader terms

A.5 Literature review: Databases

Database	Website
Scopus	www.scopus.com
UT Library	www.utwente.nl/library
Web of Science	www.webofscience.com

Table A.5: List of applied databases and their websites

A.6 Literature review: Included articles

Author	Article title	Keywords
Afrifa, G. A., & Padachi, K.	Working capital level influence on SME profitability	Working capital, SME profitability, liquidity, financial management
Andreeva, G., Calabrese, R., & Osmetti, S. A.	A comparative analysis of the UK and Italian small businesses using Generalised Extreme Value models	SME credit risk, financial modeling, extreme value theory
Bagna, E., & Cotta Ramusino, E.	Market Multiples and the Valuation of Cyclical Companies	Valuation multiples, cyclical firms, financial performance, market trends
Bancel, F., & Mittoo, U. R.	The Gap between the Theory and Practice of Corporate Valuation: Survey of European Experts	Corporate valuation, theory-practice gap, financial modeling
Bartlett, W., & Bukvič, V.	Barriers to SME Growth in Slovenia	SME growth, financial constraints, market barriers
Batrancea, I., Morar, I.-D., Masca, E., Catalin, S., & Bechis, L.	Econometric Modeling of SME Performance. Case of Romania	SME performance, econometrics, financial modeling
Beranova, M.	The Problem of Accounting Methods in Company Valuation	Accounting methods, financial reporting, company valuation
Bowman, R. G., & Bush, S. R.	Using Comparable Companies to Estimate the Betas of Private Companies	Comparable analysis, private company valuation, beta estimation
Damodaran, A.	Investment Valuation: Tools and Techniques for Determining the Value of Any Asset	Valuation methods, investment analysis, asset valuation
Dunne, P., & Hughes, A.	Age, Size, Growth and Survival: UK Companies in the 1980s	Firm growth, company size, business lifecycle, survival analysis
European Parliament and Council of the EU	Directive 2013/34/EU on annual financial statements	Financial disclosure, SME reporting standards, EU regulations
Gama, A. P. M., & Geraldes, H. S. A.	Credit risk assessment and the impact of the New Basel Capital Accord on SMEs: An empirical analysis	Credit risk, Basel Accord, SME finance
Haron, R.	Modelling Debt Financing Behaviour of Malaysia SMEs	SME financing, debt management, financial behavior
Jenkins, D. S., & Kane, G. D.	A Contextual Analysis of Income and Asset-Based Approaches to Private Equity Valuation	Private equity, valuation approaches, asset-based valuation
Koller, T., Goedhart, M., & Wessels, D.	Valuation: Measuring and Managing the Value of Companies	Business valuation, financial modeling, corporate finance

Table A.6: List of articles, including their authors and keywords

Author	Article title	Keywords
Li, L., Yousif, M., & El-Kanj, N.	Prediction of corporate financial distress based on digital signal processing and multiple regression analysis	Financial distress prediction, regression analysis, corporate finance
Lwango, A., Coeurderoy, R., & Giménez Roche, G. A.	Family influence and SME performance under conditions of firm size and age	Family businesses, SME performance, firm age
Mauboussin, M. J.	What Does an EV/EBITDA Multiple Mean?	Valuation multiples, EBITDA, investment analysis
Malakauskas, A., & Lakštutienė, A.	The Application of Artificial Intelligence Tools in Creditworthiness Modelling for SME Entities	AI in finance, SME creditworthiness, financial modeling
Markus, G., & Rideg, A.	Understanding the connection between SMEs' competitiveness and cash flow generation: an empirical analysis from Hungary	SME competitiveness, cash flow, financial stability
Meitner, M.	The Market Approach to Comparable Company Valuation	Market valuation, company comparison, financial modeling
Nenkov, D., & Hristozov, Y.	DCF Valuation of Companies: Exploring the Interrelation Between Revenue and Operating Expenditures	Discounted Cash Flow (DCF), revenue analysis, SME valuation
Palepu, K. G., & Healy, P. M.	Business Analysis and Valuation: Using Financial Statements	Business analysis, financial statements, valuation techniques
Ribal, J., Blasco, A., & Segura, B.	Estimation of Valuation Multiples of Spanish Unlisted Food Companies	Valuation multiples, food industry, private companies
Schammo, P.	The EU Securities Law Framework for SMEs: Can Firms and Investors Meet?	SME capital markets, securities law, investment barriers
Steiger, F.	The Validity of Company Valuation Using Discounted Cash Flow Methods	DCF valuation, financial forecasting, valuation accuracy
Strategic Direction Editorial Team	Ten Top Tips for Small to Medium Enterprise (SME) Success	SME strategy, business growth, operational efficiency
Tong, Y., & Serrasqueiro, Z.	A Study on the Influence of Financial Factors on the Growth of Small and Medium-Sized Enterprises in Portuguese High Technology and Medium-High Technology Sectors	SME growth, financial factors, high-tech industries
Zelazowski, K.	Application of Multiple-Based Methods in Valuation of Real Estate Development Companies	Real estate valuation, multiple-based methods, financial performance

Table A.6: List of articles, including their authors and keywords

A.7 Descriptive statistics per sector

Feature	Business services		Wholesale		Construction	
	Mean	Std	Mean	Std	Mean	Std
EBITDA	1,101,583	1,241,788	971,889	1,097,087	551,373	803,921
Number of employees	44.01	50.44	23.51	28.42	19.63	26.23
Intangible assets	714,203	3,945,820	124,289	1,080,174	29,788	281,292
Tangible fixed assets	2,076,804	4,943,041	1,110,910	2,623,561	1,230,841	6,293,688
Current assets	7,069,352	34,598,428	7,365,656	13,499,410	3,149,471	6,873,453
Debtors	2,837,886	18,081,691	2,760,975	6,453,675	1,197,419	2,713,922
Total assets	12,889,768	63,148,574	11,662,296	127,473,700	4,853,148	15,764,463
Shareholders funds	5,808,205	42,645,420	6,090,630	121,294,100	1,968,888	7,264,150
Non-current liabilities	2,033,232	14,254,832	887,202	5,594,890	866,770	6,174,099
Current liabilities	5,041,110	28,960,078	4,684,463	12,016,740	2,017,491	5,184,023
Total shareholders' funds and liabilities	12,889,768	63,148,574	11,662,296	127,473,700	4,853,148	15,764,463
Working capital	1,701,652	6,589,498	2,760,910	9,191,816	1,060,934	3,510,428
Solvency ratio	38.21	26.05	42.25	24.99	39.82	23.12
Current ratio	2.03	2.87	2.34	2.85	2.13	2.29
Liquidity ratio	1.76	2.72	1.58	2.45	1.75	2.09
Gearing	58.31	135.69	66.20	122.23	91.52	132.34
Age	15.98	16.19	27.07	18.35	22.04	17.96

TABLE A.7: Descriptive statistics (mean and standard deviation) across business services, wholesale, and construction sectors

A.8 Distribution of variables: Business services

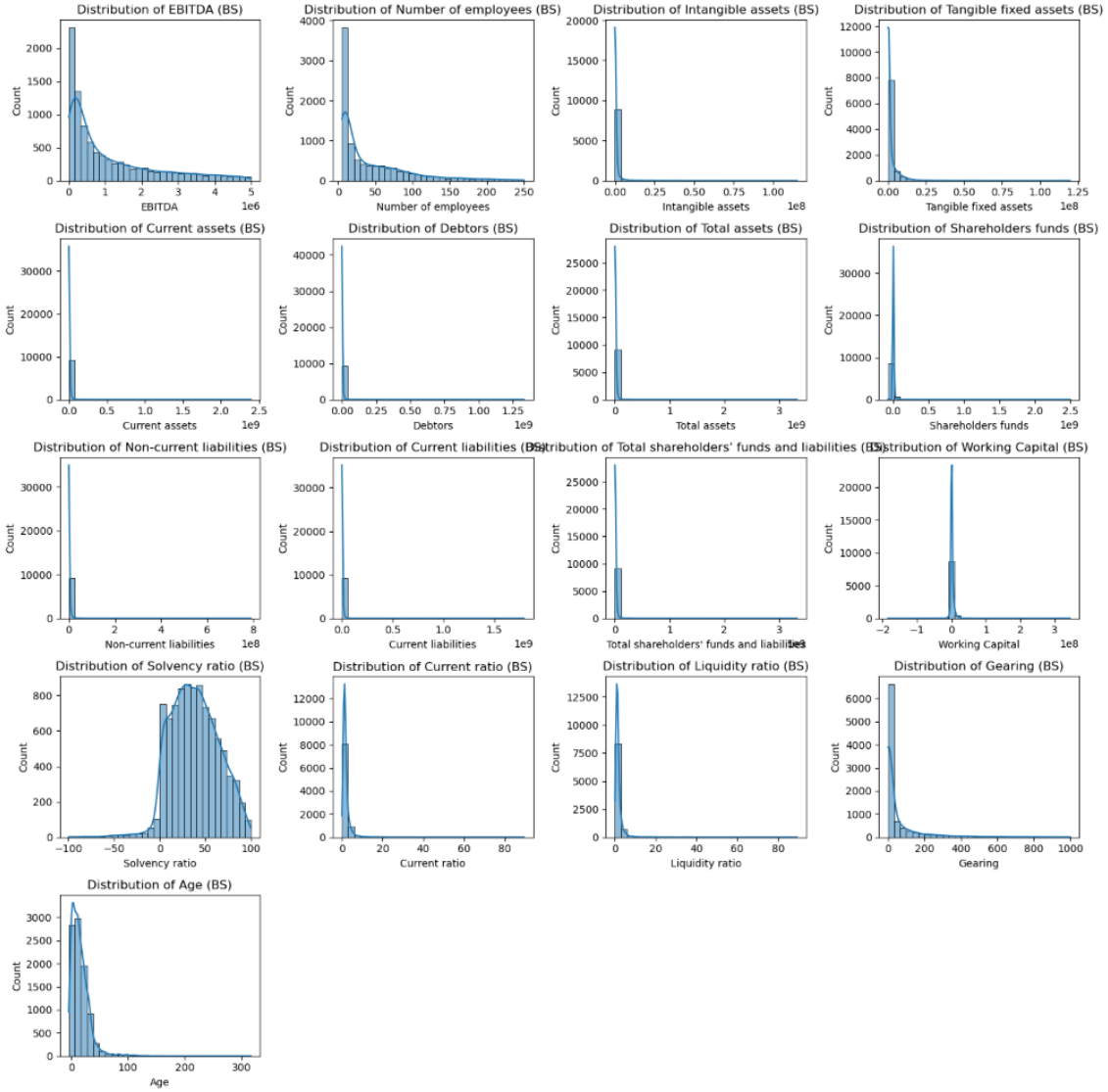


FIGURE A.1: Baseline scenario: Distribution of each feature for the business services sector

A.9 Distribution of variables: Wholesale

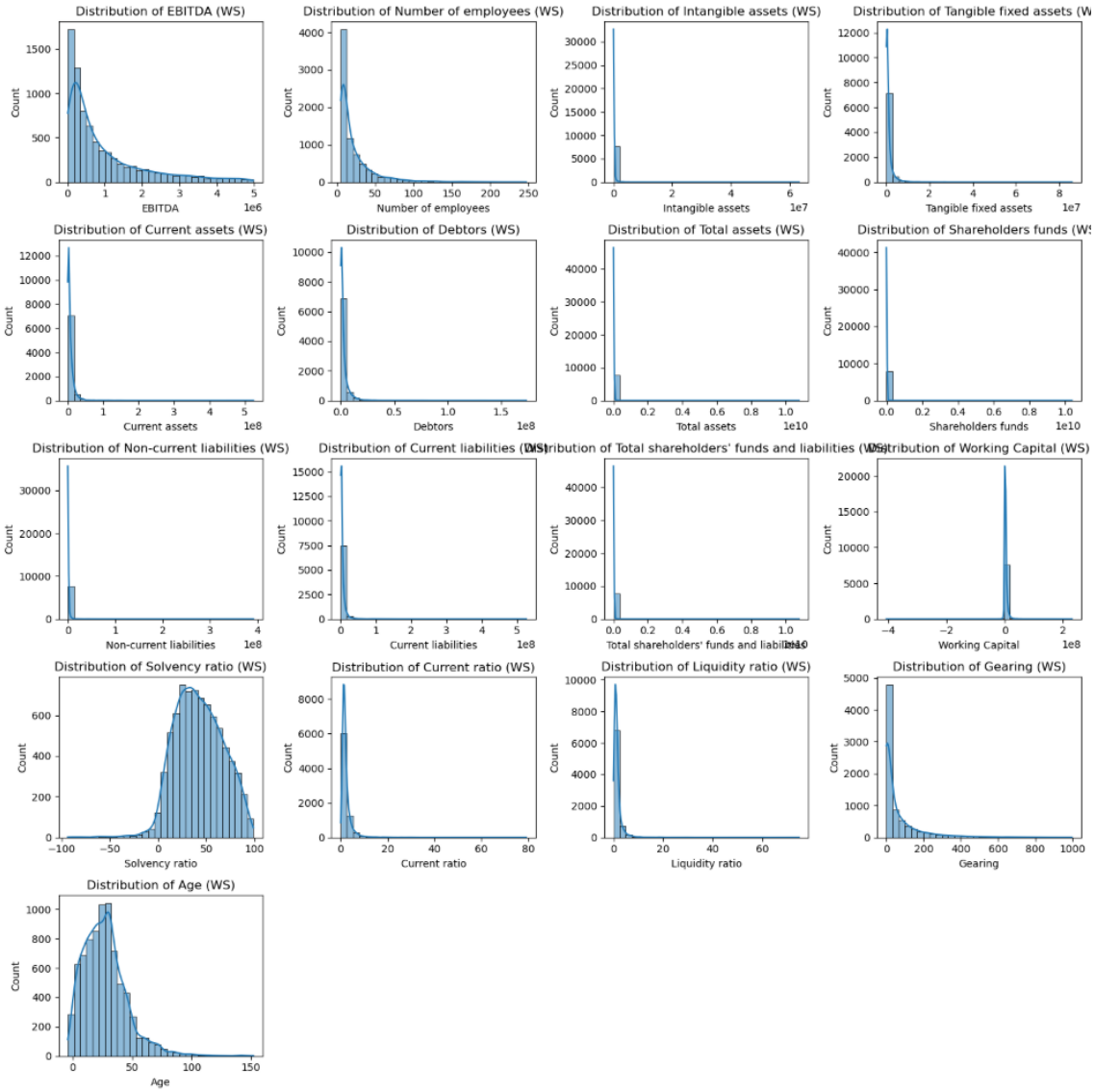


FIGURE A.2: Baseline scenario: Distribution of each feature for the wholesale sector

A.10 Distribution of variables: Construction

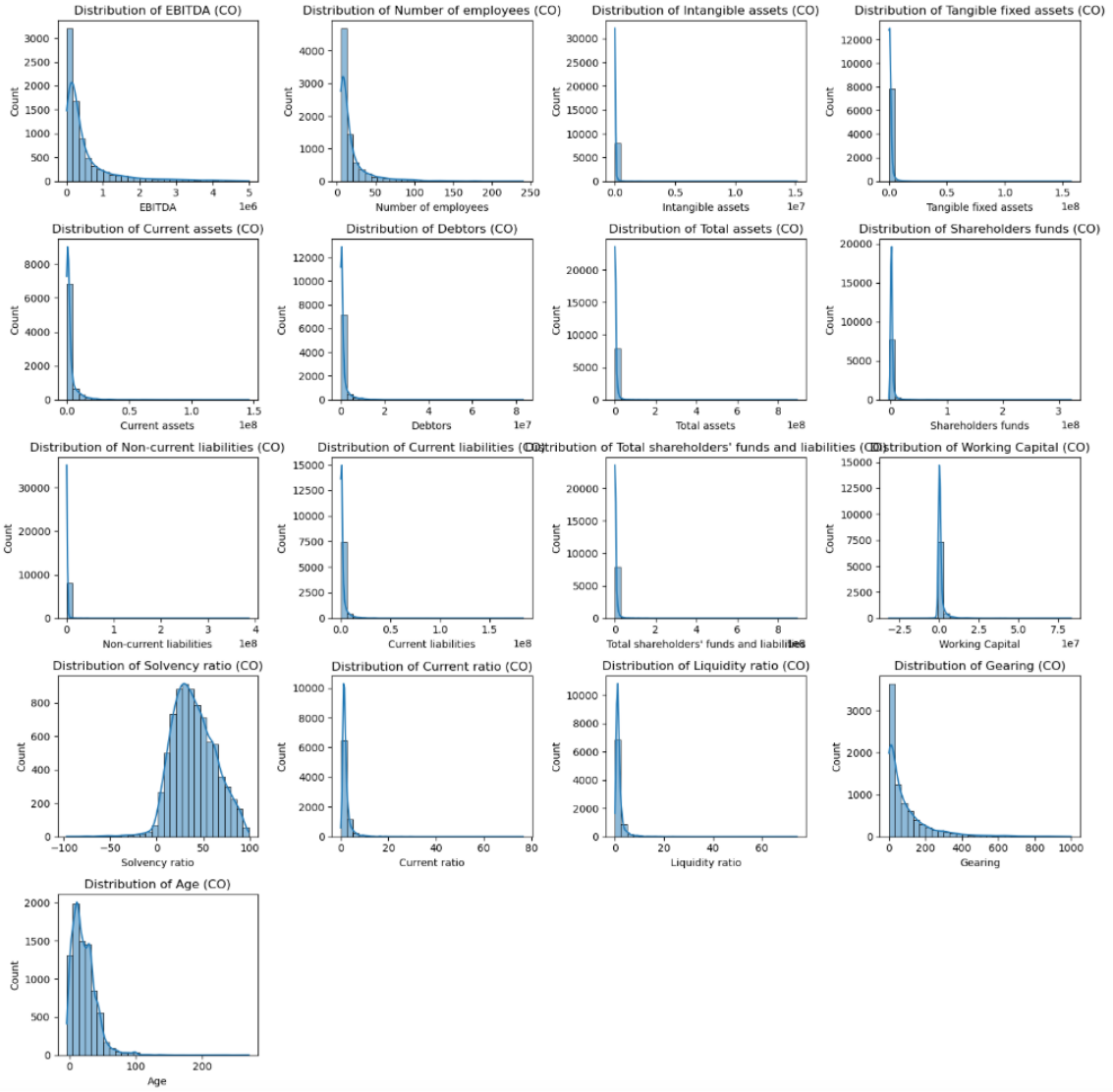


FIGURE A.3: Baseline scenario: Distribution of each feature for the construction sector

A.11 Backtesting of Pearson and Spearman correlations between input variables and EBITDA

Variable	Business services		Wholesale		Construction	
	Pearson Δ	Spearman Δ	Pearson Δ	Spearman Δ	Pearson Δ	Spearman Δ
Number of employees	+0.016	-0.002	-0.007	-0.007	+0.001	+0.024
Intangible assets	-0.090	-0.014	+0.051	+0.014	+0.076	+0.006
Tangible fixed assets	+0.024	+0.025	+0.008	+0.004	+0.072	+0.034
Current assets	-0.125	+0.015	-0.092	-0.016	+0.004	-0.008
Debtors	+0.006	-0.013	+0.031	-0.004	-0.063	-0.007
Total assets	-0.131	+0.024	+0.113	-0.010	+0.046	+0.005
Shareholders' funds	-0.096	+0.014	+0.091	-0.006	-0.111	-0.013
Non-current liabilities	-0.016	+0.043	+0.012	+0.037	+0.115	+0.057
Current liabilities	-0.097	+0.024	-0.076	-0.016	+0.050	+0.004
Total shareholders' funds	-0.131	+0.024	+0.113	-0.010	+0.046	+0.005
Working capital	-0.129	-0.059	+0.067	-0.038	-0.032	-0.020
Solvency ratio	-0.022	-0.020	+0.021	+0.023	-0.014	-0.016
Current ratio	-0.020	-0.031	+0.012	+0.013	-0.026	-0.027
Liquidity ratio	-0.001	+0.009	+0.025	+0.047	-0.012	-0.001
Gearing	-0.009	-0.025	+0.014	+0.002	+0.020	+0.037
Age	+0.013	-0.002	-0.026	-0.031	-0.002	+0.010

TABLE A.8: Backtesting results: Pearson and Spearman correlation differences between baseline and 2019 scenarios

A.12 The effect of Winsorizing: Correlation between input variables and EBITDA

Variable	Business services		Wholesale		Construction	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
Number of employees	0.522	0.604	0.496	0.569	0.579	0.523
Intangible assets	0.285	0.288	0.203	0.174	0.189	0.175
Tangible fixed assets	0.491	0.604	0.393	0.426	0.460	0.565
Current assets	0.524	0.761	0.597	0.756	0.692	0.750
Debtors	0.505	0.694	0.523	0.665	0.624	0.613
Total assets	0.510	0.788	0.570	0.780	0.647	0.796
Shareholders' funds	0.476	0.698	0.554	0.729	0.622	0.746
Non-current liabilities	0.374	0.440	0.260	0.166	0.374	0.365
Current liabilities	0.408	0.712	0.499	0.669	0.621	0.674
Working capital	0.473	0.493	0.515	0.576	0.569	0.510
Solvency ratio	0.189	0.190	0.132	0.159	0.089	0.139
Current ratio	0.088	0.120	0.061	0.127	0.053	0.114
Liquidity ratio	-0.024	-0.085	0.037	0.088	0.046	0.098
Gearing	-0.228	-0.337	-0.159	-0.196	-0.119	-0.110
Age	0.071	0.064	0.155	0.167	0.290	0.212

TABLE A.9: The effect of Winsorizing: Pearson and Spearman correlation coefficients between EBITDA and each feature across business services, wholesale, and construction sectors

A.13 Pearson correlations after log transformations

Variable	Business services	Wholesale	Construction
	Pearson Log (Δ)		
Number of employees	0.600 (+0.081)	0.555 (+0.074)	0.569 (+0.006)
Intangible assets	0.247 (+0.048)	0.175 (+0.059)	0.186 (+0.091)
Tangible fixed assets	0.574 (+0.166)	0.411 (+0.089)	0.544 (+0.254)
Current assets	0.771 (+0.633)	0.742 (+0.304)	0.755 (+0.167)
Debtors	0.676 (+0.558)	0.606 (+0.229)	0.633 (+0.103)
Total assets	0.784 (+0.633)	0.762 (+0.683)	0.791 (+0.392)
Shareholders' funds	0.705 (+0.593)	0.718 (+0.670)	0.737 (+0.379)
Non-current liabilities	0.274 (+0.157)	0.067 (-0.050)	0.193 (-0.007)
Current liabilities	0.720 (+0.615)	0.653 (+0.345)	0.686 (+0.214)
Working capital	0.684 (+0.402)	0.621 (+0.357)	0.598 (+0.211)
Solvency ratio	0.189 (+0.004)	0.153 (+0.020)	0.108 (+0.017)
Current ratio	0.129 (+0.074)	0.091 (+0.042)	0.057 (+0.016)
Liquidity ratio	-0.042 (-0.027)	0.075 (+0.045)	0.052 (+0.016)
Gearing	-0.420 (-0.200)	-0.233 (-0.079)	-0.154 (-0.041)
Age	0.085 (+0.014)	0.135 (-0.024)	0.196 (-0.095)

TABLE A.10: Pearson correlation between EBITDA and each Feature (Log Transformed) across Sectors, with Delta to Baseline Pearson Values

A.14 Pearson correlation after square root-transformations

Variable	Business services	Wholesale	Construction
	Pearson Sqrt (Δ)		
Number of employees	0.592 (+0.073)	0.558 (+0.077)	0.614 (+0.051)
Intangible assets	0.311 (+0.112)	0.234 (+0.118)	0.223 (+0.128)
Tangible fixed assets	0.591 (+0.183)	0.422 (+0.100)	0.533 (+0.243)
Current assets	0.704 (+0.566)	0.709 (+0.271)	0.764 (+0.176)
Debtors	0.650 (+0.532)	0.632 (+0.255)	0.688 (+0.158)
Total assets	0.694 (+0.543)	0.708 (+0.629)	0.763 (+0.364)
Shareholders' funds	0.672 (+0.560)	0.690 (+0.642)	0.739 (+0.381)
Non-current liabilities	0.426 (+0.309)	0.240 (+0.123)	0.399 (+0.199)
Current liabilities	0.605 (+0.500)	0.609 (+0.301)	0.697 (+0.225)
Working capital	0.689 (+0.407)	0.641 (+0.377)	0.670 (+0.283)
Solvency ratio	0.220 (+0.035)	0.143 (+0.010)	0.090 (-0.001)
Current ratio	0.135 (+0.080)	0.079 (+0.030)	0.049 (+0.008)
Liquidity ratio	-0.032 (-0.017)	0.063 (+0.033)	0.043 (+0.007)
Gearing	-0.369 (-0.149)	-0.232 (-0.078)	-0.178 (-0.065)
Age	0.094 (+0.023)	0.141 (-0.018)	0.243 (-0.048)

TABLE A.11: Pearson correlation between EBITDA and each feature (square root-transformed) across sectors, with Delta to baseline Pearson values

A.15 Confusion matrices of the best performing classification tests per sector

A.15.1 Business services

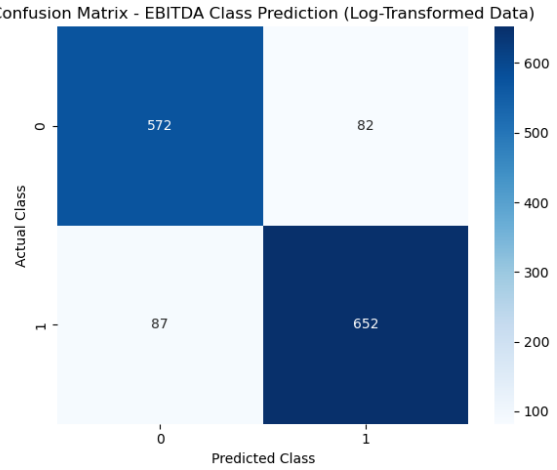
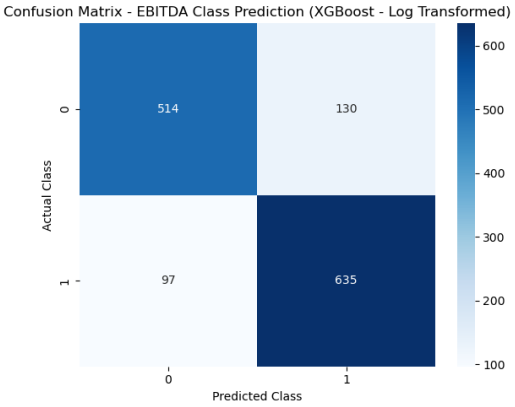
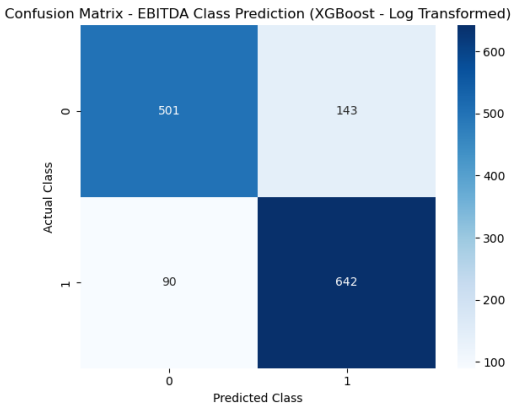


FIGURE A.4: EBITDA classification in the business services sector: Confusion matrix of test 3

A.15.2 Wholesale



(A) Confusion matrix test 7



(B) Confusion matrix test 8

FIGURE A.5: EBITDA classification in the wholesale sector: Confusion matrix of test 7 and 8

A.15.3 Construction

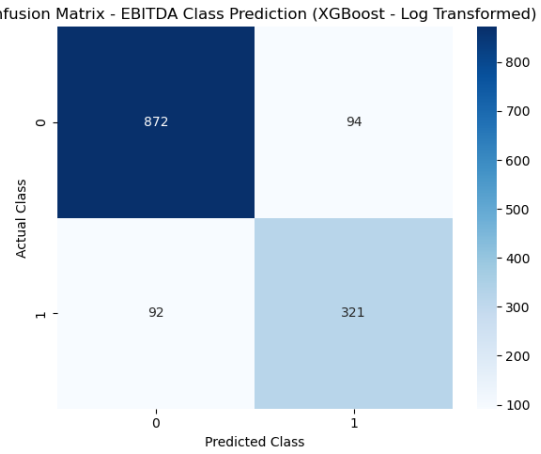
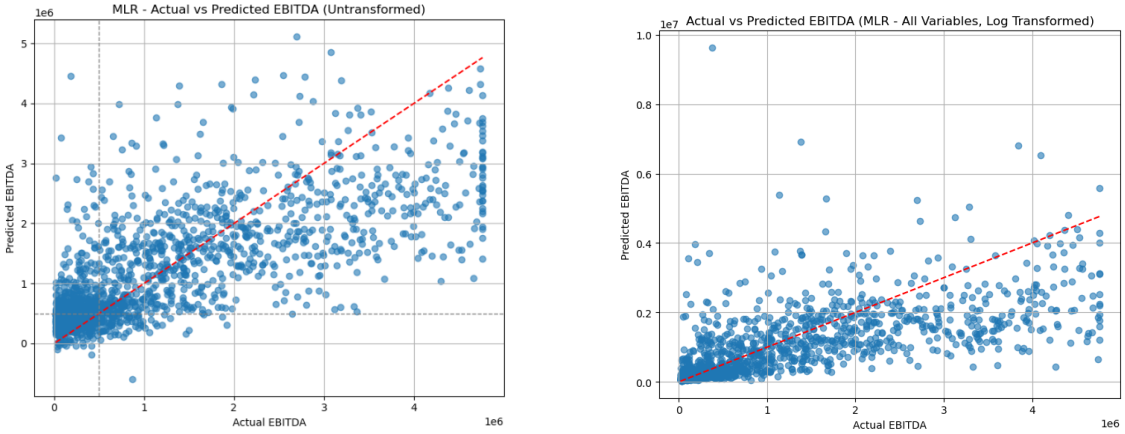


FIGURE A.6: EBITDA classification in the construction sector: Confusion matrix of test 5

A.16 MLR scatter plots actual vs predicted EBITDA values of the best performing tests per sector

A.16.1 Business services

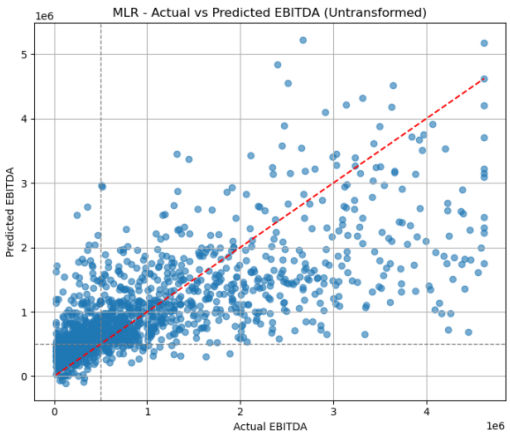


(A) MLR on normal data

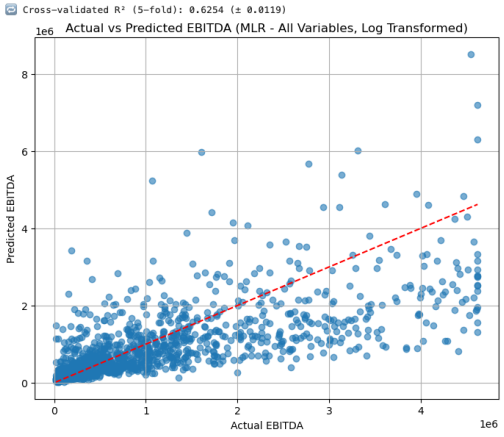
(B) MLR on log data

FIGURE A.7: Scatter plot of tests 1 and 3, actual vs predicted outcomes MLR for the business services sector

A.16.2 Wholesale



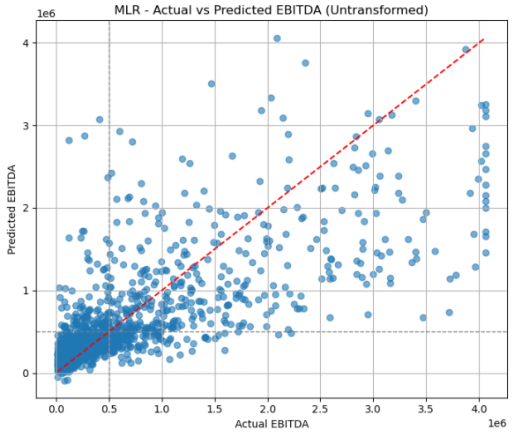
(A) MLR on normal data



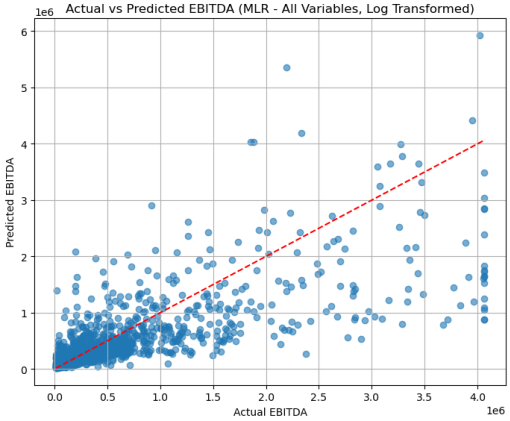
(B) MLR on log data

FIGURE A.8: Scatter plot of tests 1 and 3, actual vs predicted outcomes MLR for the wholesale sector

A.16.3 Construction



(A) MLR on normal data



(B) MLR on log data

FIGURE A.9: Scatter plot of tests 1 and 3, actual vs predicted outcomes MLR for the construction sector

A.17 Feature coefficients of MLR predictions including all variables

Feature	Business services		Wholesale		Construction	
	Normal	Log	Normal	Log	Normal	Log
β_0 (Intercept)	318101.98	2.440	175280.42	2.380	59601.69	1.050
1. Number of employees	4310.08	0.004	6405.90	0.076	4281.12	0.054
2. Intangible assets	0.116	0.008	0.178	0.002	0.188	0.004
3. Tangible fixed assets	0.099	0.153	0.136	0.134	0.103	0.204
4. Current assets	0.028	0.220	0.029	0.248	0.065	0.474
5. Debtors	0.045	0.020	0.024	-0.011	0.067	-0.107
6. Total assets	-0.024	0.081	-0.036	-0.114	-0.053	0.091
7. Shareholders funds	0.024	0.105	0.062	0.433	0.071	0.202
8. Non-current liabilities	0.013	0.006	0.022	0.006	0.047	0.012
9. Current liabilities	0.016	0.121	0.038	0.075	0.027	-0.031
10. Working capital	0.047	0.088	0.023	0.075	0.035	0.060
11. Solvency ratio	4635.47	-0.002	3781.25	-0.179	3158.89	-0.009
12. Current ratio	136649.18	-0.491	-6078.75	-0.602	-78176.81	-1.262
13. Liquidity ratio	-181498.71	0.445	-12260.96	0.515	64879.91	1.052
14. Gearing	-518.30	-0.076	-624.58	-0.069	-139.54	-0.079
15. Age	-3817.13	-0.087	-168.03	-0.113	-958.88	-0.146

TABLE A.12: MLR coefficients per feature across business services, wholesale, and construction sectors in normal and log-transformed space.

Example formula:

$$\begin{aligned}\hat{y}_{\text{EBITDA}} = {} & 318{,}101.98 + 4{,}310.08 \cdot \text{Employees} + 0.116 \cdot \text{IntangibleAssets} + 0.099 \cdot \text{TangibleAssets} \\ & + 0.028 \cdot \text{CurrentAssets} + 0.045 \cdot \text{Debtors} - 0.024 \cdot \text{TotalAssets} + 0.024 \cdot \text{ShareholdersFunds} \\ & + 0.013 \cdot \text{NonCurrentLiabilities} + 0.016 \cdot \text{CurrentLiabilities} + 0.047 \cdot \text{WorkingCapital} \\ & + 4{,}635.47 \cdot \text{SolvencyRatio} + 136{,}649.18 \cdot \text{CurrentRatio} - 181{,}498.71 \cdot \text{LiquidityRatio} \\ & - 518.30 \cdot \text{Gearing} - 3{,}817.13 \cdot \text{Age} + \varepsilon\end{aligned} \tag{A.1}$$

A.18 Feature coefficients of MLR predictions including top-5 variables

Feature	Business services		Wholesale		Construction	
	Normal	Log	Normal	Log	Normal	Log
β_0 (Intercept)	337296.06	1.570	433152.31	1.010	187741.20	0.500
1. Current assets	26376	0.077	32338	-0.006	28007	-0.078
2. Total assets	3882	0.352	-2862	0.338	-14362	0.562
3. Tangible fixed assets	86808	–	–	–	–	–
4. Working capital	72279	0.103	25794	–	–	–
5. Number of employees	6400	–	–	–	–	–
6. Shareholders funds	–	0.126	52174	0.308	89914	0.254
7. Current liabilities	–	0.136	–	0.130	32022	0.068
8. Debtors	–	–	29810	0.053	109428	0.046

TABLE A.13: Top 5 MLR coefficients per feature across business services, wholesale, and construction sectors in normal and log-transformed space.

Example formula:

$$\begin{aligned}\hat{y}_{\text{EBITDA}} = 337{,}296.06 + 6{,}400.58 \cdot \text{Employees} + 0.087 \cdot \text{TangibleAssets} \\ + 0.072 \cdot \text{WorkingCapital} + 0.026 \cdot \text{CurrentAssets} + 0.004 \cdot \text{TotalAssets} + \varepsilon\end{aligned} \tag{A.2}$$

A.19 Correlation heatmaps of input features per sector

A.19.1 Business services

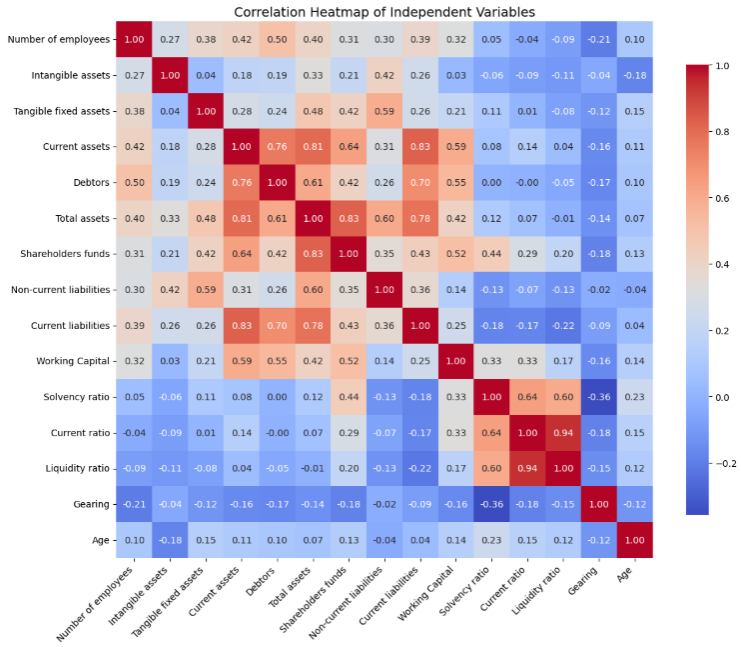


FIGURE A.10: Correlation heatmap of input features for business services sector

A.19.2 Wholesale

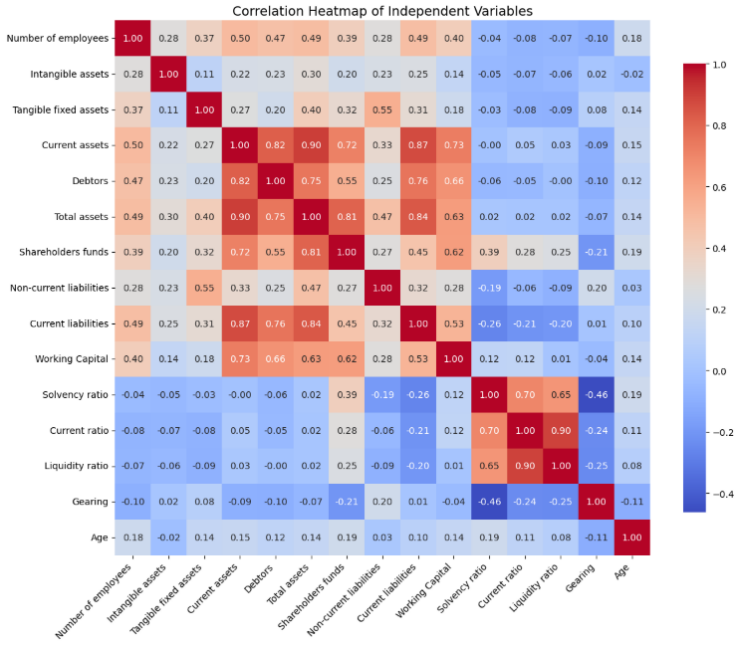


FIGURE A.11: Correlation heatmap of input features for wholesale sector

A.19.3 Construction

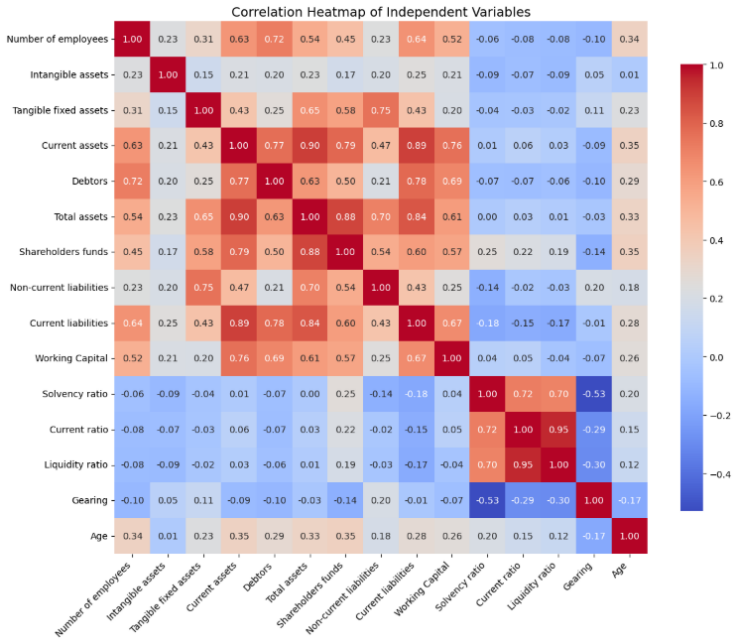


FIGURE A.12: Correlation heatmap of input features for construction sector

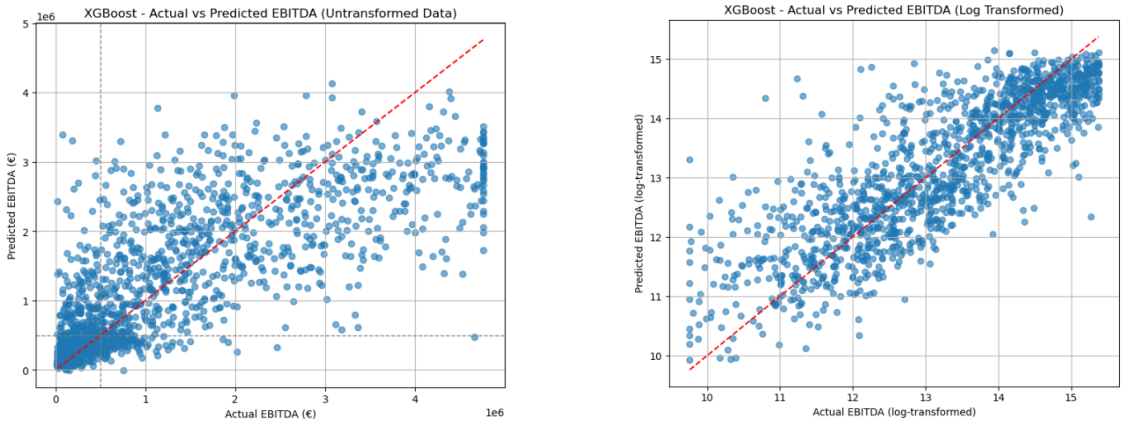
A.20 Configurations for MLR with the best performances across sectors

Sector	Variable type	Feature set	R^2	Adj. R^2	SMAPE	Accuracy	Critical %	Imprecision %
Business services	Log	All	0.65	0.65	56.51%	0.8485	8.83%	6.32%
Business services	Normal	No ratios	0.55	0.54	71.34%	0.7794	4.48%	17.58%
Wholesale	Log	All	0.65	0.64	53.24%	0.8322	9.08%	7.70%
Wholesale	Normal	No ratios	0.53	0.53	64.71%	0.7868	2.69%	18.63%
Construction	Log	All	0.66	0.65	51.22%	0.8535	9.50%	5.15%
Construction	Normal	All	0.58	0.57	60.51%	0.8575	7.43%	6.82%

TABLE A.14: MLR Performance comparison across sectors, variable types, and feature sets

A.21 XGBoost scatter plots actual vs predicted EBITDA values of the best performing tests per sector

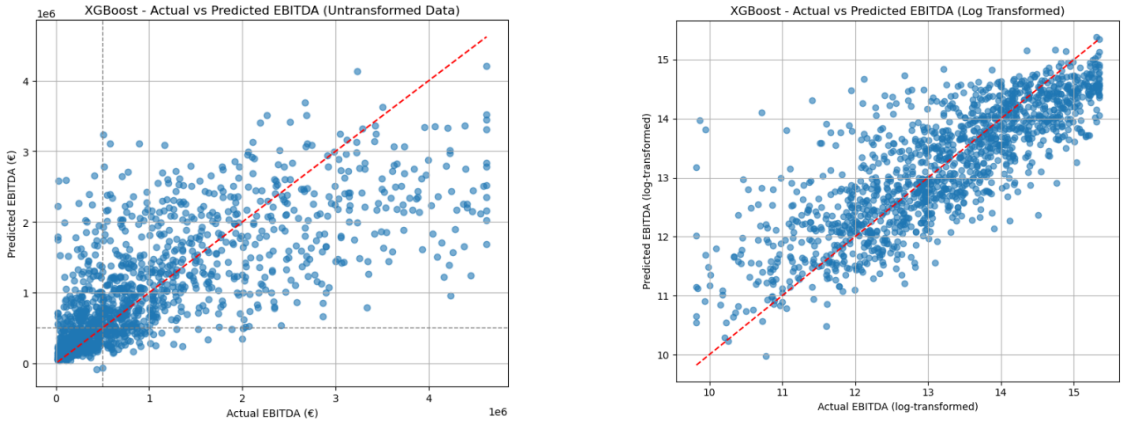
A.21.1 Business services



(A) XGBoost on normal data (B) XGBoost on log data

FIGURE A.13: XGBoost scatter plots actual vs predicted EBITDA values for business services

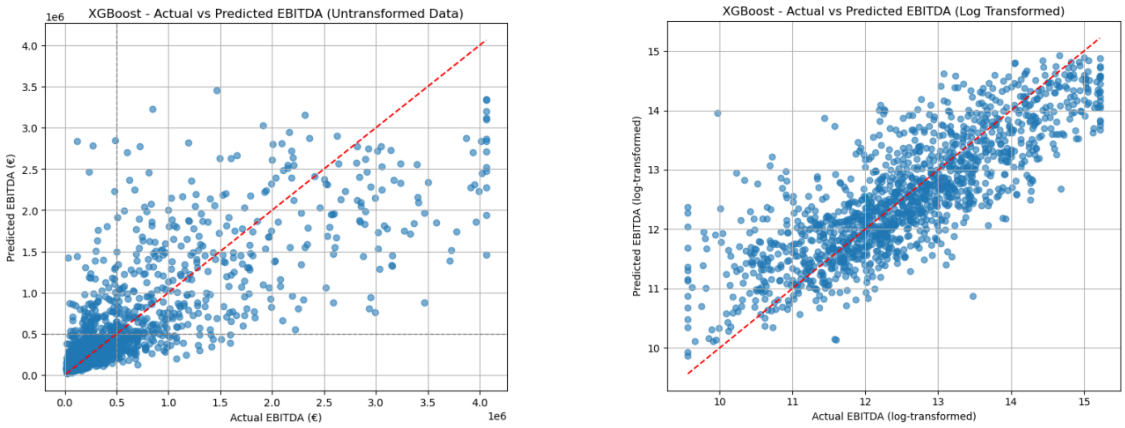
A.21.2 Wholesale



(A) XGBoost on normal data (B) XGBoost on log data

FIGURE A.14: XGBoost scatter plots actual vs predicted EBITDA values for wholesale

A.21.3 Construction



(A) XGBoost on normal data

(B) XGBoost on log data

FIGURE A.15: XGBoost scatter plots actual vs predicted EBITDA values for construction

A.22 DuPont analysis

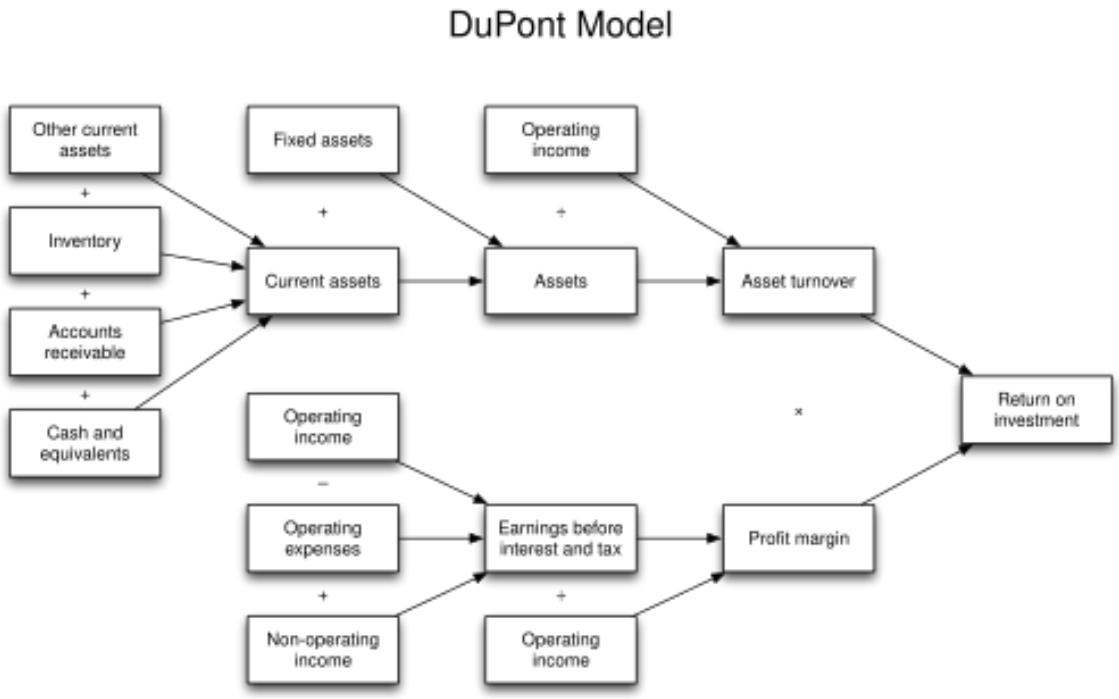


FIGURE A.16: Overview of a DuPont analysis (Wikipedia contributors, 2024)

A.23 Differences between mean and standard deviation of the critical errors versus the correct predicted values

A.23.1 Business services

Feature	μ Critical	μ Correct	μ Diff.	σ Critical	σ Correct	σ Diff.
Number of employees	2.68011	3.26480	0.58469	0.75558	1.08887	0.33330
Intangible assets	5.81299	5.18266	0.63033	5.64491	5.88453	0.23962
Tangible fixed assets	11.86378	12.74321	0.87943	1.67771	2.27678	0.59906
Current assets	14.31231	14.79006	0.47775	0.67616	1.47304	0.79688
Debtors	13.12794	13.76339	0.63544	0.98333	1.70751	0.72418
Total assets	14.71284	15.23596	0.52312	0.84828	1.49624	0.64796
Shareholders funds	13.18206	14.19117	1.00911	1.88062	1.88914	0.00851
Non-current liabilities	8.30576	9.60253	1.29677	5.97587	5.85201	0.12385
Current liabilities	13.85947	14.23636	0.37689	0.92681	1.45440	0.52759
Working Capital	12.52233	13.47728	0.95495	1.29413	1.85580	0.56166
Solvency ratio	3.24488	3.61818	0.37330	1.08047	0.73714	0.34333
Current ratio	0.99035	1.04839	0.05804	0.43317	0.40403	0.02915
Liquidity ratio	0.96016	0.92244	0.03772	0.43936	0.41655	0.02282
Gearing	2.73943	1.90848	0.83095	2.08236	2.02791	0.05445
Age	2.44014	2.56559	0.12545	0.82697	0.95146	0.12449
EBITDA	13.54999	13.19995	0.35004	0.41837	1.48477	1.06640

TABLE A.15: Comparison of feature means and standard deviations between critical and correctly predicted groups in log scale, for business services sector.

A.23.2 Wholesale

Feature	μ Critical	μ Correct	μ Diff.	σ Critical	σ Correct	σ Diff.
Number of employees	2.48186	2.86274	0.38088	0.46746	0.86030	0.39284
Intangible assets	4.39675	4.13578	0.26098	5.24118	5.34635	0.10517
Tangible fixed assets	12.20469	12.78143	0.57674	1.58787	1.74927	0.16141
Current assets	14.79813	15.15342	0.35529	0.50610	1.30660	0.80050
Debtors	13.30808	13.78490	0.47683	1.53078	1.80404	0.27326
Total assets	15.01905	15.41653	0.39748	0.42785	1.23402	0.80618
Shareholders funds	13.93881	14.44095	0.50214	0.70362	1.44865	0.74503
Non-current liabilities	8.51360	9.37003	0.85643	5.92185	5.63449	0.28736
Current liabilities	14.11502	14.47754	0.36252	0.79263	1.36303	0.57040
Working Capital	13.65387	14.13704	0.48316	0.99340	1.57285	0.57945
Solvency ratio	3.56996	3.67309	0.10313	0.74543	0.64163	0.10380
Current ratio	1.12372	1.12685	0.00313	0.42149	0.43175	0.01026
Liquidity ratio	0.89684	0.82295	0.07389	0.42967	0.46183	0.03216
Gearing	2.79349	2.58291	0.21058	2.06794	2.02709	0.04085
Age	3.06493	3.12645	0.06153	0.92737	0.75694	0.17044
EBITDA	13.56563	13.22978	0.33585	0.34720	1.34749	1.00029

TABLE A.16: Comparison of feature means and standard deviations between critical and correctly predicted groups in log scale, for wholesale sector.

A.23.3 Construction

Feature	μ Critical	μ Correct	μ Diff.	σ Critical	σ Correct	σ Diff.
Number of employees	2.71194	2.65660	0.05534	0.64460	0.79967	0.15507
Intangible assets	3.24873	2.86519	0.38354	4.67256	4.44468	0.22787
Tangible fixed assets	12.58873	12.60999	0.02126	1.36168	1.53623	0.17455
Current assets	14.44426	14.07056	0.37370	0.64578	1.24448	0.59870
Debtors	13.22945	13.00888	0.22058	1.09729	1.43549	0.33820
Total assets	14.76989	14.47155	0.29835	0.50029	1.16653	0.66625
Shareholders funds	13.72422	13.41261	0.31161	0.71459	1.35784	0.64325
Non-current liabilities	9.63428	10.38907	0.75479	5.23540	4.53477	0.70063
Current liabilities	13.86351	13.49943	0.36408	0.86349	1.32212	0.45863
Working Capital	13.29026	12.82688	0.46338	1.08514	1.55079	0.46565
Solvency ratio	3.59575	3.59152	0.00424	0.62506	0.64420	0.01913
Current ratio	1.06395	1.05151	0.01244	0.40034	0.38953	0.01081
Liquidity ratio	0.91414	0.91145	0.00269	0.42032	0.41202	0.00831
Gearing	3.03861	3.38504	0.34643	1.92224	1.84389	0.07835
Age	3.02909	2.87749	0.15160	0.82238	0.90126	0.07888
EBITDA	13.50947	12.43323	1.07623	0.34127	1.30593	0.96466

TABLE A.17: Comparison of feature means and standard deviations between critical and correctly predicted groups in log scale, for construction sector.

A.24 Cross-validation of best performing models

Model	Sector	# Test	Variable type	R^2		Bias		SMAPE	
				Mean (μ)	St. Dev. (σ)	Mean (μ)	St. Dev. (σ)	Mean (μ)	St. Dev. (σ)
MLR	Construction	3	Log	0.6803	0.0169	-0.0004	0.0178	48.84%	1.21%
RF	Business services	2	Log	0.7080	0.0161	-0.0161	0.0358	4.51%	0.08%
XGBoost	Business services	4	Log	0.7130	0.0156	-0.0012	0.0261	4.45%	0.08%

TABLE A.18: Cross-validation performance per model: mean (μ) and standard deviation (σ) of R^2, bias, and SMAPE across folds