Cover Page – MSc Business Analytics Consultancy Project/Dissertation 2022-23

Candidate #: BSHG7

Title of Project:

Relative Valuation with Gradient Boosting Machines and FinBERT: *The London Stock Exchange Case Study*

Date: 7th August 2023

Word Count: 13153

Disclaimer:

I hereby declare that this dissertation is my individual work and to the best of my knowledge and confidence, it has not already been accepted in substance for the award of any other degree and is not concurrently submitted in candidature for any degree. It is the end product of my own independent study except where other acknowledgement has been stated in the text.

MSc Business Analytics Project/Dissertation 2022-23
Marking Sheet

Criteria/Weight	Supervisor's comments
Topic, theoretical framework, literature, and methodology (35%): Topic is clearly identified and boundaries are asserted. Knowledge of relevant theories and their limitations. Current and relevant literature coming from reliable sources. Appropriate and adequate methodology for topics. Detailed methodology facilitating replication of project and reproducibility of results.	
Analysis and conclusions /recommendations (35%): Use of primary and/or secondary data. Rigorous analysis and interpretations. Alternative interpretations/arguments are considered. Limitations are identified and justified by reasonable arguments. Conclusions/recommendations are fully consistent with evidence presented.	
Structure, originality and presentation (10%): Provides a concise summary. Demonstrates an understanding of business context. Coherent and appropriate structure. Adequate presentation, language, style, graphs, tables, and referencing. Appropriate use of visualisation. Presents business recommendations.	
Complexity of project scope and progress made towards business goals (10%): Progress made towards overcoming technical and operational challenges encountered during the project. Progress made in overcoming problem framing and theoretical and data related problems encountered during the project.	
Project Management (10%): Good use of project management and communication tools. Evidence of objectives being broken down in appropriate tasks and timely engagement with primary supervisor.	

General marking guidelines

85+ Outstanding work of publishable standard.
70-84 Excellent work showing mastery of the subject matter and excellent analytical skills.
60-69 Very good work. Interesting analysis with original insights. Some minor errors.
50-59 Good work which only covers a basic analysis. Some problems but no major omissions.
40-49 Inadequate work. Not sufficiently analytical. Some major omissions.
39- Work seriously flawed. Lack of clarity and argumentation. Too descriptive.

Mark: _____

Acknowledgements

I extend heartfelt thanks to my supervisor, whose guidance and expertise were valuable in this research. My gratitude also goes to the LSEG team. Their insights and collaboration elevated this project and pushed me to produce real-world results. Their contributions further fuelled my enthusiasm for finance and data analysis.

To my dear friend Benedetta, who has been with me since the start of my academic journey, your unwavering support and invaluable feedback were central, every shared moment is cherished. Additionally, I'm thankful for all the beautiful connections I made at UCL, which enriched my experience and broadened my horizons.

Finally, to my family, your enduring care, encouragement, and support have been my foundation. Your faith in me has been my anchor.

Table of Contents

List of Tables and Figures

Abstract

In an era where data-driven decision-making is increasingly pivotal in financial markets and strategic planning, machine learning (ML) techniques have emerged as an indispensable tool for enhancing traditional financial valuation methods. This paper posits that supervised machine learning models, particularly Gradient Boosting Machines, can offer valuable insights for predicting and analysing forward financial valuation metrics such as Enterprise Value over Earning Before Interest, Tax, Depreciation, and Amortisation (EV/EBITDA). By leveraging diversified data sources, ranging from company filings to media announcements, this research exploits quantitative and qualitative metrics. Utilising FinBERT, a pre-trained Natural Language Processing (NLP) model, the paper aims to integrate sentiment and classification features into financial forecasting models. The findings highlight the forecasting efficiency of GBMs, demonstrated across both the broader market— represented by the S&P 500—and individual firms whose valuation is inherently complex, such as the London Stock Exchange Group (LSEG). This dissertation focuses on the application of LightGBM, with categorised data on firms undergoing a business transformation. This approach highlights the capacity of machine learning to provide precise, future-oriented financial analysis useful for financial institutions and decision-makers.

1. Introduction

1.1 Research question

Financial valuation is a complex process that requires a deep understanding of market dynamics and shareholder perception. In this dynamic environment, traditional financial valuation models, such as discounted cash flow (DCF) analysis, multiples, and comparable company analysis, often fall short of providing a comprehensive and accurate depiction of a company's intrinsic value due to their limitations in flexibility and inability to capture different factors. (Goodell, 2021)

In recent years, the rise of machine learning (ML) has presented new opportunities to improve the accuracy and efficiency of financial valuation. ML is an artificial intelligence that allows computers to learn from data without being explicitly programmed. (Milner, 1997) These models offer enhanced flexibility and the ability to incorporate many features beyond what traditional models can capture (Flood M.D., 2016). ML algorithms can leverage large volumes of structured and unstructured data, including market trends, sentiment analysis, and other aspects critical to understanding a company's value dynamics. This flexibility allows models to consider multiple dimensions simultaneously and capture complex interactions that are difficult to quantify using traditional approaches like DCF multiples and relative valuation field (Dixon M.F., 2020). Further strategic considerations can be made by analysing the models' parameters. For instance, feature importance can provide a better understanding of the drivers of financial valuation. The demand is therefore rising for models that can carry out financial valuation tasks, where the data constantly changes, and new insights are constantly generated. (Holzinger A., 2018)

This research paper investigates the application of GBMs to bridge today's gap in financial valuation practices, leveraging the potential of these flexible models and diverse data sources by extracting features with FinBERT. More specifically, upon establishing LightGBM's predictive power on the broader market, we will examine the London Stock Exchange Group (LSEG) case, assessing how GBMs can contribute to a more comprehensive understanding of market dynamics and shareholder perception.

LSEG, as a former provider in capital markets and post-trade industry, is an ideal case study because it has recently acquired Refinitiv, a leading data and analytics provider. Through this acquisition, LSEG's perceived value has transformed, shifting from being primarily an exchange provider to emerging as a competitive data player. LSEG's value proposition has transitioned its capital structure

from primarily relying on transactional revenues to recurrent revenue model accounting for approximately 70% of its total revenues. This represents a significant change, as this business restructuring strategy impacts the intrinsic valuation of the company (LSEG, 2023). This core repositioning in LSEG's business model presents a complex challenge for traditional valuation methods, which may struggle to capture its complex dynamics.

1.2 Significance of the Study

This research seeks to provide insights into how predictive analytics can enhance the valuation process in the financial industry. In doing so, it contributes to bridging the gap between the constant evolution typical of the financial industry and traditional valuation practices. As argued above, standard techniques are no longer sufficient to capture the full extent and nuances of the data available to financial institutions and enterprises. While they do provide significant information, this paper argues that they should be combined with ML valuation to provide a more complete picture. Lastly, LSEG serves as a compelling case study. It demonstrates the practical implications of ML-driven valuation and its potential to enhance decision-making in the financial industry.

LSEG stands to benefit from the application of ML techniques in multiple aspects of its operations. Firstly, ML can greatly assist in corporate strategy, enabling more effective planning and decision-making processes. By leveraging ML algorithms, LSEG can gain valuable insights into market trends, competitor analysis, and investor behaviour, thereby enhancing its strategic positioning in the industry. Furthermore, ML can prove instrumental for the Investor Relations teams at LSEG. It can serve as the voice of the company in the market, facilitating effective communication and helping to shape the narrative around LSEG's current and future prospects. This attracts potential investors and strengthens relationships with existing stakeholders. ML-driven analysis and data-driven storytelling can significantly enhance LSEG's brand perception compared to its peers, presenting it as a forward-thinking and technologically advanced company.

Finally, this dissertation strives to highlight the potential of a closer and more dynamic relationship between academic research and industry. An increased collaboration between these two fields, in fact, can yield mutually beneficial outcomes for both. By fostering and harnessing research findings, the industry can drive innovation and enhance its operational efficiency. Conversely, real-world challenges posed by industry can steer academic research towards addressing practical problems, hence contributing to the advancement of both domains.

1.3 Structure of the dissertation

This paper begins with a literature review that considers the limitations and advantages of traditional methods in capturing the full scope of market dynamics. It analyses the transition in valuation techniques from discounted cash flow to the relative valuation method. Then, it discusses in detail the approaches to the latter and its benefits, along with its relative shortcomings and how ML academics are tackling those limitations. Lastly, it explores the choice of Gradient Boosting Machines and how the model's properties best complement the current challenges.

Considering the literature, the paper follows Microsoft Team Data Science Process (TDSP) methodology developed by Microsoft (Aparicio, 2020). It begins by evaluating the business framework, in our case, for the financial valuation of LSEG and its relative peers. Then, it considers the data acquisition and categorisation of textual data through FinBERT, followed by data understanding and GBM modelling. Subsequently, the paper delves into the specific applications of GBMs and analyses their potential in reshaping current valuation techniques. Leveraging a vast financial database of the Standard and Poors' 500 companies, this paper tests and benchmarks several models. The algorithms tested are different architectures of gradient boosting. These have been chosen because they are the most effective learning models in the literature on financial valuation techniques.

By examining the performance of the supervised learning models against these traditional approaches and expert knowledge, the research seeks to identify which approach yields better results in capturing the underlying dynamics and providing more accurate predictions. Having established LightGBM can determine the most accurate valuation, the dissertation discusses the implications of these findings in the context of the case study. This paper applies the valuation model to LSEG and the identified peers, in order to ascertain their performance during the transition and how it overall positions itself on the market.

In its final section, this dissertation explores potential future applications of ML techniques in the broader context of financial valuation beyond the case of LSEG. The aim is to investigate how ML techniques can be applied to companies operating in ever-changing environments, undergoing business model transitions, or experiencing significant mergers and acquisitions (M&A) activity. While this case study presents some limitations, in fact, it still opens the door for valuable further research and discussion in the field.

2. Literature review

This dissertation extends the ongoing discussion on financial valuation evolution, including traditional techniques and innovative ML approaches. Traditional methods such as discounted cash flow (DCF) and relative valuation often fail to capture the complex dynamics of today's financial markets. Concurrently, ML techniques are gaining traction in financial valuation, although with disagreements on their optimal application. This includes debates between advocates of unsupervised models [(Dameri, 2020), (Qian, 2006)] and proponents of supervised ones [(Özlem, 2022), (X. Yang, 2019), (Hüseyin, 2017)]. This gap presents challenges in accurately assessing a company's intrinsic value, emphasising the need for enhanced financial valuation methods crucial for researchers, analysts, and financial decision-makers. In this section, we acknowledge the extensive research and theories related to financial valuation, however focusing specifically on the area of relative valuation. The choice to concentrate on this particular aspect is strategic and intentional. As argued by Damodaran many valuation models could be classified as an extension of relative valuation (Damodaran, 2012).

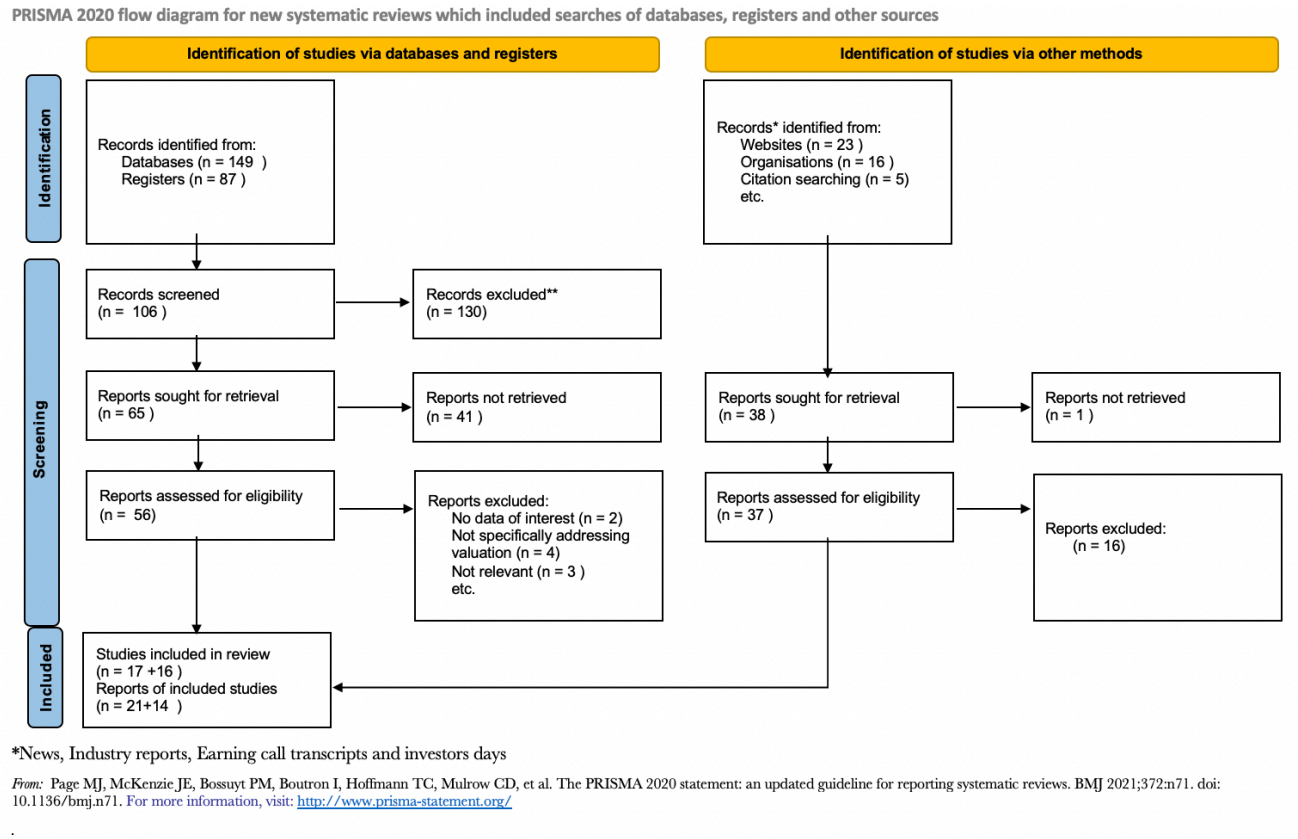
This chapter first states the theoretical framework by exploring traditional valuation techniques, the shift from DCF to relative valuation, the pros and cons of traditional methods, and how ML can address their limitations. Then, by assessing the current literature, it delves into the potential of using GBMs as a suitable model aligned with financial theory as argued by Geertsema (2023). By analysing and synthesising the findings from a range of academic articles, research papers, and industry reports, this review aims to support the subsequent analysis and modelling endeavours on the specific case of LSEG financial valuation.

2.1 Scope and inclusion criteria

As visible from Figure 1, articles were sourced from various academic databases, including but not limited to JSTOR, Scopus, and IEEE. The systematic literature review conducted for this study was guided by relevance, recency, and citation analysis. Initial papers were identified through structured keyword searches and Literature Mapping (LitMap), a tool that provides a network visualisation of citations, on various databases. Keywords were derived from the LitMap's conceptual framework (see Appendix A, Figure 1).

The articles considered were screened based on their direct applicability to financial valuation and machine learning applications. Recent publications were prioritised to keep the review contemporary and the LitMap was utilized for identifying seminal works and citation patterns. The final selection includes 33 articles, chosen for their relevance, substantial contribution to the body of knowledge, and recognition within the academic community.

Figure 1: PRISMA 2020 diagram for systematic review (Prisma, 2020)



As LSEG is a global business, the geographical focus of the review was not restricted to any particular region or market, making the review globally inclusive and providing a broad perspective on the subject. The review's scope was inclusive of all sectors and industries, to provide a comprehensive analysis of the application of machine learning techniques in financial valuation.

However, this review does have some limitations. First, the rapid advancement in machine learning techniques and financial valuation practices may not fully represent some of the newest developments in the reviewed literature. In addition, excluding articles not directly related to the subject may have inadvertently left out peripheral discussions that could add depth to the understanding of the topic.

Finally, the complexities of the topic considered and the limitations in the availability of data and resources could constrain this review's comprehensiveness.

2.2 Traditional valuation techniques

Financial valuation can be considered the heart of business, as its practice is essential in corporate finance, portfolio management and market study. Financial literature illustrates a wide spectrum of models, that often make very different assumptions about the drivers of financial value. However, they share some common characteristics that allow for their classification. The most common approaches to financial valuation can be clustered into four categories, presented here from a broader to a narrower applicability.

The first approach considered is discounted cashflow valuation (DCF). In this model, the value of an asset is the present value of the expected cashflows on the asset, discounted back at a rate that reflects the riskiness of these cashflows (Kruschwitz, 2006) (French, 2005). The application of DCF valuation dates back to the 1800s. This method is based on the intrinsic value of the business or asset and is the oldest with the broadest applicability. The second approach with such a broad use is relative valuation, also known as the method of comparable or multiples. It estimates the value of an asset by looking at the pricing of comparable assets relative to a common variable like earnings, cashflows, book value or sales (Cohen, 2000).

The third and fourth approaches are mainly utilised in context-specific scenarios as they focus on specific characteristics. The third approach is liquidation and accounting valuation, which is built around valuing the existing assets of a firm, with accounting estimates of value often used as a starting point. This approach values a business based on what it could be sold for if its assets were sold separately and the company ceased operations. It is often used in bankruptcy or in distressed situations (Poborský, 2015). The fourth and final approach is contingent claim valuation. It is commonly applied for assets that have option-like characteristics, using models developed for financial options to value businesses, equity in businesses, or investments in projects (Glosten, 1994).

As outlined above, there are multiple methods that can be used to conduct financial valuation. Since the focus of this paper is to analyse the LSEG case, it will now discuss the first two approaches as they are the most fit to capture the firm's financial position, as the business is not in liquidation nor it

presents option-like characteristics. (DCF and Relative Valuation mathematical derivations in Appendix B)

Given its history, DCF valuation is the most recognised in academia. However, the industry has been questioning its practicality. The main assumption behind this model is that assets with high and predictable cash flows should have higher values than assets with low and volatile cash flows. In fact quoting Damodaran 'using discounted cash flow models is in some sense an act of faith.' (Damodaran, 2002) . The DCF's underlying hypothesis is that each asset carries an intrinsic value that can be approximated through careful examination of the asset's core features. The issue with the concept of intrinsic value is that it assumes omniscient knowledge on the nature of the asset. This gives rise to an unavoidable dilemma, because of our inability to uncover the true intrinsic value of an asset, rendering us uncertain about whether our estimations based on discounted cash flow valuations align closely with reality or not.

As a result, alternative valuation methodologies have gained traction, with relative valuation being notably prevalent. In relative valuation, the value of an asset is determined based on the price of comparable assets in the market (Damodaran, 2002). There is a substantial philosophical difference between discounted cash flow and relative valuation. The former seeks to discern the intrinsic value of an asset based on its future cash flow generating capabilities. Conversely, the latter involves measuring an asset's worth by comparing it with the market's valuation of similar assets. If the market correctly assigns accurate values to similar assets, the values derived from both discounted cash flow and relative valuation methods may converge.

In conclusion, in financial valuation DCF is a search (albeit unfulfilled) for intrinsic value. Instead, relative valuation gives up on estimating intrinsic value and essentially puts its trust in markets getting it right, and assumes the market is on average correct. It can be argued that most valuations are relative. Damodaran notes that almost 90% of equity research valuations and 50% of acquisition valuations use some combination of multiples and comparable companies and are thus relative valuations (Damodaran, 2002). Although this approach appears to be the most successful, it solely relies on the assumption that the market is correct. However, given the intricacies and complexities inherent in the industry, it is not accurate to assume that the market is always correct. Hence new methodologies that grant a more comprehensive approach and entail more features capturing otherwise unseen patterns must be developed.

2.3 Innovative valuation techniques

Based on the reviewed literature, relative valuation primarily revolves around three critical steps: identifying comparable companies, scaling to standard variables, and determining a valuation range while accounting for differences when comparing standardised values. These steps inherently present specific challenges that are the focal points of contemporary ML research. Additionally, this approach faces comprehensive limitations due to its inability to incorporate non-numeric data and strategic shifts, which require further exploration.

1. Identifying comparable companies

Identifying comparables entails the selection of firms akin to the subject company in terms of industry, size, growth potential, and risk profile. However, it is worth noting that this step presents challenges as determining similarities among firms is not always straightforward. The initial issue stems from the common practice of assuming that firms within the same sector are comparable. However, from a valuation theory perspective, a firm is only comparable if it shares similar fundamental characteristics, regardless of industry. Therefore, a company could be compared to another in a different business, given that they share similar risk, growth, and cash flow characteristics. To tackle this issue, many ML researchers leverage the potential of unsupervised methods, by identifying structures, relationships, and patterns within data without needing pre-existing labels or categories.

One exemplary study is the research conducted by Dameri, Garelli, and Resta (2020) that brings forward a compelling perspective on the complex task of understanding firms' performance. They advocate for using self-organising maps (SOMs), a type of artificial neural network, for analysing and clustering firms' financial performance -hence identifying comparables. Their approach diverges from the conventional practice of reducing the number of examined variables or neglecting their interrelations. Instead, they highlight the importance of maintain data complexity to retain the full set of heterogeneous features involved in financial performance evaluation.

2. Scaling to standard variables

Next, normalisation to common variables plays a crucial role. It allows for a valid comparison by calibrating financial ratios to account for disparities in size, risk, and growth. Yet, the decision on which parameters to use, such as earnings multiples or book values, often remains ambiguous. The challenge is not as frequently addressed by ML researchers as it is in industry. The choice of the common variable appears to be made by sector preference and common practice rather than by financial reasons.

A decidedly unconventional approach to variable normalisation was the one taken by Charles M.C. Lee, and Changyi Wang. These authors introduce a new approach to identify related peer firms using internet search data (Charles M.C. Lee, 2015). They propose that companies appearing in successive searches by the same user, referred as Search-Based Peers (SBPs), share various fundamental similarities. The study reveals that SBPs better explain changes in a company's metrics, such as stock returns, valuation multiples, growth rates, and profitability ratios, compared to traditional industry classifications. This approach aims at reflecting the investors' perceptions of peer firms and can capture real-time market dynamics and sentiment shifts. Although the study reported several beneficial insights, firms often exhibit resistance towards deviating from established valuation metrics, as it may entail transparency and communication challenges.

3. Accounting for differences

The last step of setting a valuation range aims at providing a bigger-picture view by comparing various valuation multiples with industry standards. However, it is not clear what exact factors need to be considered during this step. Another critical aspect is adjusting for differences in fundamentals between companies. However, finding a perfect match for the valued company is almost impossible. This highlights the inherent difficulties in relative valuation. Given the complexity of relative valuation, it is not surprising that there is a lack of studies examining which factors to control for. Literature either controls for macroeconomic factors or applies computational techniques to assess the potential outcomes by simulating numerous iterations of uncertain variables, applying techniques such as the Monte-Carlo Scenario analysis.

The process of relative valuation involves comparing the value of a company to that of its peers in the same industry or sector. While this may seem straightforward in theory, it becomes substantially more complicated in practice due to the factors to consider and the variations among firms within an industry. Adding macroeconomic considerations into the mix only magnifies this complexity, as these broad economic factors have diverse and sometimes unpredictable impacts on different industries and firms. For instance, changes in interest rates or inflation could have varying effects on different sectors depending on factors like debt levels within the industry, pricing power of firms, and the industry's sensitivity to consumer demand.

On the other hand, many academics and industry experts have pivoted to applying stochastic modelling techniques to capture uncertainty. The research of Aubrey Clayton (2020) explores the

application of Monte Carlo simulation in the insurance world, specifically in the context of cash flow testing and projected liability valuation for complex variable annuities. The approach involves utilising deep learning techniques to develop a proxy model for cash flow, which can be applied to various valuation scenarios. Using this proxy model, the article demonstrates that the results compared favourably to traditional 'brute force' Monte Carlo simulation, requiring significantly fewer computational resources.

In conclusion, the literature review reveals that relative valuation, while widely practised in the industry, poses inherent challenges that have become focal points of contemporary ML research. The three critical steps involved in relative valuation—identifying comparable companies, scaling to standard variables, and controlling for differences—present complex issues researchers have sought to address.

2.4 Harnessing GBMs for Financial Valuation

Navigating the plethora of ML solutions highlighted in the literature, it is imperative to choose an algorithm aligned with the study objectives, data availability, and practical feasibility. Key considerations include handling complex data, identifying comparables accurately, scaling to standard variables, and robustness in valuation range determination.

One notable paper aiming to address all three challenges highlighted in Section 2.3 is presented by P. Geertsema and H. Lu (2023). The study demonstrates that GBM models offer more accurate valuations than traditional ones, with valuations closely resembling fundamental value. Proving how overvalued stocks tend to drop in price while undervalued stocks show an increase in price over the next month. The researchers also utilise feature importance and SHAP values to examine the relationship between accounting information and valuation multiples, highlighting that the influential variables in GBM predictions align closely with variables derived from theoretical DCF models. Ultimately, the study demonstrates the ability of GBM models to identify complex relationships between financial information and firm value. However, this method presents limitations in fully capturing strategic effects.

This issue is addressed by Stankevičienė's (2012) research, which presents a comprehensive examination on the use of numerous valuation methodologies on strategic effects. More specifically, this work focuses on the impact of companies' restructuring strategies on their financial performance.

While the author acknowledges the importance of traditional techniques, she also emphasises their limitations and suggests the inclusion of more unconventional metrics. Subsequent studies, inspired by Stankevičienė's work, have successfully integrated textual data into pricing and valuation methodologies, yielding both insightful and efficacious outcomes utilising FinBERT [(Halder, 2022), (Chen Q. , 2021), (Liapis CM, 2023)]. Guided by literature insights, the present study amalgamates traditional and unconventional metrics as essential features in ML models. The intent is to enhance the comprehension and predictive ability of strategic business shifts, thereby addressing the limitations of traditional valuation methods.

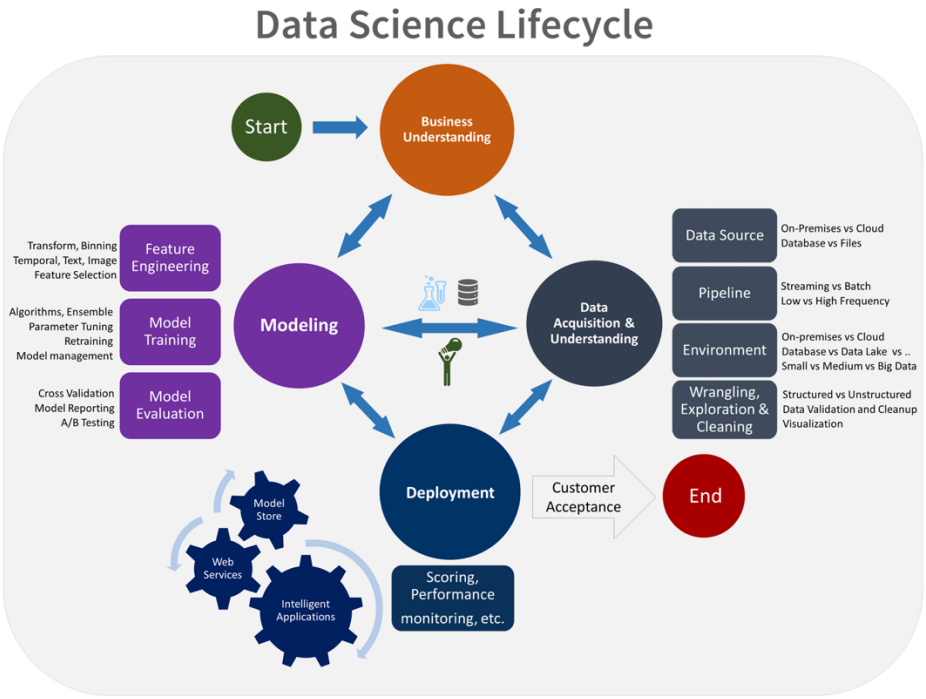
This paper focuses on the repositioning and subsequent valuation of LSEG, a task that stretches the capabilities of traditional valuation methods. Drawing on relevant literature, it applies these models to a company undergoing significant financial restructuring. This focused approach aims to enhance our understanding of GBM's application in valuing companies amid substantial transformations, incorporating unstructured data sources. Furthermore, while it is acknowledged in the field and evident in literature trends that unstructured data can yield valuable insights, such data is not commonly integrated into firm-specific valuation scenarios. This work builds upon existing research by demonstrating how this inclusion can enhance the accuracy and richness of the valuation process. Our ultimate goal is to refine the methodology and provide insightful analysis beneficial to investors, decision-makers, and stakeholders involved in repositioning and valuation.

3. Methodology

3.1 The TDSP approach

This paper applies the Team Data Science Process (TDSP) methodology developed by Microsoft (Aparicio, 2020). This iterative methodology deploys predictive analytics solutions for decision-making. The TDSP lifecycle is composed of five major stages: Business Understanding, Data Acquisition and Understanding, Modelling and Deployment (Figure 2). This methodology imposes an exercise of continuous research and discovery, enabling the paper to reach broader findings and deploy tangible results. (Project Management details Appendix C)

Figure 2 Team Data Science Process (TDSP)- Microsoft



A unique feature of the TDSP is seen in the use of SMART criteria (Specific, Measurable, Achievable, Relevant, Time-bound) to define objectives (Martinez, 2021). SMART objectives are essential in creating specific, attainable goals for this project since it strongly emphasises on producing business-relevant results. It is also important to remember that the TDSP depends on Microsoft's services and policies, which can be a downside in broader scenarios. But in the scope of this paper, it is advantageous as LSEG and Microsoft have established a partnership in 2023. By utilising Microsoft's resources and technology, this cooperation increases the likelihood that the project will achieve LSEG's SMART goals.

This standardised methodology grants clarity and reproducibility by ensuring robust data lineage. It does so by tracking data from its original source through each transformation, thus ensuring the integrity and traceability of the information used. Moreover, the TDSP's iterative approach implies regular re-evaluation and fine-tuning of models and hypotheses, ensuring that the research stays adaptable and responsive to new insights. Lastly, TDSP's emphasis on comprehending business implications and fostering stakeholder involvement ensures that the research remains aligned with real-world needs and can be conveyed to a corporate audience. In summary, the TDSP methodology has been chosen because it can increase this research's potential for robust results, tangible application, and broader impact.

3.2 Business Understanding

Although the primary focus of this study is the valuation of LSEG, creating a reliable and generalisable predictive model necessitates training and evaluation across a comprehensive and diverse dataset. For this reason, this paper employs data from the Standard & Poor's 500 companies. This is because the S&P index considers a statistically significant number of companies across different industries and business contexts. This methodological approach provided a robust basis for developing the paper's GBMs and served as a rigorous benchmark validating their effectiveness in predicting the EV/EBITDA.

3.2.1 The target variable: EV/EBITDA

Enterprise Value (EV) represents a company's total value, including its market capitalisation minus its cash and cash equivalents, and then added to its debt. EBITDA, an acronym for 'earnings before interest, taxes, depreciation, and amortisation', provides a measure of a company's operational earnings. (Ohlson, 1995)

The EV/EBITDA ratio is a financial metric comparable to the Price/Earnings (P/E) ratio but with a significant difference: it is indifferent to a company's capital structure (its debt and equity distribution). This feature makes it especially valuable when comparing companies with different debt levels, such as in our case. In LSEG's industry, peers are characterised by high debt levels or unusual non-cash expenses compared to the average. This could distort its earnings representation using the P/E ratio. In fact, the P/E ratio offers insight into what the market is willing to pay today for a company's future earnings. A high P/E ratio could suggest that the market has high expectations for a company's future growth, while a low P/E ratio may indicate the opposite. (Gottwald, 2012). EV/EBITDA concentrates on a company's gross profit, offering investors an impartial insight into its earning potential while standardising the effects of its capital structure and accounting methods. Therefore, this ratio helps to provide a more accurate representation of a company's economic reality (Chan, 2010).

3.2.2 The S&P index

The S&P index is widely regarded as the best single gauge of large-capitalisation U.S. equities. According to the Annual Survey of Assets, an estimated USD 15.6 trillion is benchmarked to the index

(as of Dec. 31, 2021). The index includes 500 leading companies and covers approximately 80% of available market capitalisation. It is the basis of many listed and over-the-counter investment instruments and is considered an overall benchmark of the global equity markets [(Frino, 2001), (S&P Dow Jones Indices, 2022)].

The models are trained on a robust dataset consisting of financial metrics and related features from the S&P500 companies over 20 quarters. This dataset was specifically chosen due to its robust statistical reference, generalizability, comparability, overall quality, and credibility of reference point. Training the models on this dataset provides statistical reference for financial valuation metrics. The dataset encompasses various industries and company sizes (Figure 3, Table 1), offering a representative sample to capture the overall market dynamics.

Figure 3 S&P 500 Sector Breakdown

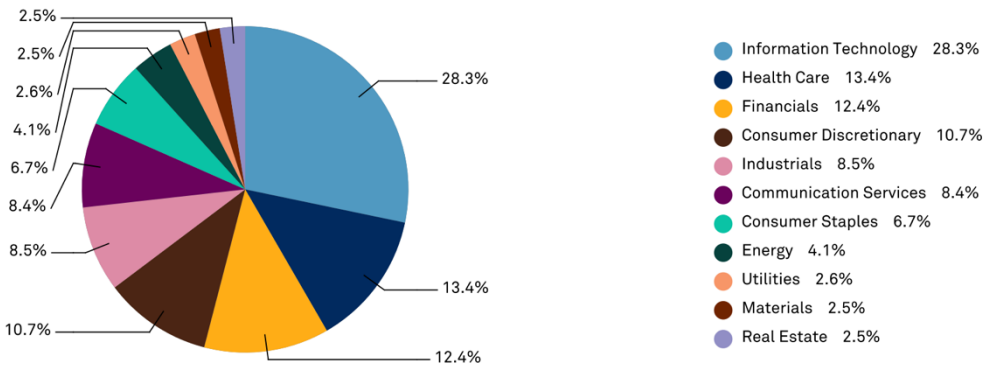


Table 1 S&P Characteristics expressed in GBP

NUMBER OF CONSTITUENTS	503
CONSTITUENT MARKET [USD MILLION]	
MEAN TOTAL MARKET CAP	80,195.51
LARGEST TOTAL MARKET CAP	3,089,903.51
SMALLEST TOTAL MARKET CAP	3,871.48
MEDIAN TOTAL MARKET CAP	32,156.59
WEIGHT LARGEST CONSTITUENT [%]	7.6
WEIGHT TOP 10 CONSTITUENTS [%]	30.6

Utilising the S&P dataset increases the likelihood of capturing patterns and relationships that apply to a broader range of companies, allowing this paper to test the significance of its method. In addition, training the models on this dataset enables comparative analysis among the companies and sectors. The models can capture similarities and differences in the financial valuation metrics across various

industries, sectors, and company sizes. This allows for deeper insights into the factors influencing valuation metrics and the ability to identify outliers or anomalies. Lastly, using this dataset builds the reproducibility and credibility of the paper as the data is widely available, recognised and trusted by academia and industry.

3.2.3 The LSEG case

Having elucidated the rationale and methodology behind training and validating our models on the broad dataset of the S&P 500, this paper now focuses on the specific case study of the London Stock Exchange Group and the selected peers (Appendix D).

The LSEG case study is a compelling example due to its current strategic shift in the business landscape. As the company endeavours to transform from a traditional stock exchange to a comprehensive platform offering data services, exchange, and post-trade services, it is positioning itself to raise its market competitiveness, profitability, and valuation. This intricate process, however, requires precise measurement. Traditional valuation methods fail to capture the entirety of the strategic decisions taken by LSEG, thereby underscoring the necessity for an innovative, data-driven approach to valuation, such as the one this paper implements through GBM.

The acquisition of Refinitiv, a data business, by LSEG in 2020 resulted in a significant shift in its revenue model from transactional to recurring sources (LSEG, 2023). This transition challenges traditional valuation methods, which exclusively rely on historical financial statements. However, with LSEG's strategy shift, business dynamics have substantially changed. Integrating a data business into the firm's operations introduces new revenue streams and alters the company's growth prospects and perception. Hence, in this context, relying exclusively on past financial records is faulty.

Furthermore, traditional valuation techniques do not incorporate qualitative factors associated with the shift to recurring revenue (Stankevičienė, 2012). Factors such as market sentiment and market positioning become crucial considerations in assessing the value generated by a recurring revenue model. Traditional methods, which primarily rely on quantitative financial data, do not account for these qualitative factors and their impact on the company's valuation. Given these challenges, traditional valuation techniques fall short of capturing the complete value created by LSEG's shift. To overcome these limitations, alternative approaches may be necessary to assess the value generated by

the transformed revenue model accurately. In this context, a pool of data and exchange peers was selected by experts to provide an appropriate benchmark for this ongoing shift.

3.2.4 Overview of the business framework

The past sections have highlighted the importance of using structured and unstructured data in forecast financial valuation to gain deeper insights into the factors driving these valuations. The rationale behind the decision to predict EV/EBITDA rests on the need for a comprehensive assessment of a company's value and the facility to benchmark our prediction against traditional financial forecasts. Simultaneously, it justifies the decision to train and test GBMs using data from a broad range of S&P 500 companies before focusing on LSEG as a single entity. This decision aims to ensure the robustness and generalizability of the paper whilst also acknowledging the unique business context and complexities that come with LSEG's strategic repositioning in the market.

Given the iterative methodology of the study, it is crucial to remain consistent with the SMART objectives. The aim of this paper is not merely to demonstrate a technical exercise in ML, but also to illustrate its practical application in a real-world financial scenario, understand global market influences, and examine the impact of strategic business decisions on a single company's valuation. The results of this study are expected to provide relevant insights that could aid in strategic decision-making and successful business repositioning for entities similar to LSEG.

4. Data Acquisition and Understanding

As the TDSP methodology outlines, the second step is Data Acquisition and Understanding. The following chapter will start by presenting the data-gathering process. It then expands on data validation, lineage concerns, and the architecture and rationale behind the pipeline and storage choices. Then, it will discuss data wrangling, exploration and cleaning processes to best prepare the data for the GBM algorithms. This section aims to ensure transparency and reproducibility, ensuring business-grounded results.

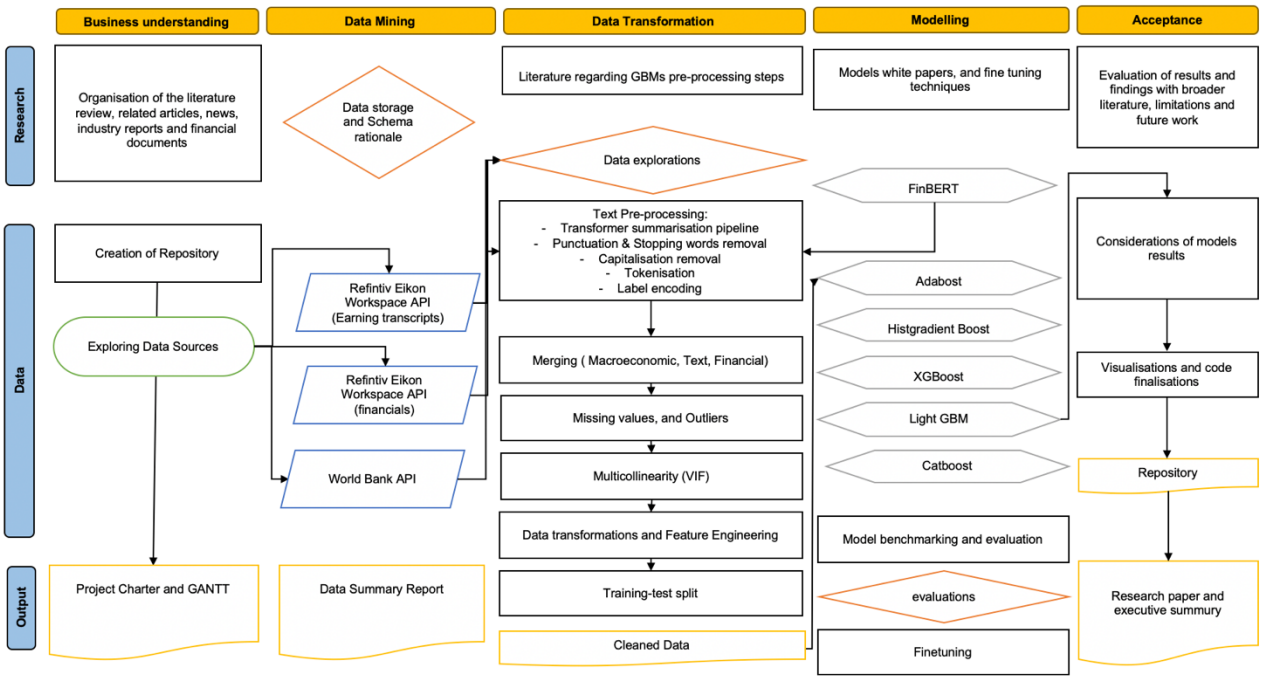
4.1 Data Acquisition and Architecture

The dataset used for this task was obtained from the EIKON Workspace API, a platform provided by Refinitiv, now owned by LSEG. The API grants access to various financial and market data services.

EIKON, formerly known as Thomson Reuters EIKON, is recognised as a flagship product of Refinitiv/LSEG, widely utilised by over 200,000 analysts worldwide (LSEG, 2023). Leveraging such a reputable and reliable financial data source ensures the optimal validity and quality of the data utilised in this study.[1] In addition to EIKON, macroeconomic data was obtained from the World Bank API, further enriching the dataset with a broad range of global economic indicators.

The data pipeline design and implementation involved a systematic and structured approach to handle data flow from various sources to the final analysis and reporting stages. The pipeline encompassed several key steps: data mining, transformation, and ML processing. More details on the architecture rationale can be seen in Figure 4 below (Appendix E.1 and E.2).

Figure 4 Project Process Flowchart



4.2 Data Cleaning and Wrangling

The pre-processing pipeline comprises critical steps to prepare the data optimally for modelling. Initially, missing values and outliers are addressed to maintain data integrity. Text data is appropriately

[1] The data acquisition and analysis process adhered to strict corporate privacy guidelines. The data was obtained from authorised sources with proper permissions and licenses. In order to protect proprietary data, trade secrets, financial records, and other sensitive company information from unauthorised disclosure, private data was excluded to meet privacy and confidentiality concerns.

handled by transforming it into numerical representations with FinBERT, as GBMs only processes numerical inputs. Then, the dataset is divided into training and testing sets to enable accurate model development and evaluation. Finally, data distributions and statistical properties are explored to potentially implement transformations.

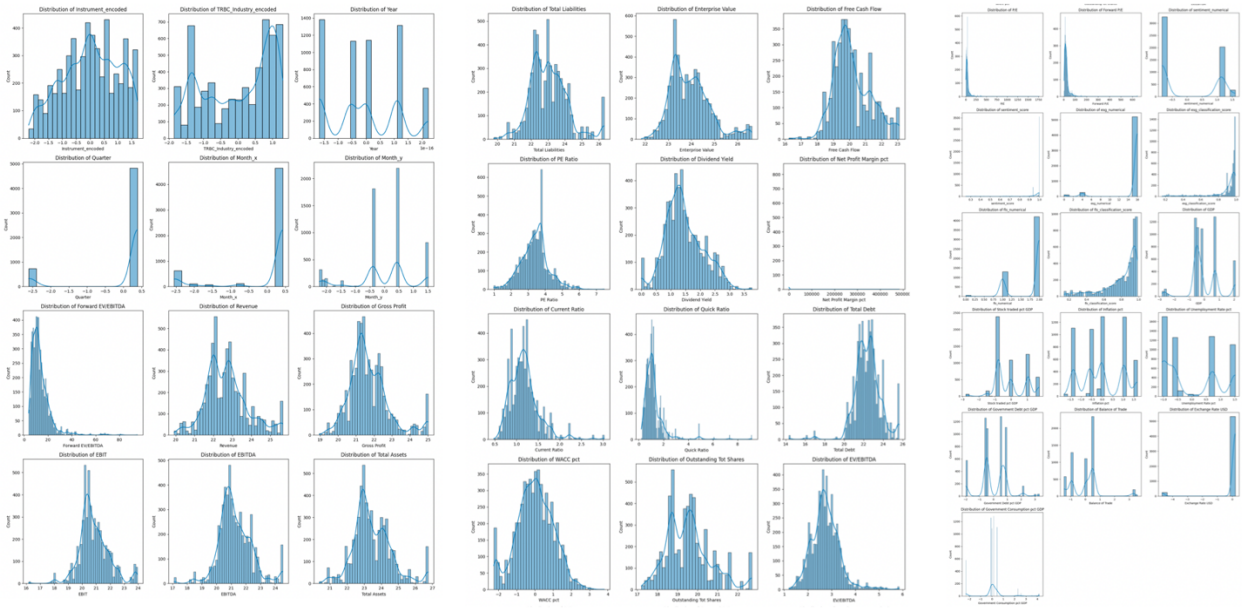
4.2.1 Missing data and outliers

To handle missing data, Iterative Imputation techniques were applied to estimate missing values. Imputation is a technique used to estimate missing values in a dataset, enabling the inclusion of otherwise incomplete data points. It preserves the overall structure and integrity of the dataset while minimising bias introduced by complete case analysis. The iterative technique is an advanced imputation methodology that iteratively estimates missing values based on the relationships between variables. This method fills missing data using initial estimates and a predictive model trained on observed values. This process is repeated until convergence or a stopping point is met (Jäger, 2021). In the context of financial valuation, outliers play a crucial role and have been retained in the model. Outliers represent extreme or exceptional events that can significantly impact valuation metrics. The analysis captures these unique instances, allowing for a more accurate representation of the market dynamics and potential disruptions. (Frecka, 1983) Removing outliers could lead to a loss of valuable insights and distort the overall understanding of the financial landscape, making their inclusion essential for a comprehensive and robust valuation analysis. (Zellner, 1981)

4.2.2 Data Transformation and Feature Engineering Considerations

To better understand the data, the distributions of the variables have been studied. This was done by analysing the statistical properties and identifying skewness and kurtosis. (Royston, 1992) The dataset paints a complex, multi-faceted picture of economic and financial conditions across various sectors and companies. This paper employed numerous feature transformations with the intention of refining our dataset to maximize its compatibility with GBMs. Despite exhaustive trials (Appendix E.3), these transformations did not lead to improvements in model performance or acceleration in convergence speed. Future research could provide additional insights by exploring various other types of transformations and building on the foundation established in this paper. Not implementing transformations in this context simplifies the data preparation process and enhances the model's interpretability, thereby improving its applicability for stakeholders.

Figure 5 Data Distribution



4.2.3 Multicollinearity

Multicollinearity, a statistical phenomenon where independent variables in a regression model exhibit high correlation, was found in our dataset through the Variance Inflation Factor analysis (Mansfield, 1982). A host of variables, including financial metrics as well as several macroeconomic indicators, were correlated. Addressing multicollinearity is critical as it could potentially affect the stability and reliability of the model's predictions. Traditional methods of handling multicollinearity, such as removing correlated variables, were considered but dismissed due to the need to maintain the interpretability of the model. As a solution, this paper turned to regularisation techniques such as L1 and L2 in the algorithms.

4.2.4 FinBERT and textual data

As the paper aims to capture the complexities and dynamics of the market beyond traditional data, we incorporated a scoring system based on analysing, labelling and scoring earning transcripts and press releases from each quarter (Weiss, 2010). GBMs do not handle text data directly (Bentéjac, 2021), so all the text information has been transformed into a numerical representation with label encoding.

The text data underwent a series of pre-processing steps to ensure its suitability for analysis. Firstly, this paper utilised the 'Transformers summarization pipeline' to condense the text into concise summaries (Gupta, 2022). Additionally, various pre-processing techniques have been applied using the Natural Language Toolkit (NLTK), including removing punctuation, stopping words, and capitalization (Bird,

2006). Furthermore, the tokens were lemmatized to reduce words to their base or root form, allowing for more meaningful comparisons. To extract label and score the pre-processed text, we employed the FINBERT language model (Yang, 2020). FINBERT is specifically designed for financial text analysis, enabling the extraction of sentiment (TONE), environmental, social, and governance (ESG) factors, and forward-looking statements (FLS). [2] It is based on the powerful BERT (Bidirectional Encoder Representations from Transformers) architecture, which utilizes a transformer network to learn contextual representations of words in a text corpus. (Peng, 2021)

When using FINBERT for sentiment analysis and classification tasks, the model assigns a label and a corresponding score to each input text. The label represents the predicted sentiment or classification category, such as positive, negative, or neutral sentiment, or specific categories like 'Environment' , 'Social', 'Governmental', or 'None' for ESG classification and 'Non-specific FLS', 'Specific FLS' and 'Not FLS' for Forward-Looking Statements. The score indicates the confidence or probability associated with the assigned label. (Huang, 2020) Higher scores indicate higher confidence in the assigned sentiment or classification. This allowed us to incorporate qualitative information and sentiment analysis into the predictive models, complementing the quantitative financial and macroeconomic data (Figure 6).

Figure 6 Classified Text Dataframe

Instrument	Guidance Measure	The Doc Type	Year	Month	preprocessed_text	num_tokens	sentiment_results	sentiment_score
POOL.OQ	EBITDA	Transcript	2022	2	strategic addition business e	44	{'label': 'Positive', 'score': 0.8	0.8201017379760742
POOL.OQ	EBITDA	Transcript	2022	2	strategic addition business e	44	{'label': 'Positive', 'score': 0.8	0.8201017379760742
POOL.OQ	EBITDA	Transcript	2022	2	strategic addition business e	44	{'label': 'Positive', 'score': 0.8	0.8201017379760742
POOL.OQ	EBITDA	Transcript	2022	2	strategic addition business e	44	{'label': 'Positive', 'score': 0.8	0.8201017379760742
CHRW.OQ	EBITDA	Transcript	2023	4	would expect , cash used re	55	{'label': 'Neutral', 'score': 0.9	0.9978664517402649
AJG.N	EBITDA	Transcript	2023	4	look forward , 're seeing rest	37	{'label': 'Neutral', 'score': 0.5	0.5438294410705566
AJG.N	EBITDA	Transcript	2023	4	look forward , 're seeing rest	37	{'label': 'Neutral', 'score': 0.5	0.5438294410705566
AJG.N	EBITDA	Transcript	2023	4	think accounting deferred re	114	{'label': 'Neutral', 'score': 0.9	0.9024152159690857
AJG.N	EBITDA	Transcript	2023	1	another year double-digit gr	29	{'label': 'Positive', 'score': 1.0	1.0
AJG.N	EBITDA	Transcript	2023	1	another year double-digit gr	29	{'label': 'Positive', 'score': 1.0	1.0
AJG.N	EBITDA	Transcript	2023	1	another year double-digit gr	29	{'label': 'Positive', 'score': 1.0	1.0
AJG.N	EBITDA	Transcript	2023	1	another year double-digit gr	29	{'label': 'Positive', 'score': 1.0	1.0
AJG.N	EBITDA	Transcript	2022	7	delivering $ 730 million rever	47	{'label': 'Positive', 'score': 1.0	1.0
AJG.N	EBITDA	Transcript	2022	10	right saying ?	8	{'label': 'Neutral', 'score': 0.9	0.999784529209137

[2] Forward-looking statements, in the context of financial analysis, refer to statements made by companies or individuals that anticipate future events, outcomes, or financial performance. These statements are often included in financial reports, earning transcripts, or press releases and provide insights into the company's expectations, projections, or plans for the future.

esg_classification	esg_classification_score	fls_classification	fls_classification_score	sentiment_numerical	esg_numerical	fls_numerical
{'label': 'None', 'score': 0.9865126609	0.9865126609802246	{'label': 'Specific FLS', 'score': 0.8	0.8901234269142151	1	4	1
{'label': 'None', 'score': 0.9865126609	0.9865126609802246	{'label': 'Specific FLS', 'score': 0.8	0.8901234269142151	1	4	1
{'label': 'None', 'score': 0.9865126609	0.9865126609802246	{'label': 'Specific FLS', 'score': 0.8	0.8901234269142151	1	4	1
{'label': 'None', 'score': 0.9865126609	0.9865126609802246	{'label': 'Specific FLS', 'score': 0.8	0.8901234269142151	1	4	1
{'label': 'None', 'score': 0.9947546720	0.9947546720504761	{'label': 'Not FLS', 'score': 0.65203	0.6520322561264038	0	4	2
{'label': 'None', 'score': 0.9812947511	0.9812947511672974	{'label': 'Not FLS', 'score': 0.57290	0.5729066729545593	0	4	2
{'label': 'None', 'score': 0.9812947511	0.9812947511672974	{'label': 'Not FLS', 'score': 0.57290	0.5729066729545593	0	4	2
{'label': 'None', 'score': 0.9882681965	0.9882681965827942	{'label': 'Not FLS', 'score': 0.94220	0.942201554775238	0	4	2
{'label': 'None', 'score': 0.9881935715	0.9881935715675354	{'label': 'Not FLS', 'score': 0.95483	0.954839825630188	1	4	2
{'label': 'None', 'score': 0.9881935715	0.9881935715675354	{'label': 'Not FLS', 'score': 0.95483	0.954839825630188	1	4	2
{'label': 'None', 'score': 0.9881935715	0.9881935715675354	{'label': 'Not FLS', 'score': 0.95483	0.954839825630188	1	4	2
{'label': 'None', 'score': 0.9881935715	0.9881935715675354	{'label': 'Not FLS', 'score': 0.95483	0.954839825630188	1	4	2
{'label': 'None', 'score': 0.9834732413	0.9834732413291931	{'label': 'Not FLS', 'score': 0.91375	0.9137547016143799	1	4	2
{'label': 'None', 'score': 0.9723539352	0.9723539352416992	{'label': 'Not FLS', 'score': 0.96173	0.9617305397987366	0	4	2

By incorporating text data through the aforementioned pre-processing, this paper aimed to capture a more comprehensive understanding of market dynamics. This hybrid approach combining numerical and textual information enhances the predictive capabilities of the models, enabling to forecast standard financial valuation metrics with a deeper understanding of the underlying qualitative factors. Moreover, it allows the model to capture factors impossible to consider with traditional financial valuation methods. Including text-based scoring and sentiment analysis strengthens the richness of the analysis of this paper, providing valuable insights into the market that extend beyond traditional numerical factors.

Finally, categorical variables have been transformed into numerical representations using Label Encoding to allow the GBMs to interpret the input. Each unique category or level in a categorical variable is assigned a unique numeric label. The label encoding process involves assigning integers starting from 0 to the categories in the variable. For instance, a variable with three categories may be encoded as 0, 1, and 2, respectively. This method has been selected as it is straightforward and does not introduce additional complexity or dimensionality to the data.

Utilizing FinBERT in text analysis comes with certain limitations. It is important to acknowledge these limitations, which include increased complexity and the potential introduction of biases in the predictive model. Firstly, not further fine-tuning the model was a deliberate choice to simplify the implementation process. While this decision offers convenience, it may impact the model's performance by limiting its ability to adapt to the specific dataset at hand. In addition, one notable limitation of FinBERT is the complexity it introduces in terms of interpretability. The model's inner workings can be intricate, making it challenging to explain its predictions clearly. This can hinder the transparency and comprehensibility of the results obtained from the model. Lastly, like any model, FinBERT is susceptible to biases in the training data. If the training data contains biases or lacks

diversity, these biases can manifest in the model's predictions and influence the analysis. By acknowledging the challenges these limitations present, researchers can consider their implications on the interpretation and generalizability of the results obtained from the overall paper. Despite these limitations, FinBERT remains a valuable tool in financial text analysis. Its ability to capture financial domain-specific patterns and extract insights from complex textual data outweighs the challenges it presents.

5. Modelling

Grounded in the comprehensive literature review, this paper determines that GBMs present a viable methodology for its research question. Their proven robustness and ability to manage intricate, non-linear relationships between variables render them an excellent choice for this study. GBMs represent an umbrella term for many algorithms: for this reason, this assumption is merely the tip of the iceberg in this exploration. In fact, this paper now delves deeper on the genealogy of these models, considering their unique attributes, strengths and limitations to determine which GBM architecture best fits the research question. In the subsequent sections, we will explain the model selection process and its congruence to financial theory, followed by an evaluation of the algorithms tested and a description of the training phase, shedding light on the strategies adopted for hyperparameter tuning and parameter optimization. In this context, Section 5 concludes that LightGBM leads to the most meaningful results.

5.1 Model Selection

GBM is part of the Boosting class of ensemble learning algorithms, designed to enhance the predictive performance of ML models. This technique employs many simpler models, weak learners or base models, which individually only provide limited predictive ability. However, when aggregated using a boosting algorithm, these models produce a more robust and accurate ensemble model. This ensemble approach significantly enhances prediction accuracy by leveraging the strengths of individual base models. (Schapire, 2012)

This paper now considers the various boosting models available. Boosting was born out of a question postulated by Kearns in 1988 and Valiant in 1989. They wondered if a set of weak classifiers could be transformed into a single strong classifier (Kearns, 1988) (Kearns M. a., 1989). Schapire answered this query in 1990 with a positive response (Schapire R. E., 1990). Subsequently, Freund and Schapire proposed AdaBoost in 1996, the first practical boosting algorithm (Freund, 1996). AdaBoost iteratively

fits a weak classifier to weighted versions of the data, reweighting data at each iteration such that misclassified data points receive larger weights. Boosting algorithms underwent further evolution when Breiman (1997, 1998) introduced a view of boosting algorithms as numerical optimization techniques in function space (Ridgeway, 1999). This perspective allowed Mason et al. (1999) and Friedman (2001) to develop a more generalised version of boosting, which could optimize any differentiable loss function (Friedman, 2001). This significant advancement made boosting applicable to a broader range of regression problems, not just classification tasks. Finally, this paper considers the algorithms presented by Friedman in 2001, known as gradient boosting, which have remained popular in practice. It is motivated as a gradient descent method in function space. (Chen, 2016)

Building on the work of Friedman, other researchers developed multiple variations, creating a broad family of models for different datasets and tasks. In this paper, we engage in an extensive modelling phase to identify the most appropriate algorithm, testing different models individually and then comparing their performance. Through this comparative analysis, LightGBM has been identified the most suitable algorithm and has been fine-tuned to improve performance further. This approach not only provides confirmation of the suitability of GBMs, as highlighted in the literature review, but also allowed a refined choice of a specific algorithm that worked best for our unique data and research objectives. To further ascertain the suitability of GBMs in this specific context, this study empirically evaluated the model, focusing on its compatibility with financial theory and its predictive accuracy. The results, outlined below, corroborate the claims made in the literature and support the decision to employ GBMs in this research. This paper takes the use of GBMs a step further by incorporating sentiment analysis and factor classification using FinBERT. By considering the sentiment derived from unstructured textual data in company filings and media announcements, it aims to capture the nuanced dynamics not readily observable in structured data. This enhances the model's predictive capabilities and allows a more informed interpretation of how various factors contribute to a firm's valuation. It creates a more robust and sophisticated valuation model that can handle the complexities of firms repositioning in the market and changing their revenue structures.

With these considerations in mind, this paper acknowledges the suitability of GBMs in a broader and narrow context— S&P500 and firms repositioning in the market and altering their revenue structures. Furthermore, it explores improvements by incorporating additional unstructured data sources into the model. By doing so, it increases the model's accuracy and predictive power, providing more reliable and nuanced insights into firm valuation.

5.2 Model training and evaluation

Having expanded on the model selection, we systematically explored the six following models that fall under the GBM umbrella: Adaboost, Histgradient boost, Xgboost, Light GBM, Catboost (more details on their performance can be found in the appendix). To comprehensively evaluate the models' predictive capabilities, each model was trained separately with the same dataset, using EV/EBITDA as a target variable (Appendix F for complete features list). The models were initialized with default parameters, creating a baseline evaluation. Then the best-performing algorithms were fine-tuned using a cross-validation process to ensure robust performance across different subsets of data.

Considerations of computational efficiency primarily drive the decision to fine-tune only the best-performing model. Fine-tuning, while crucial for optimizing model performance, is a computationally intensive process. It demands significant resources, both in terms of time and computational power. Given these constraints, the study adopted a more resource-efficient strategy. By focusing our fine-tuning efforts only on the most promising models, the papers aims to balance computational efficiency and performance optimization, thereby ensuring that our study is both effective and resource-conscious.

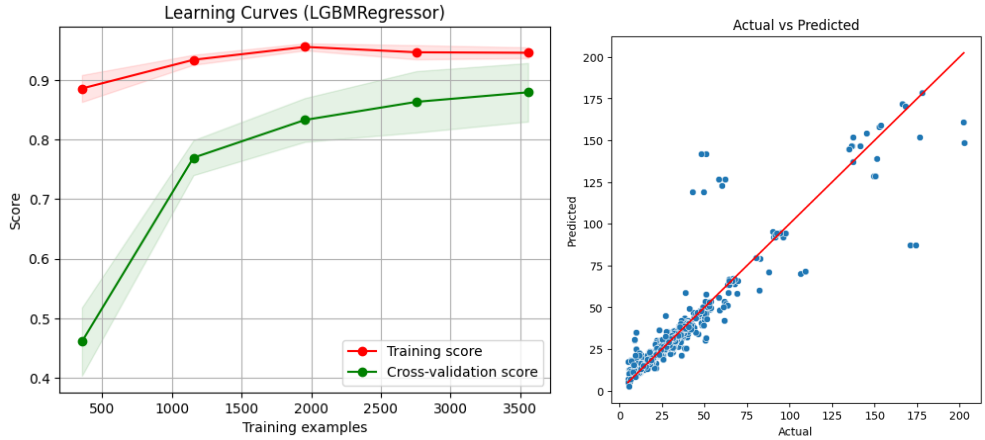
Given the regression task, we relied on regression-oriented metrics to evaluate our models' performances. The key metrics used are Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2). After training each model, metrics were computed for both the training and validation data, allowing to monitor how well the model fits the data and its ability to generalize to new and unseen data, visible in table 2 ('Learning Curves' and 'Actual vs Predicted' Plots Appendix G). The results of these evaluation metrics for each of the six models on both the training and validation datasets are summarized in Table 2 below:

Table 2 Models and Performance on S&P

Model	*RMSE*	*MAE*	*R^2*
Adaboost	15.2582	10.6813	0.5616
Histgradient boost	8.5624	2.9755	0.8600
Xgboost	10.2151	2.3119	0.8046
Light GBM	8.4419	2.9880	0.8624
Catboost	9.7310	2.8700	0.8136

The table provides a clear comparison of the predictive capabilities of each model, giving us a comprehensive understanding of their EV/EBITDA prediction. LightGBM appears to be the most efficient (lowest RMSE). Nonetheless, a deeper analysis of the models must be conducted. As shown in Figure 7, exploring the models' learning curves and plots can draw precious insights on bias-vs-variance tradeoff, model complexity, Overfitting/Underfitting Detection and data sufficiency. Furthermore, the 'Actual vs Predicted' plot can reveal accuracy, fit, outliers handling and systematic bias.

Figure 7 Learning Curves and Actual vs Predicted



Among the models evaluated, LightGBM stands out as the best performer. Its learning curve steadily improves as the training set size increases, indicating effective learning and generalization. The actual vs. predicted plot reveals accurate predictions, with points closely aligned to the diagonal red line. Although S&P 500's outliers generate slight deviations, the model captures well underlying patterns. LightGBM's RMSE of 8.4419 and R-squared of 0.86 signify strong predictive power, outperforming other models like Adaboost. HistGradientBoost, XGBoost, and Catboost also exhibit respectable performances but fall short compared to Light GBM. Overall, LightGBM proves to be a reliable and accurate choice, outperforming other algorithms in both RMSE and R-squared metrics.

5.3 LightGBM, the best performing model

LightGBM stands out as the optimal model due to its multifaceted advantages. These include its advanced boosting mechanism, its approach to the dimensionality curse, its adaptability, its many hyper parameters and regularization and randomization techniques, along with its computational efficiency.

5.3.1 Advanced Boosting Mechanism:

Like HistGradientBoost, LightGBM employs histogram-based algorithms that convert continuous feature values into discrete bins, both algorithms for instace present a high R^2 (around 86%).These algorithms offer a performance that nearly matches the accuracy of Newton Boosting (employed by XGBoost, which is reflected in the lowest MAE (2.3119), while being more efficient in terms of memory and computation. Furthermore, LightGBM employs a hybrid approach combining gradient boosting with histogram-based techniques, allowing it to deliver impressive accuracy without compromising efficiency. In addition, LightGBM employs Gradient-based One-Side Sampling (GOSS), a technique that aims at resolving the potential loss of information that often occurs in gradient boosting machines (Al Daoud, 2019).

5.3.2 Mitigation of Dimensionality Curse:

Contrary to numerous models that succumb to the degradation of performance with escalating problem dimensionality, LightGBM possesses a unique capability to handle high-dimensional data, like the dataset handled in this paper. This is realised through the use of Exclusive Feature Bundling (EFB) (Hancock, 2021). Dealing with high-dimensional data can be challenging due to its inherent sparsity and exclusiveness, where features often do not take non-zero values simultaneously. EFB capitalizes on this exclusiveness by bundling these features together. By reducing dimensionality, EFB not only enhances computational efficiency but also decreases memory usage, allowing the model to handle more complex datasets.

5.3.3 Versatility and Adaptability:

One of LightGBM's cardinal strengths lies in its adaptability which is essential to tackle the bias-variance trade-off. The methodology accounts for the bias-variance trade-off at every stage of the learning process. Unlike other gradient boosting frameworks that grow trees level-wise, LightGBM grows trees leaf-wise, meaning it chooses the leaf with the maximum delta loss to grow, leading to better optimisation of the objective function and thereby reducing bias. Furthermore, LightGBM addresses bias variance through enhanced regularization and randomization techniques. It provides an array of additional regularisation parameters such as L1 and L2 which control the complexity of the trees and thus help to avoid overfitting, thereby controlling variance.

5.3.4 Computational Efficiency:

Lastly, LightGBM also exhibits computational efficiency, which is particularly critical when working with large datasets. Computational resource considerations can heavily dictate the choice of a model.

LightGBM employs GOSS, a technique that balances computational efficiency with model accuracy. GOSS samples data instances based on their gradients to prioritize high-error instances for model improvement. It uses a dual sampling approach to maintain the original data distribution, thereby ensuring accuracy. This method allows to focus on the most valuable data for training without significantly increasing computational load, hence improving computational efficiency.

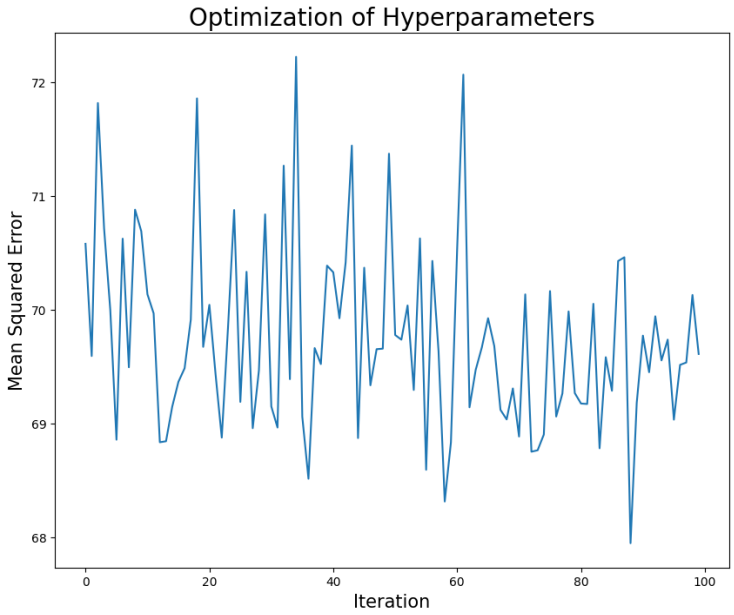
In light of these considerations, the decision to exclusively fine-tune the LightGBM model was made for this study. The computational demands of model fine-tuning, often a resource-intensive process, were pivotal in this choice. Concentrating on the most promising model, we aimed to strike an effective balance between computational efficiency and performance optimisation.

5.4 Hyper Parameter tuning

Friedman, the architect of GBMs articulated the necessity of regularisation for tackling the overfitting and underfitting issues. He emphasized the adjustment of three meta-parameters: the number of iterations, the learning rate, and the number of terminal nodes/leaves (Friedman, 2000, pp. 1203, 1214–1215). Drawing upon this, LightGBM carries forward the principle of regularisation from the GBMs and broadens its application. The model offers a rich array of tunable hyperparameters to regulate decision trees' complexity. Among these, this paper has considered learning rate adjustments, sub-sampling, and L1 and L2 regularisation techniques. These were the best-fitting hyperparameters to increase the model's resilience against overfitting and underfitting.

Employing Friedman's 'intelligent trial and error' approach, the hyperparameter optimization process in this paper begins with a broad search space. It was gradually narrowed down using the best found parameters as guides, resulting in a finely tuned LightGBM model. This process was made efficient by using the Hyperopt library that applies Bayesian optimization (more details in Appendix H), specifically the Tree of Parzen Estimators algorithm, to iteratively build a probabilistic model of the objective function and select the most promising parameters for evaluation (Pelikan, 1999). The final optimal parameters, which allowed a faster minimum convergence were {'learning_rate': 0.101, 'min_child_samples': 48.0, 'min_data_in_leaf': 40.0, 'num_leaves': 62.0} (Figure 7).

Figure 8 Optimisation of Hyperparameters



This configuration, shaped by Friedman's guidance and LightGBM's advanced regularization capabilities, was instrumental in achieving robust predictive performance, effectively navigating the bias-variance trade-off, and ensuring the model was neither underfitting nor overfitting the training data. After hyperparameter optimization, the performance metrics improved (Table 4).

Table 3 Fine tuning results

Metric	*Score*	*Improvement*
Mean Squared Error (MSE):	59.1855	- 18.79%
Root Mean Squared Error (RMSE):	7.6932	- 9.91%
Mean Absolute Error (MAE):	2.5857	- 0.09%
R^2	0.8858	+2.65%

Table 4 shows a notable reduction in both MSE and RMSE, which implies a decrease in large errors. However, the near-constant MAE indicates that the model's performance on smaller errors remained virtually the same, with average prediction deviations around 2.59 units. The increase in the R^2 score, from 85.93% to 88.58%, indicates an enhancement in the fit of the model's predictions to the actual values. Post-tuning, the model can explain approximately 88.58% of the variability in the target variable, marking a significant improvement from the pre-tuning stage.

Even if this method yields better performance, it's crucial to acknowledge the inherent uncertainties associated with these predictive models. For instance, the standard deviation of the RMSE was approximately 0.77, with a 95% confidence interval in the range of 7.19 to 10.31. This demonstrates that, despite the rigorous optimization process, there exists notable variability in our model's performance. Such variability could be attributed to data limitations or inherent biases in the model, underscoring the probabilistic nature of our predictive efforts. Future work should encompass strategies to minimize this variability, including procuring more diversified datasets or investigating more intricate model architectures.

This exploration of GBM options has employed a systematic approach to model understanding and comparative analysis. It has particularly focused on LightGBM for its robustness and ability to handle complex relationships. Experimenting with different hyperparameters and using Bayesian optimization to balance efficiency and thoroughness, this paper achieved an optimal model that minimised underfitting and overfitting. While presenting limitations, this rigorous approach resulted in a high-performing model that promises to provide reliable results, and significantly contribute to the research objectives.

6. Findings

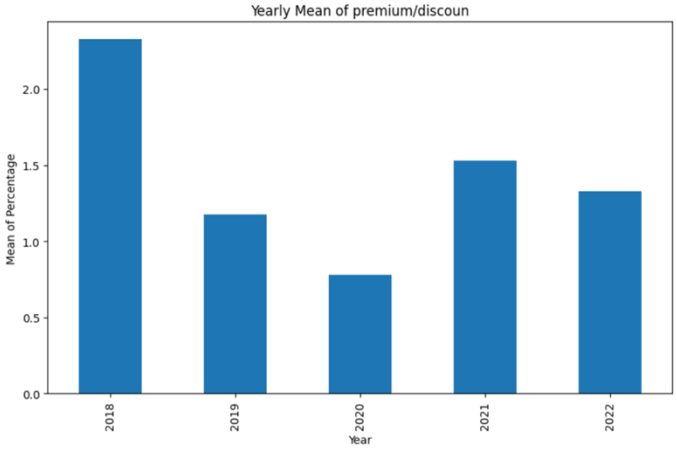
This section delves into the study's findings, keeping sight of the objectives established in the business framework. As outlined above, the methodology imposed is not merely an abstract exercise in ML but a pertinent exploration of its real-world applicability. This chapter examines the intricate mechanisms driving company valuations on a global scale, represented by the analysis of S&P 500 companies, and the specific dynamics within a single entity embodied in the study of LSEG. The section begins with the analysis of S&P 500 model results, and its feature importance. This broad market perspective sets the stage for a deeper exploration of the particular case of LSEG, which dissects once again the predictive factors. This comprehensive examination is designed to provide crucial insights that can influence strategic decision-making and aid in successfully repositioning businesses, especially those in situations similar to LSEG.

6.1 S&P 500

The examination of the S&P 500 dataset led to several interesting findings concerning the performance of different sectors, valuation trends, and their correlation with real-world events. The most prominent finding is that the S&P 500 index has consistently traded at a premium, highlighting a market value higher than an intrinsic one. This indicates a trend of optimistic market valuation, which might be attributed to market conditions, growth expectations, and investor confidence in the S&P 500 companies. Indeed, the S&P 500 index comprises leading companies in various industries, representing a significant portion of the U.S. equity market's value. As such, these companies typically exhibit robust financial performance, innovative strengths, and resilience to market volatility, which tend to attract investor confidence and contribute to optimistic market valuations. These findings are in line with the efficient markets hypothesis, and echo the expectation that passive investing through indexing yields comparable returns to active fund management [(Lo, 2007), (Clarke, 2001) (Malkiel, 2005)]. Literature states that the growth of indexing since the mid-1980s has led to a notable value premium for firms in the S&P 500 index, indicating higher demand and, consequently, elevated prices for these stocks (Morck, 2001). This phenomenon, resulting in consistent trading at a premium, affirms our model's findings and underlines the adherence of market dynamics to traditional finance theory.

A notable deviation in this upward trend is observed during the year 2020, where there was a decrease in the premium (Figure 8). This drop coincides with the onset of the COVID-19 pandemic—a period marked by widespread economic uncertainty and volatility (Dayong Zhang, 2020). The market's reaction to the pandemic reflects the susceptibility of financial valuations to unexpected global events. Post-2020, the premium has shown signs of recovery, indicating a resurgence of market optimism. This is likely fuelled by the adaptability and resilience demonstrated by businesses and the gradual recovery of the global economy.

Figure 9 S&P premium 2018-2022



Sector-wise, there was remarkable homogeneity in S&P 500's firms performance, with two exceptions: the Technology and the HealthCare sectors (Figure 9 and Table 5 below). These sectors experienced a significant rise in valuation in 2021, a year into the pandemic. This rise could be attributed to the increased relevance of health and digital solutions amidst the global crisis. Specifically, the technology-related sectors—such as Semiconductors, Software & IT, and Media & Publishing, have been consistently highly valued. The thriving demand for digital services, innovation in tech infrastructure, and the pivotal role of media in information dissemination could be driving their higher valuation. Conversely, the Financial sector remained in the lower range of valuation. This may reflect the greater regulatory scrutiny, competition, and inherent risks associated with this sector.

Figure 10 Valuation over Growth 2019-2022

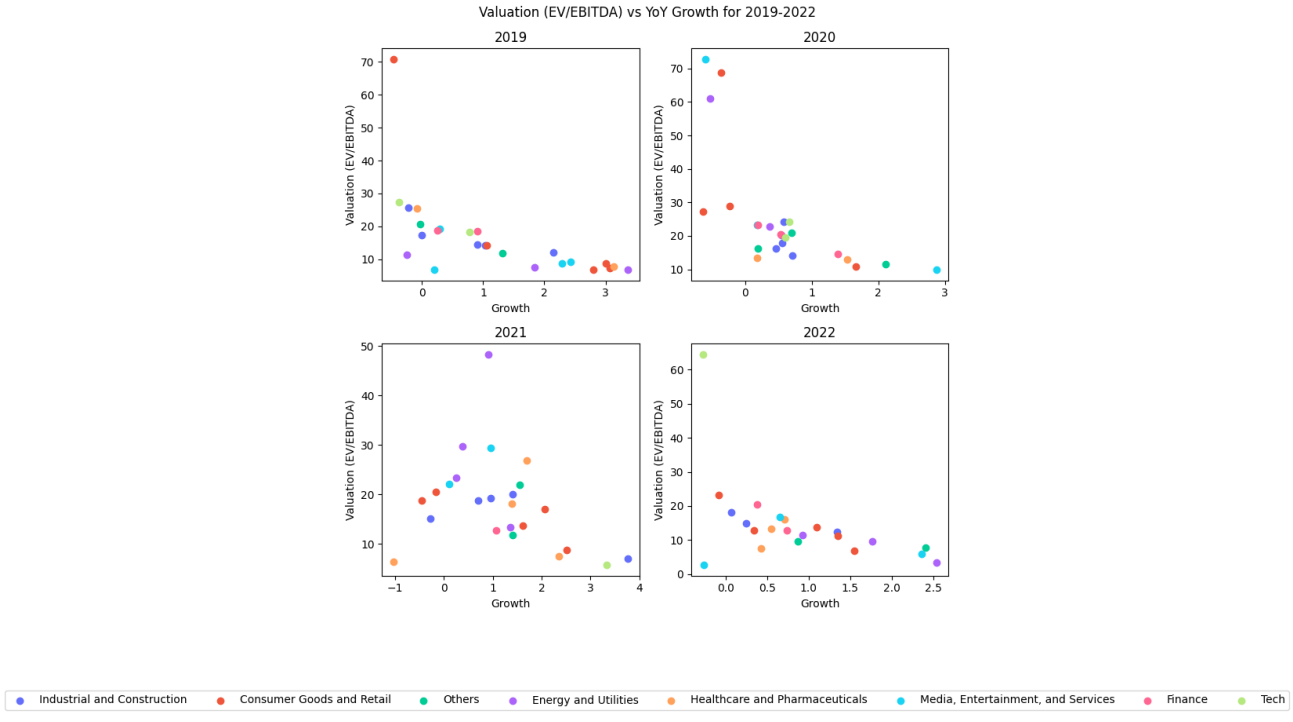


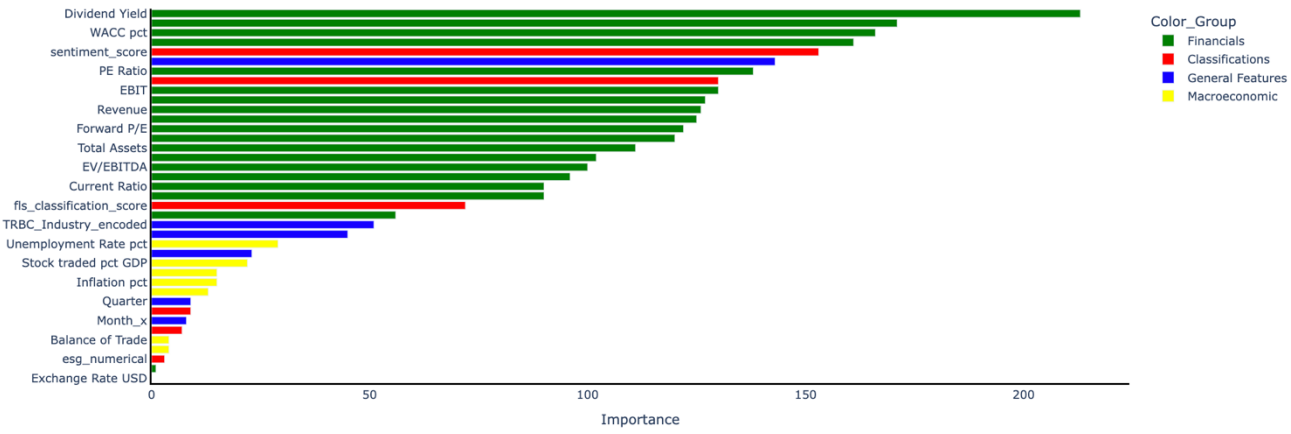
Table 4 FY0 and FY1 EV/EBITDA and Year-on-Year (YoY) Growth

Year	*FY0 EV/EBITDA*	*YoY*	*FY1 EV/EBITDA*	*YoY*
2018	13.34	/	38.78	/
2019	17.13	28.42	22.62	-43.13
2020	19.90	16.15	25.59	13.12
2021	23.99	20.52	37.06	44.78
2022	13.65	-43.16	20.14	-44.82

Understanding whether the S&P 500 is trading at a premium is essential for market analysis. It offers crucial insights into broader market sentiment, guiding investment decision-making and sheds light on potential market inefficiencies or speculative behaviour. Moreover, it establishes an overall framework and baseline for firm valuations.

Figure 10 below shows the feature importance analysis and provides valuable insights into the key determinants of a firm's financial valuation. The top-ranking features—dividend yield and WACC—are classic financial indicators recognized in the existing literature as crucial to firm valuation. Their prominence in this paper's machine learning model reaffirms the applicability of traditional financial measures in a data-driven analytical framework. These findings substantiate the relevance of traditional models, underpinned by these key variables, in predicting the EV/EBITDA ratios.

Figure 11 Feature Importance S&P500



Interestingly, the sentiment score emerged as the third most important feature, signifying the growing relevance of sentiment analysis in financial markets. This finding suggests that strong sentiments, whether positive or negative, can significantly influence firm valuation. It highlights the importance of integrating qualitative metrics, like sentiment analysis, into financial forecasting models, which

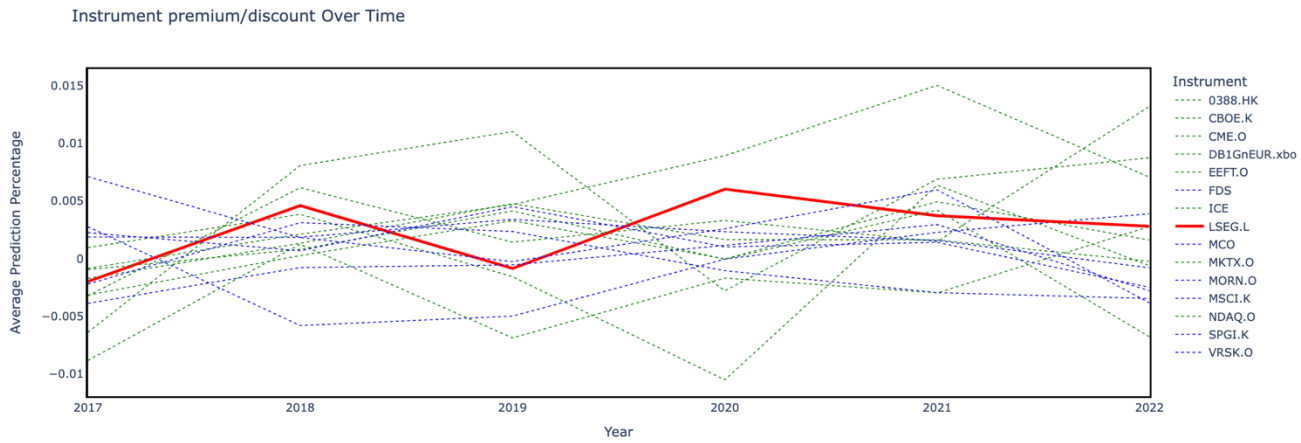
traditionally relied solely on quantitative indicators. On the other hand, the importance of the 'Instrument' feature underlines the influence of firm-specific factors in determining a firm's valuation. It suggests that given the diverse sizes and financial conditions of S&P 500 companies, firm-specific factors continue to play a significant role. This sheds light on broader research on the topic as many other aspects such as brand, reputation and perception could be investigated. While macroeconomic trends did not emerge as strong individual predictors, their collective impact is noteworthy. Although they might not hold considerable importance individually, they shape the economic landscape and indirectly influence firm valuation. This finding emphasises the need to consider a comprehensive set of factors, both macro and micro, in predicting financial valuation.

These findings support the argument that augmenting financial valuation models with sentiment and classification features derived from NLP models can enhance the predictive accuracy of forward financial valuation metrics. The use of FinBERT proved to be a valuable addition to the research framework. Moreover, it adds to the growing body of literature that emphasises the value of ML techniques in financial forecasting and decision-making. These results further establish the merit of this research in demonstrating the practical applicability of these advanced analytical techniques in enhancing traditional financial valuation methods, especially in a broad market context like the S&P 500. This novel approach paves the way for future research that could further explore integrating NLP techniques into financial valuation models, pushing the boundaries of what is possible in financial forecasting and analysis.

6.2 LSEG

The case study of LSEG constitutes a specific application of the broader trends identified in our analysis of the S&P 500 companies. The factors influencing LSEG's financial valuation are dissected, demonstrating the relevance and applicability of our machine learning model for individual firms (Figure 11).

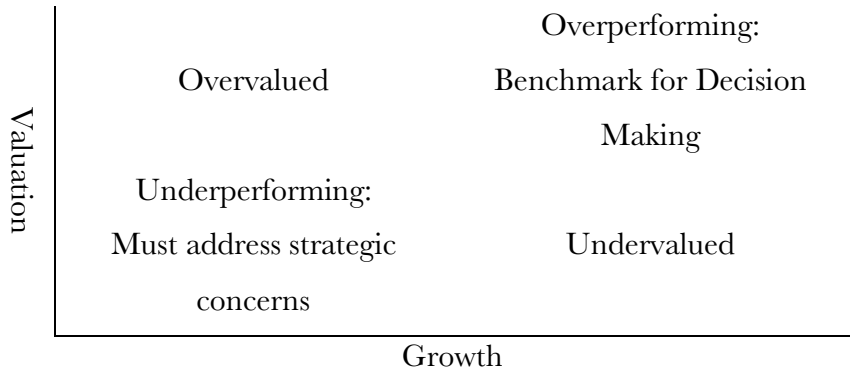
Figure 12 LSEG and peers premium/discount 2017-2022



The findings indicate that LSEG (red line) appears to be discounted relative to the chosen peers, with its intrinsic value surpassing its market value. The S&P 500 study's primary finding of the high valuation of technology-related sectors resonates with LSEG's strategic focus on technology in 2021 through the Refinitiv acquisition. While the COVID-19 pandemic had a discernible impact on the firm, 2021 posed a more significant challenge, likely due to the unexpected costs incurred from the acquisition. Notably, these costs may have prompted the market to react more harshly. Concurrently, the lower valuation attributed to the financial sector in the broader analysis bears implications for LSEG, given its role in global financial markets. These insights offer a holistic view of market trends and valuation dynamics crucial for shaping LSEG's financial strategy and defining its place in the global financial landscape.

In light of LSEG's strategic repositioning, our analysis seeks to discern LSEG's standing relative to its peers and evaluate if it aligns more closely with its heritage as an exchange provider or with its current status as a data player, following its recent strategic initiatives. When examining LSEG, its year-on-year (YoY) EV/EBITDA growth manifests similar trends to the selected peer group, reflecting a largely homogeneous market response. However, exchange providers exhibit more variability than data players, demonstrating a consistently positive trend, as highlighted in the S&P 500 analysis. The significance of this pattern becomes even more pronounced when considering valuation over YoY growth.

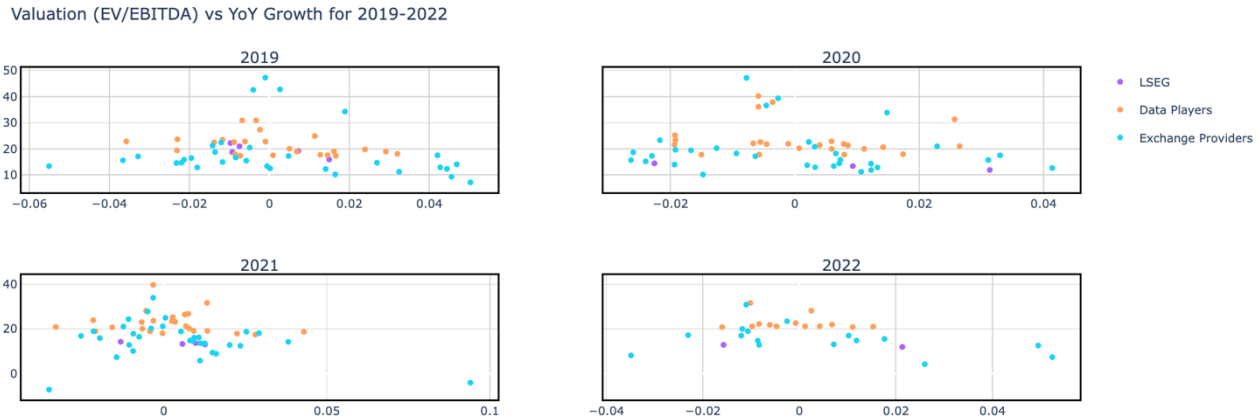
Figure 13 Valuation over Growth Matrix



Interpreting the positions of the companies on the valuation versus growth plot (Figure 12) can be extremely insightful. Companies with high value and low year-on-year growth can be considered overvalued as they demonstrate high market valuation but lack the growth potential to substantiate it. Conversely, companies with high growth but low valuation are considered undervalued, indicating that the market may not yet fully appreciate their growth potential. Firms experiencing both low growth and low valuation are underperforming and need to review their strategic plans. Those firms in the top left quadrant, with high growth and high valuation, serve as aspirational targets, providing a benchmark for strategic decision-making.

In this context, data peers predominantly appear overvalued, while exchange peers are undervalued (Figure 13). This raises an interesting question about the market's preferences. Companies like MSCI and Market Axcess, perhaps with a potential bias due to their American markets, emerge as noteworthy entities to examine. Interestingly, LSEG is grouped with exchange providers, despite its ongoing efforts to become a data player.

Figure 14 Valuation over Growth 2019-2022



The positioning of LSEG among its peers (Figure 13) suggests several hypotheses that merit further exploration. Firstly, the market has yet to fully appreciate LSEG's strategic transformation from a traditional exchange provider to a tech-oriented firm, even after significant investments and strategic acquisitions. LSEG's future growth and valuation will likely be driven by its data and analytics segment, and the analysis suggests that its strategic initiatives will contribute to its long-term revenue potential. Nonetheless, although the repositioning promises scalability, its impact on the financials are not yet evident.

Secondly, it is important to note that LSEG is not alone in its shift towards a tech-centric model. Many other exchange providers follow a similar trajectory, focusing on recurring revenue and diversification through technology-driven M&A activities, despite possible short-term volatility-driven trading. Deutsche Boerse, for instance, has been involved in ten technology-driven acquisitions in recent years aimed at augmenting its technological capabilities and broadening its revenue sources (Deutsche Boerse, 2022). Similarly, Euronext has implemented the same strategy. The European exchange demonstrates a strong interest in bolt-on deals, smaller acquisitions that complement existing operations rather than transforming them (EuroNext, 2022). The American market is also seeing similar trends. CBOE and Intercontinental Exchange have been quite active on the M&A front. They have been particularly interested in acquiring firms that offer alternative-asset, crypto, and geographical-trading add-ons, in a bid to bolster their services beyond their core trading and execution operations (CBOE, 2022) (ICE, 2022). A pursuit of technology-based firms is increasingly defining the M&A landscape in the exchange industry. The motivation behind these deals is twofold: capitalise on the opportunities provided by digital transformation and diversify revenue streams beyond traditional trading operations.

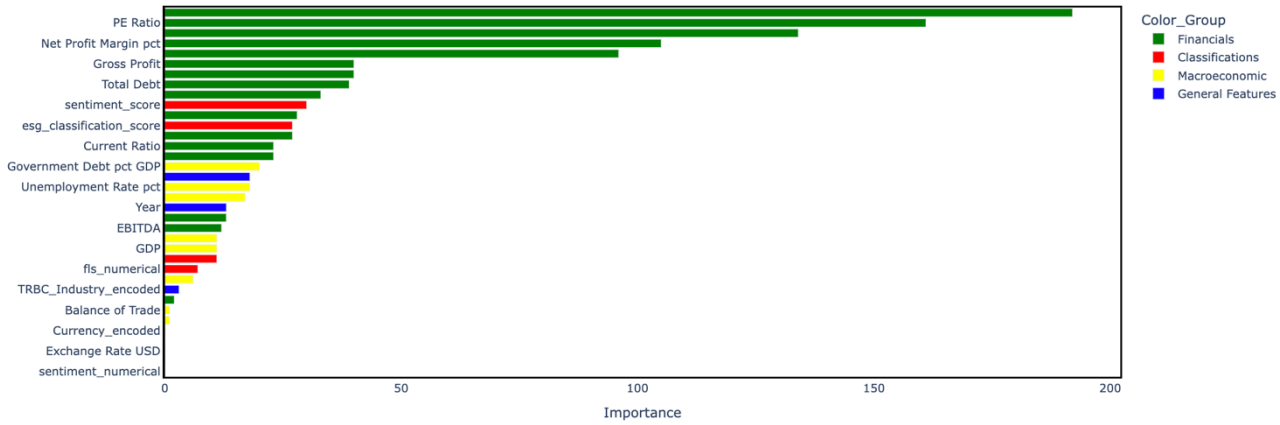
Given that LSEG's repositioning aligns with this trend, this could positively influence its valuation. The company's strategy to integrate different business segments, like linking pricing and enterprise business to Yield Book and Tradeweb, is expected to streamline data flow, improve the customer experience, and generate additional revenue synergies. On a global scale exchanges, particularly those with significant data, index, clearing, and higher-margin derivatives businesses, are likely to retain premium valuations. It is notable that LSEG's valuation multiple (EV/EBITDA) expanded recently, indicating market confidence despite the ongoing repositioning.

This widespread industry transition may blur traditional categorisations, making it more challenging to compare firms based solely on their historical market roles. As these traditionally exchange-oriented

firms continue to embrace technology, their market valuations could be influenced by both their legacy operations and their evolving tech-focused initiatives. Thus, understanding these trends is key to a more nuanced and informed analysis of LSEG's current and future market valuation. The elucidation of these trends not only underscores the need of ML analytics but also provides LSEG with insightful guidance for strategic decision-making.

The feature analysis of the most influential drivers in LSEG's valuation reveals a pattern broadly consistent with the wider market trends identified in the S&P 500 (Figure 14). Financial metrics emerged as the most significant determinant, underscoring the market and prioritizing financial health. Key factors within this category include revenue and dividend yield, which aligns with the traditional focus on profitability and return on investment in the financial sector.

Figure 15 Feature Importance LSEG and Peers



Sentiment analysis reflects the growing importance of market perception and sentiment in shaping stock valuations, even beyond the concrete financial metrics. ESG aspects were also highlighted as an essential determinant. With the rising awareness and emphasis on social responsibility, companies demonstrating strong performance in ESG aspects are increasingly recognised and valued by shareholders. Lastly, macroeconomic factors were found to have a lower yet still relevant influence. While these elements impact the broader market and all firms operating within it, their influence on individual firms' valuations can vary, and their role in LSEG's valuation reflects their universal relevance.

In conclusion, shareholders seem to be focusing on a mix of traditional financial indicators, market sentiment, ESG aspects, and macroeconomic factors when assessing LSEG. This comprehensive approach to valuation echoes the multifaceted nature of today's financial landscape and the intricate

interplay of various influencing factors. In light of these findings, it becomes evident that including unstructured data can significantly enrich financial analysis and understanding of market trends. These results paves the way for researchers and financial analysts to explore beyond traditional metrics and delve into a more comprehensive and nuanced approach to valuation.

7. Discussion: broader business implications

7.1 General implications

The findings of this paper hold significant implications across both the data science and financial sectors, spanning from wide-reaching industry trends to more specific, targeted applications. These insights carry weight in both academic discussions and practical industry applications. The importance of the broader analysis from a data science perspective cannot be understated. By ensuring that the GBM aligns with established financial theories and employs the same features as traditional valuation models, we can ensure the accuracy and reliability of predictive models. On a business level, this analysis illuminates the broader market context and identifies overarching macro trends. Understanding these market shifts is essential for contextualizing individual firm results, such as the prevalence of high valuations in the tech sector.

Furthermore, the case-specific analysis of LSEG holds significant value from both a data science and business perspective. It illustrates that LightGBM maintains its accuracy and relevance not only in broad market scenarios but also in firm-specific situations. This allows us to capture aspects that traditional valuation methods may overlook and deliver answers to intricate business questions. The paper's findings suggest that LSEG, despite its ongoing repositioning, is still largely considered as an exchange provider and is undervalued compared to its peers. However, the company seems to be making headway, with its positioning gradually aligning more closely to data-centric businesses.

This research marks a notable expansion in applying GBMs for financial valuation compared to previous studies. Traditionally, structured financial metrics have been the primary focus in this field. However, this paper emphasizes the significant role unstructured data can play in understanding financial trends and contributing to decision-making processes. This reinforces the growing recognition of alternative data's value in providing nuanced insights that standard financial metrics may not fully capture. Ultimately, this study contributes to broadening the scope of financial analysis and valuation, making it more comprehensive and informative.

The work presented here helps bridge the gap between the fast-paced evolution characterizing the financial industry and traditional valuation practices. This paper posits how ML methodologies, combined with alternative data sources, can deliver an insightful and nuanced understanding of company valuation. This sets a new precedent in financial analysis, marrying cutting-edge technology with traditional financial metrics, reflecting the broader transformations happening in the industry.

7.2 Limitations

This study, like all, is not without its limitations. From a data science perspective, adding additional unstructured data or alternative financial metrics might have yielded differing results. Including macroeconomic features inherently raises the level of uncertainty within the models. Macroeconomic indicators are subject to numerous influences and can be challenging to forecast accurately. Integrating these variables into a relative valuation model adds another layer of uncertainty, making it challenging to isolate the effects of specific macroeconomic factors in relative valuation. Moreover, alternative methods of data pre-processing or leveraging different NLP models, such as a fine-tuned FinBERT, could also have been explored, though this would require additional computational power, which was not readily available for this research.

Regarding business perspective, the study might have incorporated different indices, such as FTSE, to provide a more global outlook. An expanded market investigation beyond just the UK, Europe, and the US, could have provided different insights. In addition, according to industry experts, US equities generally hold higher valuations than their European counterparts, a phenomenon often ascribed to speculative factors, which could potentially distort comparisons. Nonetheless, despite these limitations, this study provides a unique exploration into machine learning applications in financial valuation and offers valuable insights for both the data science and financial fields. Future research could address these limitations and expand upon these findings.

7.3 Future work and recommendations

When considering future evolutions and potential areas for development, several key recommendations emerge for researchers, industry analysts and firms undergoing transformation.

As seen in this paper, the adoption and exploration of alternative data sources in research can offer a new lens through which interpret financial trends and dynamics. Combining these unconventional sources with traditional financial metrics can potentially yield richer and more nuanced insights. As the complexity of financial data grows, employing diverse ML models and techniques becomes increasingly vital for capturing and interpreting underlying market trends. Future work could explore integrating techniques such as neural networks or different NLP strategies to gain further depth in the analysis. Moreover, incorporating more varied features, such as the market's perception of M&A activities, the frequency of strategic inquiries during investor meetings, market reactions to CEO statements, or global sentiment during significant world events could further refine the model's predictive capabilities.

Financial analysts, on the other hand, should not overlook the potential value of integrating ML models into their analytical frameworks. While traditional financial analysis remains a critical pillar, supplementing these methods with machine learning can unveil hidden trends and insights which may otherwise remain obscured. Moreover, the effective leverage of available data - particularly data that cannot be succinctly encapsulated in numerical form - can augment the richness of the analysis and the understanding of complex financial dynamics.

Finally, for firms like LSEG that are in the midst of significant strategic shifts, leveraging these predictive models can yield valuable insights. These models offer the potential to gauge a firm's standing relative to peers and the broader market, and can therefore be instrumental in assessing a firm’s performance. By identifying closer comparables and shedding light on the market's perception of the firm's repositioning efforts, these models can inform strategic decisions and provide direction for future endeavours. Comprehensive analysis, incorporating both broader market and industry-specific perspectives, can be a significant differentiator in strategic planning and implementation.

In summary, the future of financial analysis lies in the confluence of traditional methods, advanced machine learning techniques, and a broad, holistic view of both market and industry. The results presented in this paper indicate the promise and potential this multi-pronged approach holds, thereby paving the way for further exploration and innovation in the field.

8. Conclusion

This study aimed to investigate how Gradient Boosting Machines, specifically LightGBM, could enhance financial valuation methods, especially regarding the valuation of companies undergoing significant transformations. The results reaffirmed the forecasting accuracy of GBMs in predicting forward financial valuation metrics such as EV/EBITDA, both at a broad market level (represented by the S&P 500) and at a firm-specific level, such as the London Stock Exchange Group. The research question sought to understand how GBMs, leveraging quantitative and qualitative metrics, extracted from diversified data sources and categorised using FinBERT, could contribute to a more comprehensive understanding of market dynamics and shareholder perception. The results demonstrated that GBMs have a significant role in financial valuation, effectively answering the research question. This study's significance lies in its ability to bridge the gap between the rapidly evolving financial industry and traditional valuation methods by applying advanced ML models and leveraging diverse data sources.

This paper's method involved the application of LightGBM, a specific architecture of GBM, to the broader market and the case of LSEG. Through meticulous analysis, it found that the models efficiently predicted the financial valuation metrics. Moreover, this study assessed LSEG's transformation and uncovered that despite undergoing a significant transition, the firm is still perceived primarily as a financial player. However, the model analysed suggests that the company's positioning is gradually aligning more closely with data-centric businesses.

The study contributes to the literature by demonstrating the use of GBMs in financial valuation and expands on the work of Geertsema and Lu by applying this approach to a company undergoing a significant transition. Inspired by the work of Stankevičienė, this paper highlighted the potential of ML and unstructured data in capturing complex dynamics with FinBERT that traditional methods might miss. Moreover, the results aligned with established financial theory, confirming the value premium for firms in the S&P 500 index and supporting previous findings.

In conclusion, this study's key results support the importance of GBMs in enhancing traditional financial valuation methods. The ability of these models to incorporate both structured and unstructured data adds a significant layer of nuance and richness to the valuation process, thus

providing a more comprehensive understanding of the complex dynamics of financial markets. The study's significance lies not just in its academic contributions but also in its strategic business impact.

In an era where data is the new oil, the financial industry's ability to leverage vast amounts of available data effectively can dictate its competitive position. This study's findings and proposed valuation approach can guide strategic decision-making to uphold a competitive standing. The research highlights the importance of continuous learning and adaptation in the financial industry. It advocates for integrating advanced ML methodologies into traditional practices, positing the need for financial institutions to stay abreast with technological advancements. As the financial landscape evolves, institutions that quickly adapt and incorporate these technologies into their strategic decision-making processes will be better equipped to navigate market uncertainties and seize emerging opportunities. Lastly, the significance of this study extends beyond its academic contributions. By introducing machine learning into financial valuation, it offers a progressive approach that enhances accuracy, ensures comprehensive analysis, and facilitates data-driven decision-making. Its findings serve as a beacon for financial institutions and decision-makers, illuminating a path towards a future where traditional finance and advanced technology coexist and enrich each other.

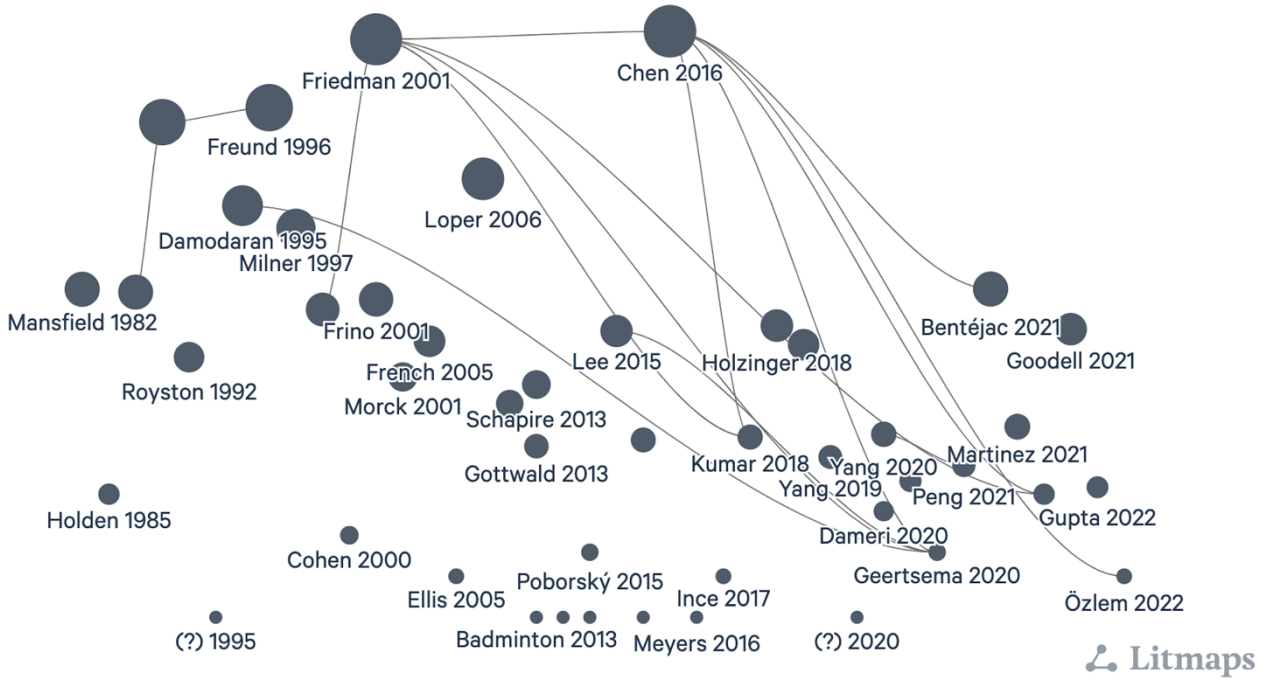
9. Appendix

The code of this paper can be found at the following [GitHub Repository]

A. LitMap

An essential aspect of this research involved conducting a comprehensive literature review. This process was not solely about studying the existing works but also understanding the intricate connections among them. To aid this endeavour, we employed LitMap, a tool that facilitated the visualization of the research field's landscape. Its ability to identify citation networks proved instrumental in illuminating how various ideas and research methodologies have evolved and influenced each other over time.

These citation networks offered invaluable insights into the most impactful works and authors within our research area. Grasping these connections allowed us to pinpoint seminal works, recurrent themes, and burgeoning trends in the field. Such understanding not only enriched our perspective but also guided our research trajectory. Moreover, the visual mapping provided by LitMap served as an effective guide through the vast corpus of literature. It ensured that our review remained systematic and all-encompassing, thereby solidifying the foundations of this research



B. DCF and relative valuation mathematical derivations

Discounted Cash Flow (DCF) is a valuation method used to estimate the value of an investment based on its future cash flows. The idea behind the DCF model is that the value of the investment is equal to the sum of its future cash flows, each discounted back to the present. (Mehta, 2020)

$$DCF = \frac{CF_1}{(1+r)^1} + \frac{CF_2}{(1+r)^2} + \cdots + \frac{CF_n}{(1+r)^n}$$

Where:

- CF1 to CFn are the cash flow of an investment for each period,
- n is the number of periods. T
- r is the interest rate or discount rate, which represents the risk and cost of capital

$$FV = DCF \cdot (1+r)^n.$$

- FV is the nominal value of a cash flow amount in a future period

$$DPV = \frac{FV}{(1+r)^n}$$

- DPV is the Discounted Present Value of the future cash flow (FV)

Where:

- EV is the Enterprise Value, which equals the Market Value of Equity (also known as Market Capitalization) plus the Market Value of Debt, minus Cash and Cash Equivalents.
- EBITDA stands for Earnings Before Interest, Taxes, Depreciation, and Amortization, which is a proxy for the company's operating cash flow.

The DCF model can be further adjusted to incorporate factors such as growth. For instance, if we assume the cash flows grow at a constant rate g, the formula becomes a model known as the Gordon Growth Model. This formula assumes that cash flows will continue to grow at a constant rate indefinitely, a somewhat strong assumption. Therefore this model is most appropriate for companies with stable growth rates. We can link the DCF (Discounted Cash Flow) method to the EV/EBITDA multiple (Relative Valuation Scaling to Standard Variables), but the process requires some assumptions. Free Cash Flow to the Firm (CF) can be approximated by EBITDA - Interest - Taxes + Depreciation and Amortization, assuming depreciation, amortization, interest, and taxes remain a

constant proportion of EBITDA. Assuming the growth rate of EBITDA is constant (g), and the terminal value is calculated using the Gordon growth model:

$$TV = EBITDA_1 * \frac{(1+g)^n}{(r-g)}$$

Assuming that the company is a going concern, and most of the company's value is derived from its terminal value (which often is the case in valuation models), the enterprise value (EV) can then be approximated as the terminal value, i.e., (EV ≈ TV). Combining these, we can express EV as a multiple of EBITDA:

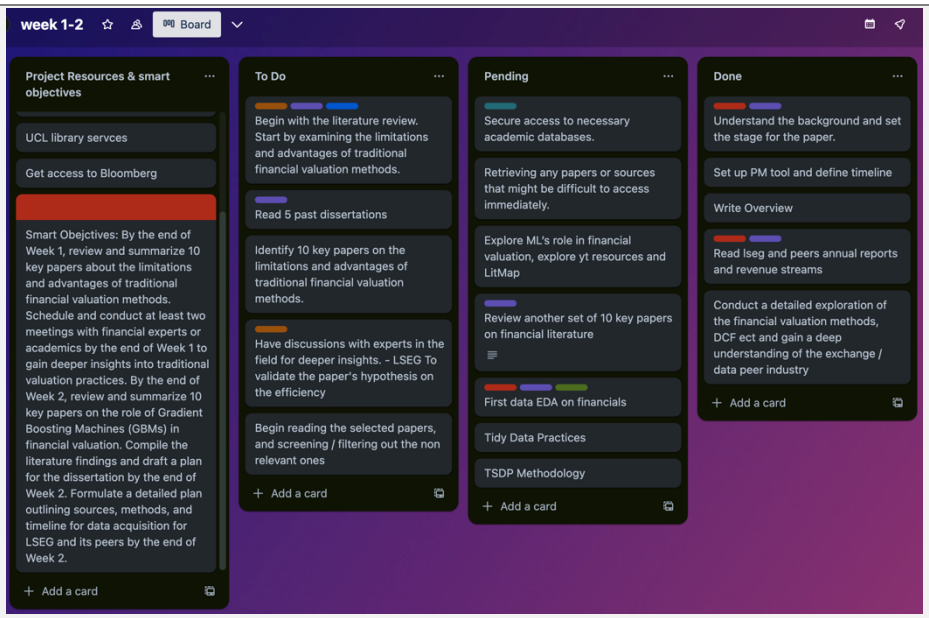
$$\frac{EV}{EBITDA} = \frac{(1+g)}{(r-g)}$$

This formula shows how the EV/EBITDA multiple can be derived from the fundamentals of the company, i.e., its cost of capital (r) and its growth rate (g). Note that this is a simplification and actual application would consider a detailed calculation of free cash flow and explicit forecast periods.

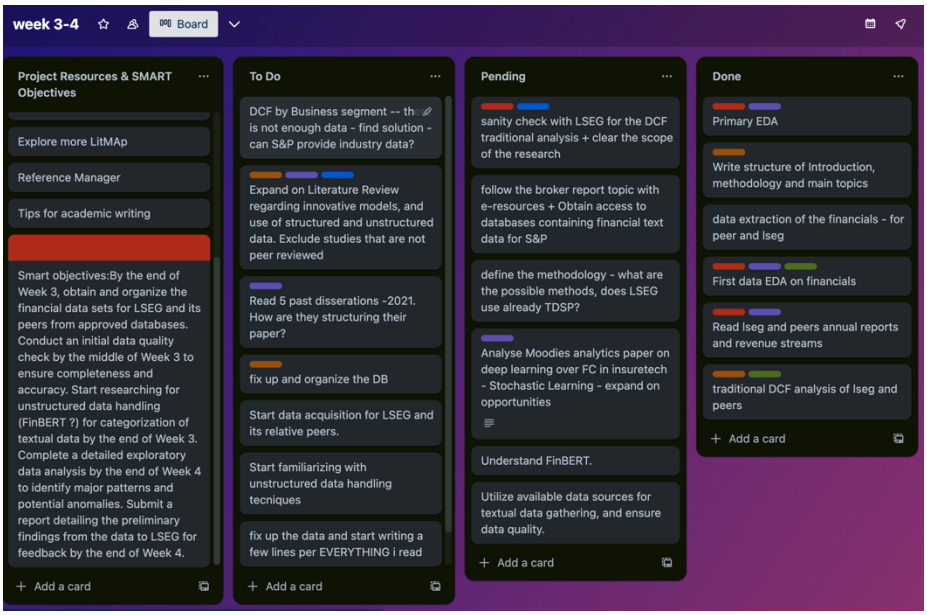
C. Project Management

Links

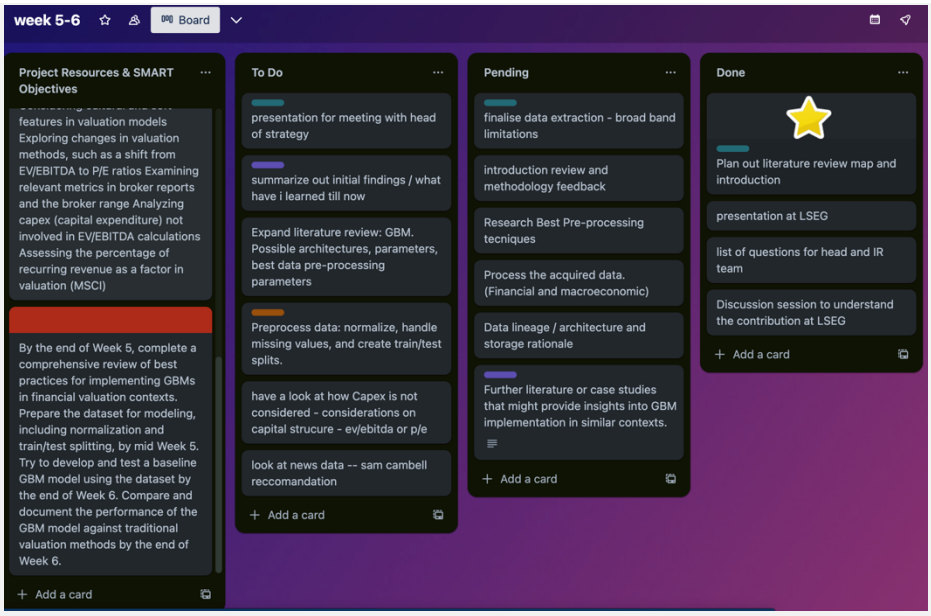
WEEKS 1-2:



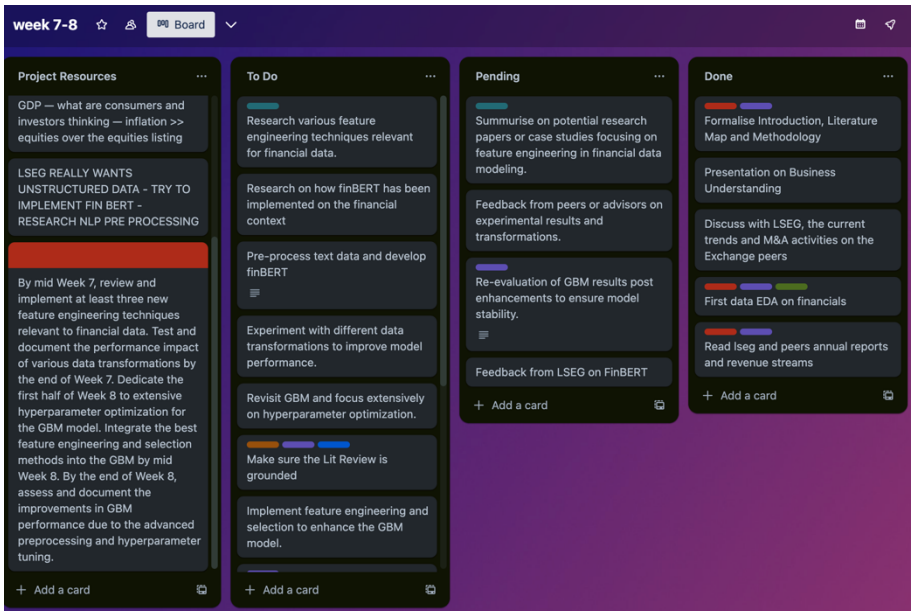
WEEKS 3-4:



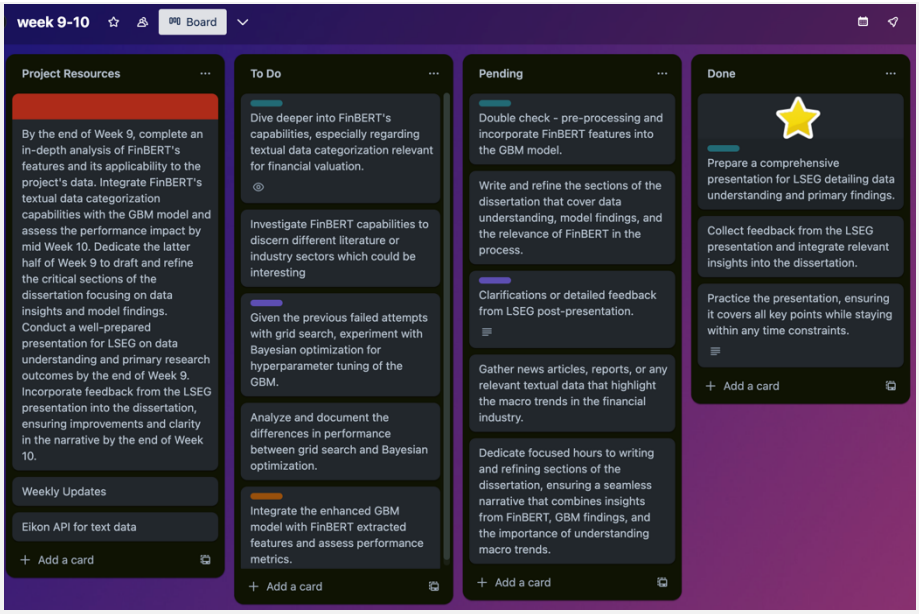
WEEKS 5-6



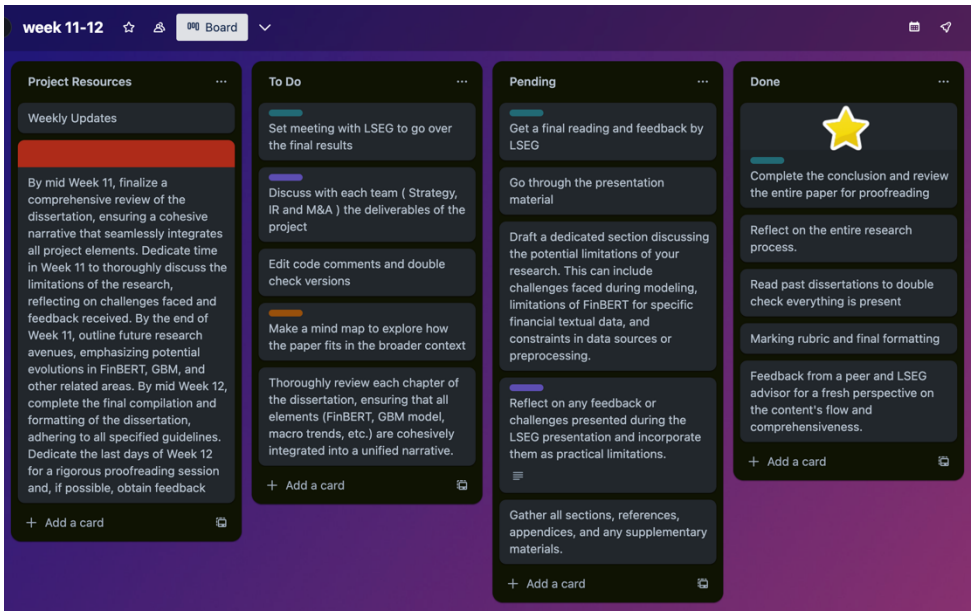
WEEKS 7-8



WEEKS 9-10:



WEEKS 11-12:



D. Business understanding, LSEG peers

The following table lists the companies identified as peers for the London Stock Exchange Group (LSEG). These companies were selected based on guidance from LSEG executives. They can be divided into two groups: Data Peers and Exchange Players.

1. Data Peers: These companies offer data services similar to LSEG.

- Moody's
- Morningstar
- S&P Global
- MSCI
- Verisk
- FactSet

2. Exchange Players: These are the leading players in the exchange market, comparable to LSEG.

- Intercontinental Exchange (ICE)
- Nasdaq
- Euronext
- CME Group
- CBOE Global Markets

- Hong Kong Exchanges and Clearing (HKEX)
- Deutsche Börse Group (DB1)
- MarketAxess

The descriptions and profiles of these companies are derived from the Eikon API, a workspace that provides detailed and updated information about companies worldwide.

Companies	REFINTIV Description
Lseg	London Stock Exchange Group plc (LSEG) is a diversified global financial markets infrastructure and data provider. LSEG operates through three divisions. Data & Analytics division provides information and data products, including indexes, benchmarks, real-time pricing data and trade reporting and reconciliation services. Data & Analytics division includes the core Refinitiv business and the FTSE Russell businesses. FTSE Russell is a global provider of index and analytics solutions. Capital Markets division provides venues/platforms for access to capital through issuance and secondary market trading for equities, fixed income, and foreign exchange (FX). Capital Markets division includes the London Stock Exchange, Tradeweb, FXall, and Turquoise businesses. Post Trade division provides clearing, risk management, capital optimization, and regulatory reporting solutions. Post Trade division consists of Over-the-counter (OTC) Derivatives, Securities & Reporting, and Non-Cash Collateral.
Moodys	Moody's Corporation is a global integrated risk assessment company. The Company operates through two segments: Moody's Investors Service (MIS), and Moody's Analytics (MA). The MIS segment publishes credit ratings and provides assessment services on a range of debt obligations, programs and facilities, and the entities that issue such obligations in markets worldwide, including various corporate, financial institution and governmental obligations, and structured finance securities. The MIS segment consists of five lines of business, which include corporate finance group; structured finance group; financial institutions group; public, project and infrastructure finance, and MIS Other. The MA segment develops a range of products and services that support the risk management activities of institutional participants in global financial markets. The MA segment consists of three lines of business, such as decision solutions, research and insights, and data and information.

Morningstar	Morningstar, Inc. is a provider of independent investment insights in North America, Europe, Australia and Asia. The Company offers a line of products and services for individual investors, financial advisors, asset managers and owners, retirement plan providers and sponsors, and institutional investors in the debt and private capital markets. It provides data and research insights on a range of investment offerings, including managed investment products, publicly listed companies, private capital markets, debt securities and real-time global market data. It also offers investment management services. It operates through its subsidiaries in 32 countries, including Australia, Brazil, Canada, Cayman Islands, Chile, China, Cyprus, Denmark, France, Germany, Hong Kong, India, Italy, Japan, Jersey, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Romania, Singapore, South Africa, South Korea, Spain, Sweden, Switzerland, Thailand, United Arab Emirates, and the United Kingdom.
S&P global	S&P Global Inc. is a provider of credit ratings, benchmarks, analytics and workflow solutions in the global capital, commodity and automotive markets. The Company operates through five segments: S&P Global Market Intelligence (Market Intelligence), S&P Global Ratings (Ratings), S&P Global Commodity Insights (Commodity Insights), S&P Global Mobility (Mobility), and S&P Dow Jones Indices (Indices). Its Market Intelligence business lines include desktop, data and advisory solutions, enterprise solutions and credit risk solutions. Its Ratings segment is a provider of credit ratings, research and analytics to investors and other market participants. Its Commodity Insights is a provider of information and benchmark prices for the commodity and energy markets. Its Mobility provides insights, forecasts and advisory services to the automotive value chain. Its Indices is an index provider maintaining a variety of valuation and index benchmarks for investment advisors and wealth managers.
MSCI	MSCI Inc. (MSCI) is a provider of critical decision support tools and solutions for the global investment community. The Company operates through four segments: Index, Analytics, ESG and Climate, and All Other-Private Assets. Its Index segment offers products, such as MSCI Global Equity Indexes, ESG and Climate Indexes, Factor Indexes, Thematic Indexes, Custom Indexes, Fixed Income Indexes and Real Estate Indexes. Its Analytics segment offers risk management, performance attribution and portfolio management content, applications and services. Its ESG and Climate segment offerings include MSCI ESG Ratings, MSCI ESG Business Involvement Screening Research, and MSCI Climate Solutions. Its All Other-Private Assets segment comprises the Real Estate and Burgiss segments. The Real Estate segment offerings include offerings include transaction data, benchmarks, return-analytics, climate assessments and market insights for tangible assets, such as real estate and infrastructure.

Verisk	Verisk Analytics, Inc. is a data analytics provider that serves insurance industry. The Company offers predictive analytics and decision support solutions to customers in rating, underwriting, claims, catastrophe and weather risk, global risk analytics, and many other fields. The Company's Insurance segment primarily serves its property and casualty (P&C) insurance customers and focuses on the prediction of loss, the selection and pricing of risk, and compliance with their reporting requirements in each United States state in which the Company operate. The Company also develop and utilize machine learned and artificially intelligent models to forecast scenarios and produce both standard and customized analytics that help its customers better manage their businesses, including detecting fraud before and after a loss event and quantifying losses. It also helps businesses and governments better anticipate and manage climate and weather-related risks.
Factset	FactSet Research Systems Inc. is a global financial data and analytics company. The Company provides financial data and market intelligence on securities, companies and industries to enable its clients to research investment ideas, as well as offering them the capabilities to analyze, monitor and manage their portfolios. The Company also offers technologies, such as a configurable desktop and mobile platform, comprehensive data feeds, cloud-based digital solutions, and application programming interfaces (APIs). Its solutions span investment research, portfolio construction and analysis, trade execution, performance measurement, risk management, and reporting across the investment lifecycle. The Company operates through three geographical segments: the Americas, EMEA and Asia Pacific. It primarily delivers insight and information through its three workflow solutions: Research & Advisory; Analytics & Trading; and Content & Technology (CTS).
ICE	Intercontinental Exchange, Inc. is a provider of marketplace infrastructure, data services and technology solutions to a range of customers, including financial institutions, corporations and government entities. The Company operates through three segments: Exchanges, Fixed Income and Data Services and Mortgage Technology. The Exchanges segment operates regulated marketplaces for the listing, trading and clearing of an array of derivatives contracts and financial securities. The Fixed Income and Data Services segment provides fixed income pricing, reference data, indices, analytics and execution services, as well as global credit default swaps (CDS), clearing and multi-asset class data delivery solutions. The Mortgage Technology segment provides a technology platform that offers customers, digital workflow tools that aim to address the inefficiencies that exist in the United States residential mortgage market, from application through closing and the secondary market.

Nasdaq	Nasdaq, Inc. is a global technology company serving the capital markets and other industries. The Company's diverse offerings include data, analytics, and software and services. The Company's Market Platforms segment includes its Trading Services and Marketplace Technology businesses. Its Trading Services business primarily includes equity derivatives trading, cash equity trading, Nordic fixed income trading and clearing and others. Its Marketplace Technology business includes its trade management services and its market technology businesses. Trade management services provides market participants with a variety of alternatives for connecting to and accessing its markets for a fee. Its Capital Access Platforms segment includes its Data & Listing Services, Index and Workflow & Insights businesses. The Company's Anti-Financial Crime segment delivers platforms that provide software as a service (SaaS) solution for fraud detection, anti-money laundering, and trade and market surveillance.
Euronext	Euronext NV is a company based in the Netherlands that serves as a parent of the Euronext pan-European exchange group. Euronext offers a diverse range of products and services, combining equity, fixed income securities and derivatives markets in Amsterdam, Brussels, Dublin, Lisbon, London and Paris. The Company's businesses comprise listing, cash trading, derivatives trading, market data and indices, post-trade and market solutions, among others. Euronext regulated markets provide a listing venue for companies seeking to raise capital and enter the Eurozone. It provides an electronic trading platform, which enables investors to place orders directly with the exchange. The Company sells real time, historic and reference data generated from the activity on the Euronext markets. It also calculates and publishes a portfolio of more than 500 benchmark indices, including the AEX-Index and CAC 40 Index. The Company offers technology solutions and services to exchanges and market operators.
CME Group	CME Group Inc. provides products across all asset classes, by trading futures, options, cash and over-the-counter (OTC) products. The Company offers a range of interest rates, equity indexes, foreign exchange (FX), agricultural commodities, energy and metals. It also offers cash and repo fixed income trading through BrokerTec, and cash and OTC FX trading through electronic broking services (EBS). The Company's operations comprise of the businesses of Chicago Mercantile Exchange Inc. (CME), the Board of Trade of the City of Chicago, Inc. (CBOT), New York Mercantile Exchange, Inc. (NYMEX) and Commodity Exchange, Inc. (COMEX) and its cash markets business. In addition, it operates central counterparty clearing houses. The Company offers clearing, settlement and guarantees for all products cleared through the clearing house. It provides clearing and settlement services for a range of exchange-traded futures and options on futures contracts and OTC derivatives.

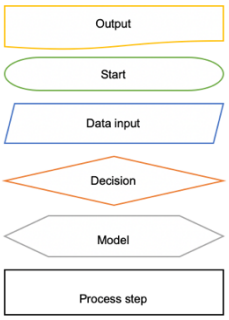
Cboe	Cboe Global Markets, Inc. is a provider of market infrastructure and tradable products, delivering trading, and clearing and investment solutions to market participants. Its segments include Options, North American Equities, Europe and Asia Pacific, Futures, Global FX., and Digital. The Options segment includes options on market indices (index options), as well as on the stocks of individual corporations (equity options). The North American Equities segment includes listed United States equities transaction services. The Europe and Asia Pacific segment includes the pan-European listed equities and derivatives transaction services, exchange-traded products (ETPs), exchange-traded commodities, and international depository receipts. The Futures segment includes transaction services provided by its electronic futures exchange, Cboe Futures Exchange, LLC (CFE). The Global FX segment includes institutional foreign exchange (FX) trading services. The Digital segment includes Cboe Digital.
Hkex	Hong Kong Exchanges and Clearing Limited is principally engaged in the operation of stock exchanges. The Company operates through five business segments. The Cash segment includes various equity products traded on the Cash Market platforms, the Shanghai Stock Exchange and the Shenzhen Stock Exchange. The Equity and Financial Derivatives segment includes derivatives products traded on Hong Kong Futures Exchange Limited (HKFE) and the Stock Exchange of Hong Kong Limited (SEHK) and other related activities. The Commodities segment includes the operations of the London Metal Exchange (LME). The Clearing segment includes the operations of various clearing houses, such as Hong Kong Securities Clearing Company Limited, the SEHK Options Clearing House Limited, HKFE Clearing Corporation Limited, over the counter (OTC) Clearing Hong Kong Limited and LME Clear Limited. The Platform and Infrastructure segment provides users with access to the platform and infrastructure of the Company.
Deutsche börse group (DB1)	Deutsche Boerse AG is a Germany-based exchange organization and an integrated provider of products and services covering the process chain of securities and derivatives trading. The Company offers listing and trading services and operates the trading platforms Xetra and Frankfurter Wertpapierboerse. It also provides clearing services through Eurex Clearing, as well as post-trade banking, settlement and custody services for fixed-income securities, shares and investment funds. In addition to that, it offers market data and technology-based services, such as data feeds, market data, news services, reference data, reporting services, external information technology (IT) services and trading infrastructure. The Company also operates through DB1 Ventures, a corporate venture capital platform that offers capital, knowledge, guidance, and connectivity to its portfolio companies.

Marketaxess	MarketAxess Holdings Inc. is engaged in operating electronic platforms for the trading of fixed-income securities and provides related data, analytics, compliance tools and post-trade services. It provides Open Trading marketplace, which offers an all-to-all trading solution in the global credit markets, creating a liquidity pool for a range of credit market participants. It is involved in drawing on a diverse set of trading protocols, including request-for-quote, live order books, sessions-based trading and portfolio trading solutions, as well as its deep data and analytical resources. It provides integrated and actionable data offerings, including Composite+ and Axess All real time pricing to assist clients with trading decisions and transaction cost analysis. The Company offers a range of post-trade services, including straight-through processing, trade matching, trade publication, regulatory transaction reporting and market and reference data across fixed-income and other products.

E. Data Acquisition, storage and cleaning

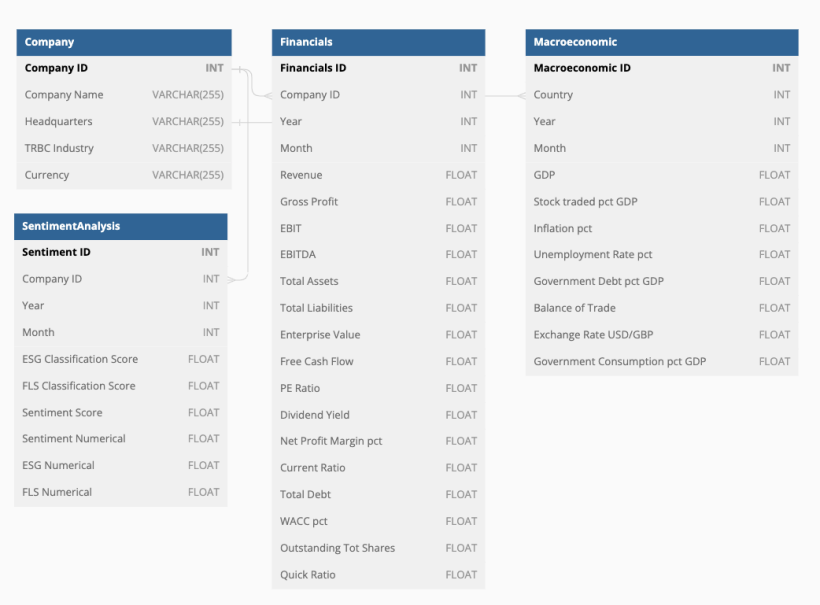
E.1 Project Process Legenda

Figure 4: Project Process Flowchart - The diagram visually represents the sequential steps involved in the methodology employed for this study. It illustrates how each stage leads to the next, offering a comprehensive overview of the research process. Here below is the legenda on each step.



E.2 Project rational:

The data is structured using illustrated schema to enable a wide range of data analysis. The chosen data schema aligns with industry standards and facilitate efficient data organization and querying for the ML task. Defining clear and simple data schemas makes aggregating, filtering, and analysing data across different dimensions and metrics easier, enabling valuable insights.



E.3 Transformations:

Gradient Boosting Machines are learning algorithms that are particularly useful due to their ability to handle heterogeneous features and resilience to outliers. Nonetheless, given the diverse dataset, feature transformation where experiment as may still be beneficial to allow the model to perform better and converge faster.

This analysis applies a range of feature transformation techniques to specific features based on their distribution characteristics, particularly skewness. Features with high positive skewness, such as 'Revenue', among others, were log-transformed, a technique that compresses the scale of the variable and reduces skewness. The log transformation is particularly effective on features that span several orders of magnitude, helping to tame their range and making them more manageable for GBMs. Meanwhile, features with moderate positive skewness, such as 'Dividend Yield' were subject to a square root transformation. This transformation is a milder form of compression that reduces positive skewness and brings the distribution closer to normal. This may help GBMs capture the patterns in these features more effectively. For features with negative skewness, such as 'Country_encoded', a power transformation was employed. By squaring each value in these features, this methodology aimed to spread out values close to zero and pull in values far from zero, improving the symmetry of the distribution. Lastly, for variables with complex distributions, the Yeo-Johnson transformation was applied, which can handle both positive and negative values.

This trials by numerous feature transformations, aimed at refining the dataset to maximize its compatibility with Gradient Boosting Machines (GBMs). Despite exhaustive trials, these transformations, unfortunately, did not lead to improvements in model performance or acceleration in convergence speed. Future research could provide additional insights by exploring various other types of transformations and building on the foundation established in this paper.

F. Variables used in the models

The model has been trained on a comprehensive dataset encompassing four types of features: general qualitive features, financial fundamentals, NLP metrics (extracted by FinBERT) and macroeconomic metrics.

1. General qualitative

These are features encoded from categorical variables that provide general information about each data point:

- Instrument_encoded: Coded representation of the financial instrument being traded.
- Country_encoded: Encoded representation of the country of origin for the instrument.
- TRBC_Industry_encoded: Encoded representation of the Thomson Reuters Business Classification (TRBC) industry category.
- Currency_encoded: Coded representation of the currency used.
- Year: The year of the data point.
- Month: The month of the data point.

2. Financial Fundamentals

These are quantitative features that provide a fundamental analysis of the financial performance of the company:

- Revenue: The total revenue generated by the company.
- Gross Profit: The profit a company makes after deducting the costs associated with making and selling its products or providing its services.
- EBIT: Earnings before interest and taxes.
- EBITDA: Earnings before interest, taxes, depreciation, and amortization.
- Total Assets: The total value of all assets owned by the company.
- Total Liabilities: The total amount of debt and financial obligations owed by the company.
- Enterprise Value: The total value of a company, including market capitalization, debt, and cash.

- Free Cash Flow: The cash a company is able to generate after accounting for capital expenditures.
- PE Ratio: Price-to-earnings ratio.
- Dividend Yield: The dividend yield or dividend-price ratio of a share.
- Net Profit Margin pct: The net profit margin as a percentage.
- Current Ratio: The current ratio, a liquidity ratio that measures a company's ability to pay short-term obligations.
- Quick Ratio: A measure of a company's ability to cover its short-term liabilities with its most liquid assets.
- Total Debt: The total amount of debt the company has.
- WACC pct: The weighted average cost of capital as a percentage.
- Outstanding Tot Shares: The total number of outstanding shares of the company.

3. NLP metrics

These are features extracted from Natural Language Processing (NLP) of company-related text data:

- `esg_classification_score`: Score derived from the classification of environmental, social, and governance (ESG) factors.
- `fls_classification_score`: Score derived from a custom classification task.
- `sentiment_score`: Overall sentiment score derived from text data.
- `sentiment_numerical`: Numerical representation of the sentiment.
- `esg_numerical`: Numerical representation of ESG factors.
- `fls_numerical`: Numerical representation of custom classification.

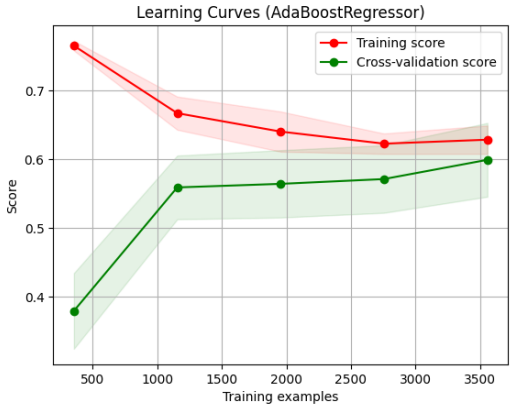
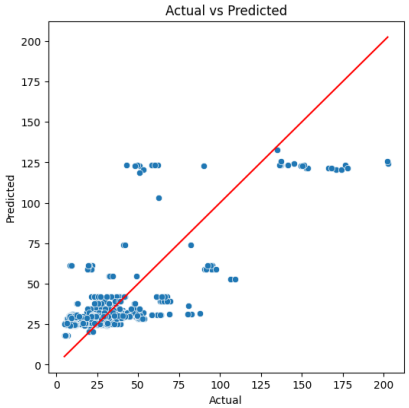
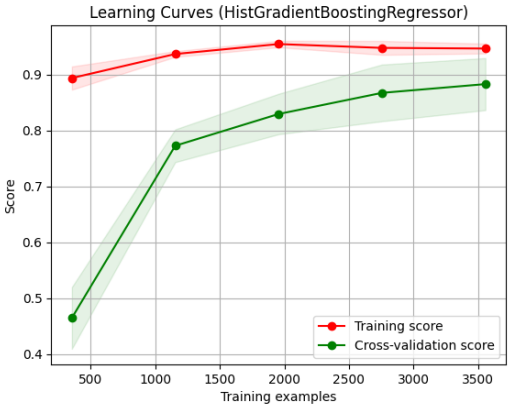
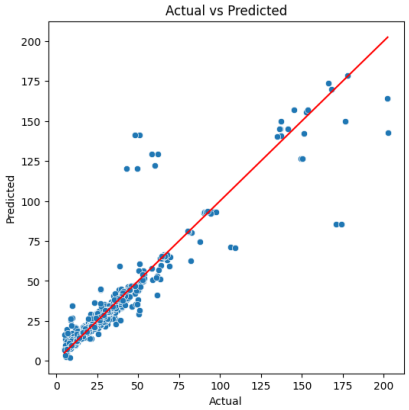
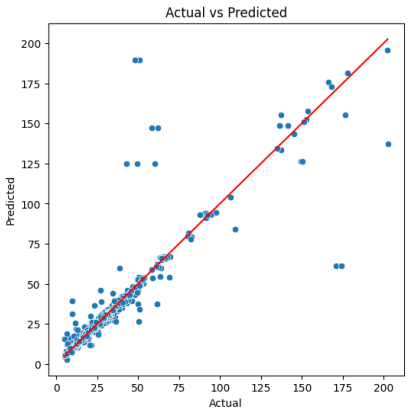
4. Macro metrics

These are broader macroeconomic indicators:

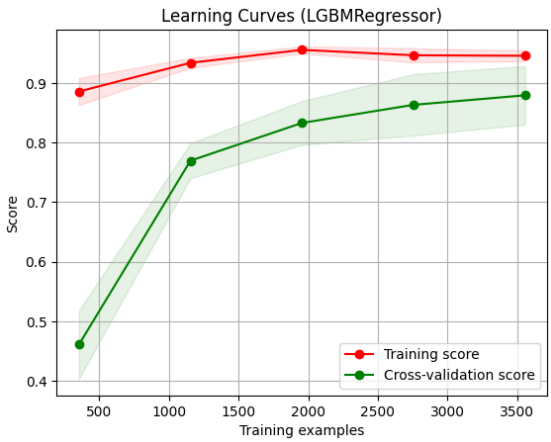
- GDP: Gross Domestic Product.
- Stock traded pct GDP: The total value of shares traded as a percentage of GDP.
- Inflation pct: The rate of inflation.
- Unemployment Rate pct: The unemployment rate as a percentage.
- Government Debt pct GDP: Government debt as a percentage of GDP.
- Balance of Trade: The difference between a country's exports and imports.
- Exchange Rate USD/GBD: The exchange rate between USD and GBP.

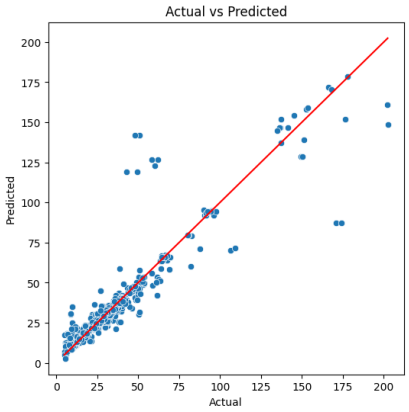
- Government Consumption pct GDP: Government consumption expenditures as a percentage of GDP.

G. Model evaluation plots

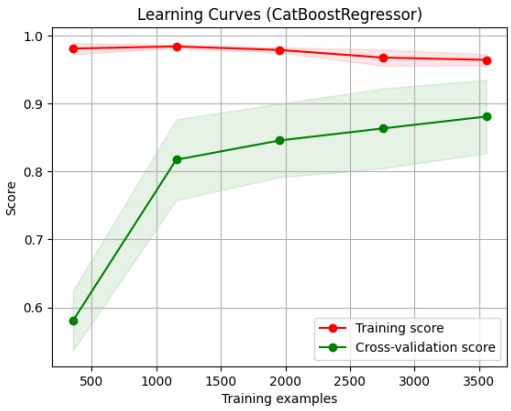
Model	Learning curves	Actual vs Predicted
Adaboost		
Histgradient boost		
XgBoost		

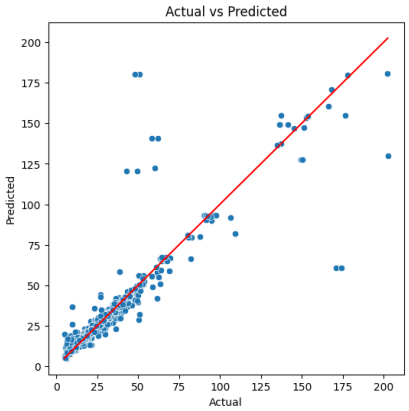
LightGBM





CatBoost





H. Hyperparameter tuning selection: Bayesian optimisation

Hyperparameter tuning is a pivotal process in designing ML models. It involves configuring the model's settings to ensure robust and generalizable predictive performance. The quality of a model's outcome can drastically differ based on these tunings, making this process a vital phase of the study.

Initially, traditional hyperparameter optimization methods, Grid Search and Random Search, were explored in this study. However, both presented limitations that prevented efficient tuning. Grid Search's strength lies in its exhaustiveness. It methodically works through multiple combinations of hyperparameters to find the optimal mix. However, this thoroughness becomes a limitation when dealing with large hyperparameter spaces. The computational cost became prohibitive as it increased proportionally with the number of hyperparameter combinations. This challenge was particularly relevant in the study, due to the high dimensionality of the hyperparameters studied and constrained computational resources. On the other hand, Random Search alleviated some of these challenges by randomly sampling the search space, enabling it to handle high-dimensional spaces more efficiently than Grid Search. Despite this advantage, it often lacked precision. The random sampling method

provided no guarantee of locating the optimal or near-optimal solution within a reasonable timeframe, leading to potential compromises on the quality of the resulting model.

Faced with these limitations, the research attention was shifted to an alternative strategy: Bayesian Optimization, specifically leveraging the Tree of Parzen Estimators (TPE) algorithm. Bayesian Optimization is a sequential design strategy for global optimization of 'black-box' functions, meaning functions without explicit form. It applies Bayesian methods to model uncertainty, cleverly navigating the trade-off between exploration of the search space and exploitation of the existing knowledge. This technique overcame the limitations encountered with Grid Search and Random Search.

In summary, the choice of Bayesian Optimization using TPE was not merely a matter of preference but driven by practical challenges and efficiency considerations. It proved to be a more effective and efficient solution for our hyperparameter tuning tasks, delivering more promising results than other methods could.

10. Bibliography

- Özlem, Ş. T. (2022). Predicting cash holdings using supervised machine learning algorithms. *Financ Innov*, 8-44.
- Al Daoud, E. (2019). Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. *International Journal of Computer and Information Engineering*, 6-10.
- Aparicio, C. a. (2020). *What is the Team Data Science Process?* Microsoft.
- Bentéjac, C. C.-M. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review, 54*, 1937-1967.
- Bird, S. (2006). NLTK: the natural language toolkit. *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, (pp. 69-72).
- CBOE, C. B. (2022). *Annual Report.*
- Chan, R. W. (2010). Gaining insight with the ev/ebitda ratio. *Better Investing 60.3*, 27-28.
- Charles M.C. Lee, P. M. (2015). Search-based peer firms: Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics 116*, 410–431.
- Chen, Q. (2021). Stock Movement Prediction with Financial News using Contextualized Embedding from BERT. *arXiv: Statistical Finance*, JEL Classification: C67, G11, G14.

- Chen, T. a. (2016). Xgboost: A scalable tree boosting system. *In Proceedings of the 22Nd ACM SIGKDD* (pp. pages 785–794). New York, NY, USA. ACM: International Conference on Knowledge Discovery and Data Mining, KDD '16.
- Clarke, J. J. (2001). The efficient markets hypothesis. *Expert financial planning: Advice from industry leaders, 7(3/4)*, 126-141.
- Clayton, A. (2020). Deep-Learning the Cash Flow Model. *Moody's Analytics*.
- Cohen, R. D. (2000). An Objective Approach to Relative Valuation. *SSB Citi Asset Management Group*, UK.
- Dameri, G. R. (2020). Neural Networks in Accounting: Clustering Firm Performance Using Financial Reporting Data. *The Journal of Information Systems, 34(2)*, 149–166.
- Dameri, G. R. (2020). Neural Networks in Accounting: Clustering Firm Performance Using Financial Reporting Data. *The Journal of Information Systems, 34(2)*, 149–166. .
- Damodaran, A. (2002). Relative valuation. *Investment Valuation*.
- Damodaran, A. (2012). *Investment valuation: Tools and techniques for determining the value of any asset. .* John Wiley & Sons.
- Dayong Zhang, M. H. (2020). Financial markets under the global pandemic of COVID-19. *Finance Research Letters,*, https://doi.org/10.1016/j.frl.2020.101528.
- Deutsche Boerse. (2022). *Annual Report.* IR Team.
- Dixon M.F., H. I. (2020). *Machine Learning in Finance: From Theory to Practice.* Springer International Publishing.
- EuroNext. (2022). *Quartely Financial Information* . Investor Relations.
- Flood M.D., J. H. (2016). Big data challenges and opportunities in financial stability monitoring . *Financ. Stab. Rev. (20)*, 129-142.
- Frecka, T. J. (1983). The effects of outliers on the cross-sectional distributional properties of financial ratios. *Accounting Review*, 115-128.
- French, G. (2005). Discounted cash flow: accounting for uncertainty. *Journal of Property Investment & Finance, 23(1)*, 75-89.
- Freund, Y. a. (1996). Experiments with a new boosting algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning (ICML 1996)* (pp. 148–156). In Saitta, L., editor, Morgan Kaufmann.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist., 29(5)*, 1189–1232.
- Frino, A. &. (2001). Tracking S&P 500 index funds. *Journal of portfolio Management*, 28(1).

- Glosten, L. R. (1994). A contingent claim approach to performance evaluation. *Journal of Empirical Finance, 1(2)*, 133-160.
- Goodell, J. W. (2021). Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance,*, 32.
- Gottwald, R. (2012). The use of the P/E ratio to stock valuation. *European Grants Projects Journals*, 21-24.
- Gupta, A. C. (2022). Automated news summarization using transformers. *In Sustainable Advanced Computing: Select Proceedings of ICSAC 2021* (pp. 249-259). Singapore: Springer Singapore.
- Hüseyin, İ. N. (2017). Investment valuation analysis with artificial neural networks. *Doğuş Üniversitesi Dergisi*, 85-96.
- Halder, S. (2022). FinBERT-LSTM: Deep Learning based stock price prediction using News Sentiment Analysis . *ArXiv*, DOI 10.48550/ARXIV.2211.07392.
- Hancock, J. &. (2021). Leveraging lightgbm for categorical big data. *IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)*, (pp. 149-154). IEEE.
- Holzinger A., K. P. (2018). Current advances, trends and challenges of machine learning and knowledge extraction: From machine learning to explainable AI. *Machine Learning and Knowledge Extraction, Springer International Publishing,*, 1-8.
- Huang, A. W. (2020). Finbert—a large language model approach to extracting information from financial text. *Available at SSRN 3910214*.
- ICE, I. E. (2022). *Annual Report.*
- Jäger, S. A. (2021). A Benchmark for Data Imputation Methods. *Frontiers in big Data*, 4, 693674.
- Kearns, M. (1988). Thoughts on hypothesis boosting. *Unpublished*.
- Kearns, M. a. (1989). Crytographic limitations on learning boolean formulae and finite automata. *In Proceedings of the Twenty-first Annual ACM Symposium on Theory of Computing*, (pp. pages 433–444). New York, NY, USA. ACM:STOC '89,
- Kruschwitz, L. &. (2006). *Discounted cash flow: a theory of the valuation of firms.* . John Wiley & Sons.
- Kumar, M. e. (2018). Parallel architecture and hyperparameter search via successive halving and classification. *arXiv preprint*, arXiv:1805.10255.
- Liapis CM, K. A. (2023). Investigating Deep Stock Market Forecasting with Sentiment Analysis *Entropy; 25(2):219*, https://doi.org/10.3390/e25020219.
- Lo, A. W. (2007). *Efficient markets hypothesis.*
- LSEG. (2023). *Annual Report* . London: Investor Relations.

- LSEG. (December 12, 2022). *Press Releases.* London: LSEG Investor Relations, Media Centre.
- Malkiel, B. G. (2005). Reflections on the efficient market hypothesis: 30 years later. *Financial review, 40(1)*, 1-9.
- Mansfield, E. R. (1982). Detecting multicollinearity. *The American Statistician, 36(3a),* 158-160.
- Martinez, I. V. (2021). Data Science Methodologies: Current Challenges and Future Approaches, *Big Data Research*.
- Milner, R. T. (1997). *The definition of standard ML: revised.* MIT Press.
- Morck, R. &. (2001). The mysterious growing value of S&P 500 membership. *NBER- National Bureau of Economic Research*, WORKING PAPER 8654 DOI 10.3386/w8654.
- Nel, W. S. (2015). An Optimal Peer Group Selection Strategy for Multiples-Based Modelling in the South African Equity Marketv. *Journal of Economics and Behavioral Studies 7.3 (J)*, 30-46.
- Ohlson, J. A. (1995). Earnings, Book Values, and Dividends in Equity Valuation. *Contemporary Accounting Research*, https://doi.org/10.1111/j.1911-3846.1995.tb00461.x
- P. Geertsema, H. L. (2023). Relative Valuation with Machine Learning. *Journal of Accounting Research 61.1*, 329-376.
- Pelikan, M. G.-P. (1999). BOA: The Bayesian optimization algorithm. *In Proceedings of the genetic and evolutionary computation conference GECCO-99 (Vol. 1, No. 1999).*
- Peng, B. C. (2021). Is domain adaptation worth your investment? Comparing BERT and FinBERT on financial tasks. *In Proceedings of the Third Workshop on Economics and Natural Language Processing.*
- Poborský, F. (2015). Fundamentals of the Liquidation Method of Business Valuation. *Procedia Economics and Finance*, 386-393.
- Qian, Y. (2006). K-means algorithm and its application for clustering companies listed in Zhejiang province. *WIT Transactions on Information and Communication Technologies*, 1-37.
- Ridgeway, G. (1999). The state of boosting. *Computing science and statistics*, 172-181.
- Royston, P. (1992). Which measures of skewness and kurtosis are best? *Statistics,* 11(3), 333-343.
- Schapire, R. E. (1990). The strength of weak learnability. *Mach. Learn., 5(2):*, 197– 227.
- Schapire, R. E. (2012). Boosting: Foundations and Algorithms. *The MIT Press.*
- Stankevičienė, J. (2012). Methods for valuation of restructuring impact on financial results of a company. *Kaunas University of Technology -Financial Economics*, DOI: http://dx.doi.org/10.5755/j01.em.17.4.2990.
- Weiss, S. M. (2010). ext mining: predictive methods for analyzing unstructured information. . *Springer Science & Business Media.*

- X. Yang, S. M. (2019). Novel Financial Capital Flow Forecast Framework Using Time Series Theory and Deep Learning: A Case Study Analysis of Yu'e Bao Transaction Data,. *IEEE, Access, vol. 7* , 70662-70672.
- Yang, Y. U. (2020). Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.
- Zellner, A. (1981). Philosophy and objectives of econometrics. *Macroeconomic analysis: Essays in macroeconomics and econometrics.*
- S&P Dow Jones Indices, S. G. (2022). *S&P 500®.* S&P Global.