Data Pipeline to Create a Comprehensive Dataset and Evaluation of the Effect that Corporate Governance has on Company Performance

Lucas Amar

4th Year Project Report
Artificial Intelligence and Computer Science
School of Informatics
University of Edinburgh

2024

Abstract

When focusing on job applications, I began diving deeper into companies and their employees on LinkedIn. This led me to ponder whether all the criteria that companies focus on when hiring affect their success. I continued by exploring whether managerial decisions and internal company structures could also play pivotal roles in success or if contrary to popular beliefs, they are not as significant as one might expect.

More and more, I was coming across the term ”corporate governance” and upon further research, I realized it is a term that includes everything I was looking at.

With this project I have built, and automated a data pipeline to combine various sources of information creating a structured and comprehensive dataset on corporate governance. I then used this to predict the year-on-year (YoY) change in return on assets (ROA) for companies in the S&P 400, S&P 500, and S&P 600, also known as the S&P 1500. Through some experimentation through random forests, gradient boosting machines, support vector machines, extreme gradient boosting, logistic regressions, and convolutional neural networks, we explored the relationship between the corporate governance and ROA changes. Although models like support vector machines, and logistic regressions were not very effective, we were able to get satisfactory results from convolutional neural networks (CNNs), with the best model reaching an accuracy of 25.60% as opposed to the 14% achieved by selecting the ROA change randomly (7 classes).

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Lucas Amar*)

Acknowledgements

With this dissertation, I would like to thank my supervisor Felipe Costa Sperb for continued guidance throughout the past year.

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Problem Statement Key Points . . . 2
- 1.3 Project Aims . . . 4

2 Literature Review **5**
- 2.1 Governance and Impact on Stock Market . . . 5
- 2.2 Key Elements of Governance . . . 6
  - 2.2.1 CEO Background . . . 6
  - 2.2.2 Board Structure . . . 6
  - 2.2.3 Diversity . . . 7
- 2.3 Impact of Good & Bad Governance . . . 7

3 Methodology **9**
- 3.1 Pipeline Overview . . . 9
- 3.2 Data Sources . . . 10
- 3.3 Data Transformation . . . 12
  - 3.3.1 CEO Background . . . 12
  - 3.3.2 Board Structure . . . 17
  - 3.3.3 Diversity . . . 18
  - 3.3.4 ROA Change . . . 19
  - 3.3.5 Final Table . . . 21

4 Experiments & Results **22**
- 4.1 Different Types of Models . . . 22
- 4.2 Models Tested . . . 24
  - 4.2.1 Random Forests . . . 24
  - 4.2.2 Gradient Boosting Machines . . . 24
  - 4.2.3 Support Vector Machines . . . 25
  - 4.2.4 Extreme Gradient Boost . . . 26
  - 4.2.5 Logistic Regressions . . . 26
  - 4.2.6 Convolutional Neural Networks . . . 26
- 4.3 Evaluation Methodology . . . 28
- 4.4 Statistical Tests . . . 29
- 4.5 Results . . . 29

5 Automation **32**

6 Discussion **34**

7 Conclusions **38**
7.1 Summary of Results 38
7.2 Future Work 39

Bibliography **40**

A Data Dictionary for Source Tables **43**

B Tables from Reviewed Papers **44**

C Data Dictionary for Final Table **45**

Chapter 1

Introduction

1.1 Motivation

Stock markets play a pivotal role in the global economy, serving as platforms where shares of publicly traded companies can be bought and sold by the public. Such markets are integral to the functioning of modern economies, facilitating capital formation for these (publicly traded) companies, resource allocation for asset managers, and providing investment opportunities to the general public. Participation in stock markets is widespread, ranging from individual retail investors to institutional investors and financial institutions [Brown et al. (2018)]. The prices of stocks are influenced by various factors, including company performance [Prastuti and Setianingrum (2019/02)], the business environment [McQueen and Roley (1993)], macroeconomic conditions [McQueen and Roley (1993)], investor sentiment [Gao et al. (2020)], market policy and frictions [Shirota et al. (2021)], and systematic investor biases [Cherono et al. (2019)] In analyzing stock prices, two main approaches are commonly employed: technical analysis, which relies on the behavior of historical price and volume data to forecast future price movements, and fundamental analysis, focusing on assessing a company's intrinsic value based on its financials and business prospects [Beyaz et al. (2018)].

Technical analysis is a method that involves analyzing a stock's historical patterns to extract information for predicting future stock prices. It offers several advantages, including the ability to identify quick short-term stock price cycles and the widespread availability of data (as market related data is widely available to the public free of charge). Additionally, technical analysis can also account for market psychology and sentiment factors, hence incorporating traders' confidence in a stock into its analyses. However, technical analysis also has drawbacks. In particular, it can lead to risky decisions, rooted in speculation, as it is prone to false signals [Zielonka (2004)]. This is because the methodology relies on the assumption that historical patterns may repeat in the future, and that insights can be learned from these historical patterns. However, the market is inherently unpredictable, and past performance is not always indicative of future outcomes [Hendry and Mizon (2014)]. Therefore, while technical analysis can provide valuable insights to inform financial decision making, it should be approached with caution and supplemented with other forms of analysis to mitigate its inherent

risks. Additionally, technical analysis neglects information related to a company's performance and future prospects, which encompasses factors such as research and development investments, patents, financial robustness of an organization, and other important factors like environmental, social, and governance (ESG) considerations, which can significantly impact a company's long-term prospects and performance. By focusing solely on historical price movements, technical analysis fails to account for these fundamental aspects of a company's value and overlooks critical information that could affect its future prospects and stock price [Drakopoulou (2015)]. In other words, it only provides information related to the patterns and trends of stock markets, but not about the intrinsic performance and value of a stock.

Fundamental analysis is a method that seeks to find a stock's real value to decide whether it is fairly priced and whether it should rise or fall in the long run. It is less risky than technical analysis, being grounded in intrinsic value and not vulnerable to speculation. Nevertheless, it presents several downsides, such as needing a deep understanding of both the market and the company. It is highly time-consuming, and presents a lesser immediate upside, as well as the challenges that come with the lack of data availability. On top of being it is also a subjective analysis as an analyst's bullish or bearish. One of the harder aspects of fundamental analysis is the fact that its factors are not precisely defined and are often measured by non-objective metrics. An example lies in corporate governance.

Defined as "the way in which companies are being governed and to what purpose" [The Chartered Governance Institute (2021)], it encompasses the non-numerical aspect of how a company is being managed: structure and practices of management, compliance with the law, salaries, among others. Given that governance includes all non numerical aspects of how a company is being run, it is a critical factor for firm performance and therefore stock performance. In this project we will be looking at a subsection of corporate governance, focusing on three main areas: CEO background, the structure of the board of directors, and diversity. CEO background means looking at CEOs' education, history at the company, and previous employment; the structure of the board encompasses information such as the committees present, the number of directors in these, and how often they meet. Finally, diversity looks at the ethnicity of directors, gender representation of the board, and if there is a board committee solely dedicated to ensuring diversity in the company.

Relative to technical analysis where data is readily available due to companies having to publish their financial information on specified dates, fundamental data is not, and the data that is available is very often segmented, needing to be combined. Furthermore, given the highly subjective interpretability of the non-numerical information, existing analyses and points of views regarding a company's financial outlook highly depend on whether the writer and whether they are more bearish or bullish.

1.2 Problem Statement Key Points

Evaluating corporate governance is a pivotal element of investment best practice due to its significant impact on organizational performance and, by extension, stock

performance. Effective governance boosts investor trust and confidence, key factors in attracting and retaining investment [Tamimi and Sebastianelli (2017)]. Moreover, sound governance practices are instrumental in mitigating various risks, including financial mismanagement, fraud, and corruption, thereby safeguarding shareholder value. Governance is a multifaceted concept which includes the quality of monitoring, ownership and leadership structure, board governance, and incentive plans used by the company [Coles et al. (2001)]. Being such a broad concept, it is challenging to identify relevant data and metrics for a thorough understanding. The creation of a dedicated database would help address this issue by providing streamlined access to governance metrics, thus facilitating analysis for investors, researchers, and regulators. This would solve both the issue of data availability and data fragmentation. A database that combines information like this could provide a standardized framework for comparing governance practices across different companies and industries.

However, as is made evident in the data collection for this project, there is a significant gap in the availability of governance-related data. Information is dispersed across multiple sources, and while our project only examines a subset of governance, the issue remains prevalent. Even third-party metrics, like those made by companies such as Kavout[1] and Bridgewise[2], are typically out of public reach or lack transparency in their methodologies, acting as a ”black box” that offers scores without explanations. In an interview with a former financial services provider company during the scoping stage of this project, the significance of fundamental analysis in supporting trading for long-term returns in today's highly speculative market was emphasized, especially due to the growing number of uninformed investors trading stocks. However, it was also noted that financial services companies often do not fully utilize governance related information due to the lack of availability of such data. This anecdotal evidence underscores the demand for governance information that remains unmet due to the unavailability or challenges in retrieving such data.

This project aims at filling these gaps by aggregating fragmented data sources to create a centralized data pipeline that makes governance information readily available. This project will allow more informed decision-making by investors regarding using Governance related information to support financial analysis and decision making. Furthermore, this project will employ state-of-the-art machine learning models to derive governance-related metrics from the collected data. This approach endeavors to help democratize actionable related insights from a large database across key governance factors. In achieving this aim, this project is aimed at paving the way for the development of new governance benchmarks and indices. In sum, this project seeks to develop a data-driven solution to enhance the understanding of governance related factors in public companies and assist investors in integrating such information into their financial decisions.

[1]Kavout is a financial technology company that leverages artificial intelligence to provide stock rankings and investment strategies.

[2]Bridgewise is an analytics firm that offers predictive financial modeling and data-driven insights for investors.

1.3 Project Aims

As previously stated, this project aims to develop a data pipeline to aggregate publicly available governance information and create transparent metrics for evaluating different factors in connection with organizational governance. This will be achieved by using best-in-class data orchestration techniques, and ML models to democratize knowledge in the field and help provide the required data to accelerate research progress on the field

The companies sampled from the study were derived from the constituents of the S&P 400, 500 and 600 indices. The governance-related aspects from these companies will focus on their CEO's professional background, the structure of the companies' boards, and the diversity of the Board of directors. More specifically, we will be collecting the following information for each of the 3 areas of focus:

- CEO Background: education (degrees and alma mater) and history at the company (tenure, previous roles, whether they are the founder, whether they are co-CEOs).
- Board Structure: number of Board committees, number of directors in them, director attendance to official Board meetings, and number of official meetings.
- Diversity: female representation, existence of a diversity board, and ethnicity of directors

We will then use this data to predict the year-on-year (YoY) change in ROA for each of the companies, and will do this entire process by following the steps outlined in Figure 1.1.

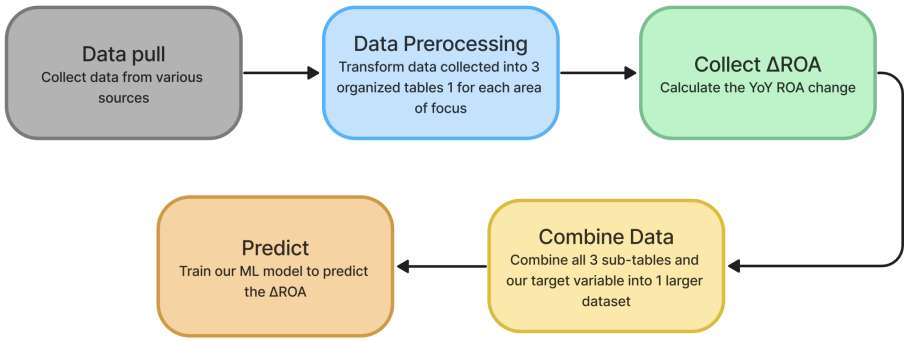

Figure 1.1: Overview of the Data Pipeline

Chapter 2

Literature Review

2.1 Governance and Impact on Stock Market

Research being conducted consistently displays the influence of corporate governance on stock prices and company performance. Many studies demonstrate this effect on firms over time, highlighting the fact that investors should not overlook it as a factor.

[Christian et al. (2020)] dives into the relationship between fundamental governance factors and how they are correlated with stock prices. With their research, the authors demonstrate the positive correlation between the frequency of board of director meetings, the number of directors, and the presence of education and training programs with key financial metrics. However, the study finds that the number of independent commissioners and the number of board of director meetings show a negative correlation with stock prices. Given that governance is such a wide concept, it is impossible to evaluate all aspects of governance in one research, but the correlations proven in this paper display the relevance of governance from the board's point of view.

In [Cremers and Ferrell (2009)] Cremers and Ferrell establish the strong positive link between high-quality corporate governance and a firm's returns over 30 years. They state that "[...] we find a robust positive association between 'good' corporate governance and abnormal returns for the 1978-2006 period [...]" and demonstrate through various figures and tables this relationship. Notably, Table IX illustrate that companies with a lower G-Index, have higher abnormal returns. The G-Index is a measure of restrictions on board members, therefore a lower value indicates the board has more liberty There have also been various attempts in the past to create a "corporate governance score". The California Public Employees' Retirement System (CalPERS) has made one of these grades by using factors such as board independence and diversity, executive compensation policies, shareholder rights, and the transparency of financial reporting. In Table 1 of [MacAvoy and Millstein (1999)] we can undeniably see the extremely strong correlation between the score and the company's percentage annual rate of returns (ROR), where the only companies that have a consistently positive ROR are those with CalPERS of A+ and the very strong average yearly ROR that companies

with scores of A+, A, and B have compared to those with C, D, or F (1.76% vs -2.87%).

2.2 Key Elements of Governance

This section will review previous research on the influence of CEO background, board structure and diversity on company performance.

2.2.1 CEO Background

Many different studies have looked at different aspects of how a CEO's background affects a company's performance. [Urquhart and Zhang (2022)] explores the correlation between a CEO's educational level, including the university rankings, and various performance metrics such as ROA. They came up with equation 2.1 below to establish the relationship between a CEO's education and the firm's performance, where βs are the coefficients for each education level. Table 6 [Urquhart and Zhang (2022)], clearly shows that CEOs with higher education can direct their companies to revenues stronger than industry averages. This table also shows that the quality of the education of CEOs is crucial. When diving into the performances of companies run by CEOs with PhDs, we see an improvement in performance by 3.03%, but a PhD from a top 100 university leads to a 4.20% improvement in performance.

$$\begin{aligned} FirmPerformance = \alpha + \beta_1(PhDEducation) + \beta_2(PGEducation) + \\ \beta_3(MBAEducation) + \beta_4(ControlVariables) + ... \end{aligned} \tag{2.1}$$

In addition, [Saidu (2019)] explores the relationship between CEO ownership, education, and professional origin positively influencing stock performance by looking at the ROA and return on equity (ROE). This study indicates that "Stock performance gets improved when the CEO has prior experience of the firm before being appointed as the chief executive officer" displaying the importance of a CEO having tenure and history at the company.

Seeing as tenure and roles at the company are deemed to be important, the line of reasoning logically extends to founder CEOs. Founders are known to have extensive tenure and a deep-rooted understanding of the industry, and more importantly, have been through every milestone of a company. In [Fahlenbrach (2009)], we can see how the performance of investment portfolios behaved between having only founder-CEOs, and other benchmarks. The findings seem to agree with [Saidu (2019)], given that "founder-CEO firms would have earned an abnormal return of 10.7% annually, and an equal-weighted strategy would have earned 8.3% annually, compared to a benchmark four-factor model" [Fahlenbrach (2009)]. They accredit this difference in performance to founder-CEOs' distinct approach to investments and overall management style.

2.2.2 Board Structure

The composition and behaviour of a company's board are essential to its future performance given that it is where all leadership decisions come from. This relationship is not only theoretical but is supported by empirical evidence. Table 5 [Lin et al. (2014)]

(available in Appendix B Table B.1) shows the result of the linear regression made to demonstrate the strong positive relationship between board attendance and ROA. Having highlighted the relevance of attendance they dive deeper into the reasons for poorer attendance, pointing out that factors like multiple directorships, high meeting frequency, and large board size tend to decrease attendance. This suggests that contrary to some popular beliefs, more meetings and larger boards might inadvertently lead to lower director attendance, negatively affecting company performance.

[Modum et al. (2013)] reached similar conclusions when analyzing board size and meeting frequency against earnings per share (EPS), finding that smaller boards –especially those including external directors– generate a better EPS. Regarding the relationship between meeting frequency and performance, they diverge from [Lin et al. (2014)], highlighting the lack of a strong correlation. Nonetheless, they do not disagree with the conclusion made, stating that excessive meetings could be counterproductive, introducing additional costs and time constraints for directors.

2.2.3 Diversity

Diversity within a company's leadership and the enforcement of a more balanced representation of genders and ethnicities is a topic of discussion that has been trending recently.

[Simionescu et al. (2021)] investigates the impact of rising female director numbers on the performance of S&P 500 companies. Figure 1 reveals the steady pace at which female presence has increased over the past decade, and when combined with Table 6 we see that this is a good thing. Table 6 reveals a positive correlation between female board representation and ROA, suggesting that gender-balanced boards are associated with improved financial performance. The research highlights an optimal level of female representation, beyond which ROA may decline, suggesting the importance of balanced gender diversity.

In line with these findings, through [Ararat et al. (2010)] which explores both gender and ethnic diversity, it finds that the diversity index they defined is positively correlated to both market-to-book (MTB) and Tobin-Q, getting p-values of 9% and 3% respectively.

In summary, the background of a CEO (encompassing their education, tenure, and professional origin), the structure of the board of directors, and the ethnic and gender diversity of a company have profound and measurable impacts on firm performance. These facets should not be overlooked in the assessment of a company's governance quality and potential for success.

2.3 Impact of Good & Bad Governance

We don't have to look back very far to find evidence of the impact corporate governance has on market capitalization. The papers discussed in this section highlight the effect of governance as an important catalyst to most unpredicted downfalls or skyrocketing stocks.

A testament to how powerful effective governance can be is Microsoft's recent change in leadership. In 2014, Satya Nadella took over as CEO, instigating a transformation in the company's work culture promoting agility and innovation. His strategic approach has been received as a gold standard and he is credited with revitalizing Microsoft. [Refaeuter (2019)] states that Nadella's leadership catalyzed a strategic renewal that is widely received as one of the greatest examples of how leadership change can positively impact a company regardless of its size. These changes are reflected in Microsoft's stock price, which soared from $37 to over $425 (current price) under his tenure, with the company becoming the most valuable in the world at the time of this dissertation.

Conversely, a lapse in governance can spell disaster, as seen in the Volkswagen emissions scandal in 2015. The company's lies regarding emissions test results for over 11 million diesel vehicles worldwide coupled with the extensive cover-up including document destruction, led to a drastic 50% drop in VW's stock price between April and September 2015. [Kano et al. (2023)] identifies this governance crisis as an aftereffect of many factors including inadequate international governance practices, insufficient shareholder oversight, and decision-making that was unduly influenced by a few dominant directors. As a result, this environment discouraged accountability, ultimately resulting in a scandal of significant proportions.

If these examples weren't enough, the 2008 financial crisis is yet another reminder of the consequences of negligent governance. [Grove and Victoravich (2012)] highlights factors such as all-powerful CEOs, weak management controls, and an obsession with short-term objectives played a critical role in precipitating the crisis. They contend that such governance failures laid the groundwork for the fraudulent reporting and excessive risk-taking that followed. The issue is that in an example like this one where the governance issue is rooted within an entire industry and affects customers on their net worth, the effect is not one stock falling but one of the worst economic downturns in recent history.

These examples should be enough evidence to prove the relevance of governance on firms' performances, highlighting the limitation of purely relying on financial KPIs to assess a firm's health. In 2007, firms like Lehman Brothers, which might have appeared robust by purely numeric metrics, ultimately proved the opposite. These examples underline the crucial importance of sound corporate governance in determining a firm's performance, positively or negatively.

Chapter 3

Methodology

3.1 Pipeline Overview

In this chapter, we will go through the entire data pipeline [1] step by step to go from our sources to the tables we will then input into our ML models. As is seen in Figure 1.1, the pipeline is broken down into five distinct phases: data pull, preprocessing of explanatory variables, calculation of the target variable, creation of the final table, and ML prediction. Each stage plays a crucial role in shaping the outcome of our analysis.

Data Pull: Given that our focus is divided across three areas: CEO background, board structure, and diversity, we need to collect data for each of them. For each of these, we collect the raw data needed for the explanatory variables we want in our future analyses.

Data preprocessing: This phase involves the meticulous cleaning and consolidation of data. The end product is one table per focus area: ceo_v0, board_v0, div_v0. Given the numerous data sources, integrating them requires many different joins, making sure to watch out for similarly named columns that contain different information, and sometimes consolidating data when the column to join on is of type text and the wording is different. To do this, string-matching algorithms are employed. Additionally, we perform the calculations needed to transform the many columns and rows pulled to have three tables with one row per ticker, and the fields ready for our predictions in the final step.

Target variable calculation: We use the change in ROA as our target variable to be predicted from the previously processed. Being a non-governance factor, it is derived from a different source, which we collect as quarterly ROA, and then calculate the difference per ticker between the last quarter recorded and the one four before.

Create final table: With all preparatory work done, we combine our three subtables and the ROA data on their ticker symbols. This will create the dataset necessary for our ML prediction phase.

ML prediction: Our pipeline finished with the application of our machine learning techniques. We select the consistently highest-performing model from the experimental

[1] The GitHub repository containing all the code discussed in this paper is available here.

chapter of the dissertation. This model is trained using our newly cleaned dataset to predict the change in ROA, effectively measuring one form of performance for how the company is governed.

This systematic approach aims to offer a nuanced understanding of how governance factors influence financial outcomes, underpinned by a robust data-driven methodology, that we hope will be used by others.

3.2 Data Sources

In this project, we are harnessing data from four different datasets, three from the Wharton Research Data Services (WRDS) and the final one from the Times Higher Education (THE) World University Ranking. The WRDS data pull spans six tables from three datasets: Directors US and Governance US from ISS ESG, Board and Director Committees, Organization Summary - Analytics, and Individual Profile Education from BoardEx, and finally, Financial Ratios by Firm Level from Financial Ratios.

ISS ESG is a robust and evolving database used by institutional investors and researchers globally, featuring data on corporate directors, governance structure, vote outcomes, as well as climate and emissions information. For our pipeline, we focus on the North American segment of the ISS ESG database, which provides us with S&P 1500 companies. Both Directors US and Governance US present us with data from January 1st 2007 to the present, updated yearly.

Directors US (DirUS) furnishes data on directors' board positions, titles, and ethnicity, among others, from which we extract their attendance, meeting dates, ethnicity, and gender. From **Governance US** (GovUS) which offers insights into committee structures, policies, and defence mechanisms, we will only record companies' tickers and their respective S&P indices.

BoardEx, the other main dataset of our data collection, archives over 1.7 million executives across more than 2.2 million organisations worldwide. It contains bibliographical data on executives and board committees, permitting investors and researchers to explore directors' backgrounds, prior employments, stock options, and political association among other factors. Given the project's scope, we will once again only select the North American dataset, having records from December 1st 1997 to the present, updated weekly.

From the **Individual Profile Education** (ProfileEducation) table, we will record the degree type and university obtained for all directors at the companies we are examining. The table has 1 row per degree per director. The **Board and Director Committtees** (BoardsDirs) table contains information on committees and directors' roles and positions through time. We will record the different committee names for each company as well as the directors in these. Lastly, the **Organization Summary Analytics** (OrgSummary) table contains data per director on their demographic and geographic data as well as their board affiliations, remuneration details and much more. We only use this table as an intermediary to join the education and committee tables, as well as get the ticker for companies mentioned in those tables, given that BoardEx only provided this information

in the summary table.

The **Financial Ratios Firm Level** (FinRatios) table is part of the newest dataset provided by WRDS. It offers a monthly updated series of over 70 financial ratios covering eight dimensions such as valuation, liquidity, profitability, solvency, capitalization, efficiency, financial, and other uncategorized ratios. From this table, we will be retrieving the ticker of companies and the date of their quarterly reports along with our target variable, ROA for each of these quarters.

THE World University Ranking (UniRanking) is a dataset containing the yearly rankings and scores of the top universities worldwide from 2011 onwards, each year being a different table. It contains over 2000 universities per year, ranking the first 200 individually. Afterwards, it assigns them to buckets of different sizes depending on how low they are ranked: 201-250, . . . , 401-500, . . . , 601-800 and so on. THE's ranking is widely recognized as one of the most used university rankings in the world, along with QS, validating the quality of the scores assigned. From this table, we will only be recording the university names and their overall score.

Together, these datasets form the backbone of our data pipeline, enabling a multifaceted analysis of corporate governance across publicly traded companies, with a particular emphasis on the S&P 1500. From these sources, we will collect specific data on our different areas of interest. As shown in Figure 3.1, we use some of these source tables for multiple of our "raw" data tables. We indicate that these tables are unprocessed by having a 010 at the end of their name. In the end, we will have collected the following columns from each of the sources:

- DirUS: ticker, director id, attend less75 pct, meeting date, ethnicity, female
- GovUS: ticker, spindex
- ProfileEducation: companyname, qualification
- BoardsDirs: directorid, directorname, rolename, datestartrole, dateendrole
- OrgSummary: ticker
- FinRatios: ticker, qdate, roa
- UniRanking: name, overall score

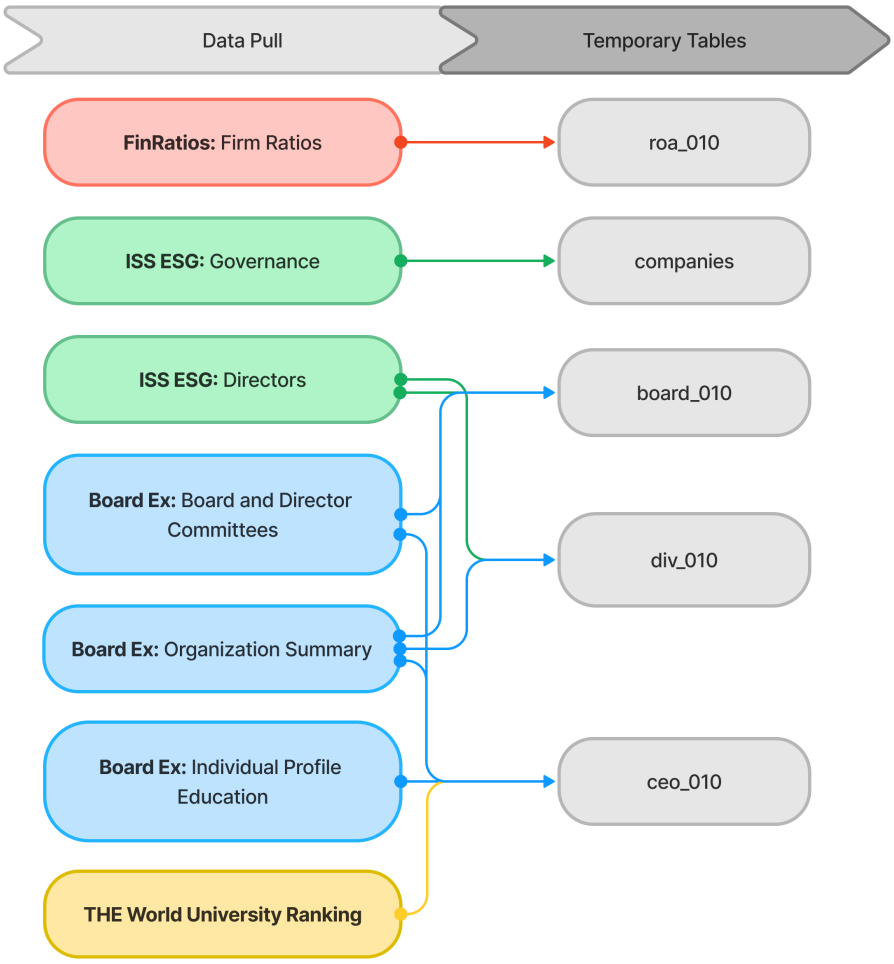

Figure 3.1: Creation of 010 Tables (Raw data pulled from sources)

3.3 Data Transformation

The data transformation to go from the raw values retrieved to the features that will be fed into ML models can be broken down into five distinct processes; one for each area of interest, one for the ROA, and one to create the final table combining all data.

3.3.1 CEO Background

The process to build the CEO subtable consists of five steps: collection of data, cleaning the data, adding the education data, identifying CEOs that were promoted from within the company, and finally keeping only the current CEO for each ticker.

Step 1: ceo_010 – As stated in the data sources section, the 010 tables are those

purely recording the data from our sources. In order to build ceo_010, we combine data from the 3 BoardEx tables mentioned. Using Query 1, we record ticker, id (directorid), name, (directorname), role (rolename), date of start and end of the role (datestartrole and dateendrole), university (companyname), degree (qualification) for all directors in BoardEx.

```
SELECT      comp.ticker
            , emp.directorid
            , emp.directorname
            , emp.rolename
            , emp.datestartrole
            , emp.dateendrole
            , edu.companyname as university
            , edu.qualification
FROM        boardex.na_wrds_dir_profile_emp emp
            INNER JOIN  boardex.na_wrds_org_summary comp
                ON comp.boardid = emp.companyid
            LEFT JOIN boardex.na_dir_profile_education edu
                ON edu.directorid = emp.directorid
GROUP BY    comp.ticker
            , emp.directorid
            , emp.directorname
            , emp.rolename
            , emp.datestartrole
            , emp.dateendrole
            , university
            , edu.qualification
;
```

Listing 3.1: Query 1 – Creation of ceo_010

Step 2: ceo_020 – The ceo_020 step consists of beginning the data cleaning. In step 1, information on all directors was collected, however, given that the interest only lies in CEOs, all rows that do not contain "CEO" in the role are removed. Additionally, we remove all roles that contain "regional" or "division" as that would mean the director is not the CEO of the company. This filtering could have been done at any point, but it was done first to remove as much data as possible and optimize runtime for future steps. Finally, the data frame (df) is filtered to only contain the companies in the S&P1500. This is done by merging the df with companies; another df that only contains tickers in one of the S&P 400, 500, and 600 as well as their corresponding S&P index.

Step 3: ceo_030 – This step tackles the completion of the education data for CEOs incorporating the data from THE World University Ranking by joining ceo_020 and UniRanking. Firstly, a new df is created by concatenating all ranking files from THE. Given that universities' rankings change from year to year and that the new df has 14 years of data, it is grouped by university name averaging the overall score THE gave each university. At this point, the 2 datasets ceo_020 and universities need to be joined to create ceo_030 with the average score of CEOs' education.

The datasets should be joined on the university name. However, given that these columns are of type string and don't correspond to an identifier like ticker. Therefore,

when taking multiple sources that reference a university name, it is highly likely that the strings won't match. For example, when mentioning the University of Edinburgh, one could write "The University of Edinburgh", "Edinburgh University", or "University of Edinburgh", etc. none of these strings match but they all refer to the same university.

To circumvent this issue, fuzzy-wuzzy is used, to define a function get_closest_match()that, given a string, will return the closest match from a whole array of strings, in this case, the closest match should be the one that references the same university. get_closest_match()uses fuzzy-wuzzy's extractOne() which is by default going to calculate the Wratio as a measure of similarity between the strings. This ratio is defined and implemented by fuzzywuzzy by combining their 3 in-house developed scorer functions.

1. fuzz.ratio(): Calculate the Levenstein distance similarity ratio.
2. fuzz.partial_ratio(): Compares subsections of the strings to find the best match.
3. fuzz.token_sort_ratio(): Splits the strings into its words, "tokens", sorts them alphabetically, and rejoins them into a string to see how many of the words match.

The Levenstein distance is essentially a value indicating the smallest number of edit operations necessary to go from one string to another using a maximum of one operation per step. The possible operations are substitution, insertion, and deletion. For example, the Levenstein distance going from "Honda" to "Hyundai" is 3:

1. Operation 1: Insert "y" after between "H" and "o" – Hyonda
2. Operation 2: Substitute "o" for "u" – Hyunda
3. Operation 3: Insert "i" at the end of the word – Hyundai

Therefore, when calculating the similarity of strings, using the Wratio that includes the three ratios mentioned together is the best way to do so given that it covers the three main issues with strings not matching:

1. Different spelling
2. Extra words i.e. "The" when comparing "The University of Edinburgh" and "University of Edinburgh"
3. Different order of words

Now having ceo_030 as described, leaves one issue to be dealt with, there is one row per degree each director got instead of one row per director. The solution used is to convert the degree column into 4 Boolean ones: Bachelor, Master, MBA, and PhD. Doing so required a categorization of the distinct values for degree into their corresponding level and, assigning the value "True" to the appropriate column. Finally, the last step consisted of removing the university names and grouping the rows by ticker, director, and role averaging the university score, applying OR to our 4 new columns and selecting the minimum start date and maximum end date for each role.

After inspecting the output data, however, it became apparent that an exception had yet to be handled. If someone is still in their position, the enddaterole is blank. Therefore, in the case that a director was in their role, then terminated, and then resumed their role, the current aggregation would ignore their newly instated role effectively removing them as the current CEO.

To circumvent this, the two date columns were converted to a date format and filled in the blanks in end dates by adding the current date.

Step 4: ceo_040 – The goal of this step is to create a column that indicates if a CEO is an internal hire. What is considered an internal hire is someone who was working at the same ticker before becoming CEO regardless of the previous role or if there was a break between the end of the previous role and the start of the CEO role. To do so, a new df promotion is defined from running query 2. This will return all combinations of roles a director (role 1) has had at a company crossed with all other ones (role 2) they have had at the same company. From this query promotion will have the columns explained in Table 3.1:

Column	Description	Purpose of column
ticker	-	Merge back to evolving ceo dataset
directorid	-	Merge back to evolving ceo dataset
rolename	role 1 title	Only keep CEO roles
non_role	role 2 title	Only keep non-ceo roles
datestartrole	start date for role 1	Ensure it comes after start date for role 2
non_startrole	start date for role 2	Ensure it comes before start date for role 1

Table 3.1: Data dictionary for promotion data frame

```
WITH temp AS (  SELECT    comp.ticker
                        , emp.directorid
                        , emp.rolename
                        , emp.datestartrole
                FROM    boardex.na_wrds_dir_profile_emp emp
                        INNER JOIN boardex.na_wrds_org_summary comp
                            ON comp.boardid = emp.companyid
                GROUP BY  comp.ticker
                        , emp.directorid
                        , emp.rolename
                        , emp.datestartrole)
SELECT      ceo.ticker
            , ceo.directorid
            , ceo.rolename
            , non_ceo.rolename AS non_role
            , ceo.datestartrole
            , non_ceo.datestartrole AS non_startrole
FROM        temp ceo
            LEFT JOIN temp AS non_ceo
                ON non_ceo.directorid = ceo.directorid
                AND ceo.ticker = non_ceo.ticker
GROUP BY    ceo.ticker
```

```
23              , ceo.directorid
24              , ceo.rolename
25              , non_role
26              , ceo.datestartrole
27              , non_startrole;
```

Listing 3.2: Query 2 – Cross all directors' roles with their other ones at the same company

With all these combinations of roles recorded we only keep the ones that match the following conditions:

1. Role 1 contains 'CEO'
2. Role 2 does not contain 'CEO'
3. Role 2 is a previous one to role 2 (Start date for role 1 is greater than the start date for role 2)

Finally, promotion is joined onto ceo_030, forming ceo_040. However, in order to make sure the internal_promotion column is properly created, the type of merge for each row was indicated, so that internal_promotion is only true if the merge type is "both", indicating that the CEO did indeed have a previous role at the same company.

Step 5: ceo_v0 – The final step in the creation of the CEO background subtable is creating columns tenure and founder and lastly cleaning it up to make sure there is only the correct CEO per ticker. Tenure is defined by calculating the difference between the years for start and end dates. Founder is indicated in the role name of the director. The BoardEx separates all titles applicable to one director by "/", therefore if a current CEO is also the founder, they would have the title "Founder/CEO". The final issue to tackle regarding this subtable is that currently there are multiple directors with the title CEO for each ticker, but the goal is to only have one. The reason for having multiple CEOs is one of three described below:

1. **Reason 1**: Different director name but CEO
   (a) **Problem**: Past CEO
   (b) **Solution**: Only keep the latest valid_start

2. **Reason 2**: Co-CEOs run the company
   (a) **Problem**: Role is Co-CEO
   (b) **Solution**: Indicate co-CEOs in a new column is_co_ceo and calculate new values in for other columns:
      i. University score: Average
      ii. Degrees: Or
      iii. Tenure: Sum
      iv. Founder: Or
      v. Internal promotion: Or
      vi. Is co-CEO: Or

3. **Reason 3**: Multiple rows for same director

(a) **Problem**: Director's role name changed while maintaining the CEO title
(b) **Solution**: Keep only one of the titles and apply the following changes to the other rows:
  i. University score: Average
  ii. Degrees: Or
  iii. Tenure: Sum
  iv. Founder: Or
  v. Internal promotion: Or
  vi. Is co-CEO: Or

3.3.2 Board Structure

In 3.3.2, we will go through the same process of transforming data from respective sources to the board_v0 suitable. The process is much simpler, requiring a separate data pull for committee and director information, followed by joining this data and filtering the tickers to only keep S&P 1500 companies.

Step 1: comms & dirs – Comms is a df containing the number of committees (boards) and directors in them. It is created by running Query 3, which pulls data directly from BoardEx's Board and Director Committees and counts the number of unique committee names and director IDs per ticker in the year 2023. Dirs is a df with the percentage of directors attending over 75% of meetings, and the number of meetings for each ticker. It pulls data directly from ISS ESG Directors US through Query 4. It calculates the percentage of directors with over 75% attendance by taking the column that indicates below 75% attendance, converting it to an integer by converting the indicator for poor attendance to a 0 and otherwise setting the value to a 1. Once this is done, it groups the rows per ticker, directorid and board to finally calculate the ratio of 1s in the column. Grouping by ticker is important because, as previously discussed, Directors US contains multiple rows for the same director, and thus we would be counting the same director's attendance multiple times.

```
SELECT      comp.ticker
            , count(distinct(comm.committeename)) as num_committees
            , count(distinct(comm.directorid)) as dirs
FROM        boardex.na_board_dir_committees comm
            LEFT JOIN boardex.na_wrds_org_summary comp
                ON comm.boardid = comp.boardid
GROUP BY    comp.ticker;
```

Listing 3.3: Query 3 – Create comms

```
WITH temp AS (      SELECT    dir.ticker
                            , case when dir.attend_less75_pct='Yes'
                            , dir.meetingdate
                              then 0
                              else 1
                            end as attend
                    FROM      risk.rmdirectors dir
                    WHERE     dir.year = 2023.0)
```

```
SELECT        t.ticker
            , sum(t.attend)/count(t.attend) as over75_pct
            , count(t.meetingdate) as num_meetings
FROM        temp t
GROUP BY    t.ticker;
```

Listing 3.4: Query 4 – Create dirs

Step 2: board_020 – This step finalizes the data preprocessing for information on the board by filtering tickers to be in the S&P 1500, merging the two datasets as well as calculating the average number of meetings per committee. To filter the tickers it merges comms into companies. Then it merges dirs onto the result of the previous merge. Finally, the average number of meetings is calculated by dividing the number of meetings by the number of committees for that ticker.

3.3.3 Diversity

The data cleaning for the diversity subtable is broken down into three steps detailed below: data collection, calculation of women representation, and calculation of diversity among directors.

Step 1: div_010 – This df contains for each director: ethnicity, whether they are a female, and all committee names for each ticker. It is done by running Query 5 below which ensures that only directors and boards still present in 2023 are being considered. It uses ISS ESG's Directors US, and BoardEx's Board and Director Committees, Organization Summary – Analytics tables.

```
WITH dir AS (   SELECT        dir.ticker
                            , dir.director_detail_id
                            , dir.female
                            , dir.ethnicity
                    FROM      risk.rmdirectors dir
                    WHERE     dir.year = 2023.0)
SELECT    dir.ticker
        , dir.director_detail_id
        , dir.female
        , dir.ethnicity
        , comm.committeename
FROM    dir
        LEFT JOIN  boardex.na_wrds_org_summary comp
            ON comp.ticker = dir.ticker
        LEFT JOIN boardex.na_board_dir_committees comm
            ON comm.boardid = comp.boardid
GROUP BY  dir.ticker
        , dir.director_detail_id
        , dir.female
        , dir.ethnicity
        , comm.committeename
;
```

Listing 3.5: Query 5 – Create div_010

Step 2: div_020 – In this step, the women percentage is calculated, as well as identifying tickers with a board committee dedicated exclusively to diversity. To calculate the women's percentage, the women's column, which currently consists of "Yes" and blanks, is converted to "Yes" and "No" values. Now counting the number of "Yes" and the total number of values the ratio of women directors is calculated and put into a column female_pct.

The field diversity_board in div_020 indicates the presence of a committee dedicated to ensuring diversity within the company. It is created by flagging any row in div_010 that has substrings "divers", "incl", or "sustai" in the committee name. The reason for these substrings is that after going through the data, most committees dedicated to diversity contained the words diversity, inclusion, or sustainability or some variation of these words like inclusivity. Once these rows are marked with a True value in the boolean column diversity_board, the rows are grouped by ticker, removing the committee name and applying an or operator to the new column.

Step 3: div_030 – In div_030 the Shannon Index \cite{ChaoShen2003} is calculated for each ticker by going back to div_010, given that div_020 is already grouped by ticker. The Shannon Index is a measure more commonly used in ecology to quantify the biodiversity in habitats. It is appreciated for its ability to measure richness (number of different categories) as well the evenness (distribution or abundance) of species being considered. In the context of corporate governance, the Shannon Index is adapted to assess the ethnic diversity in the company. The formula for the Shannon Index is seen in Equation 3.1 below:

$$H' = -\sum_{i=1}^{R} p_i \ln(p_i) \tag{3.1}$$

where H' is the Shannon Diversity Index, R is the total number of ethnic categories present within the company, p_i is the proportion of individuals belonging to the i^{th} ethnic category, and ln denotes the natural logarithm. A higher Shannon Index value indicates greater diversity, meaning that not only is there a variety of ethnic groups but also a balanced representation across them.

The Shannon Index is particularly useful in corporate settings for quantifying the level of ethnic diversity among employees or board members. It provides a comprehensive picture that considers the number of different ethnic backgrounds and their proportional representation, offering insights into the inclusivity and heterogeneity of the organizational environment. This metric aids companies in understanding the diversity of their workforce, fostering an inclusive culture, and identifying areas for improvement in their diversity and inclusion strategies. Now, the diversity subtable is complete.

3.3.4 ROA Change

The goal of this process is to prepare the target variable for the models that will be run. To do so, the quarterly ROA for all tickers is pulled from the Financial Ratios dataset using Query 6. The YoY change in ROA is calculated by finding

the difference between the columns "roa" and "roa_prev_year". "roa_prev_year" is defined by matching each row to the one 4 quarters before (qdate), and recording the roa in roa_prev_year. Finally, as per the other tables, only one row is kept per ticker, therefore only the one corresponding to the most recent quarterly report is left.

```
SELECT  DISTINCT
            fr.ticker
            , fr.qdate
            , fr.roa
FROM        wrdsapps_finratio.firm_ratio fr;
```

Listing 3.6: Query 6 – Collect ROA per quarter

The goal of the ROA as a measure of performance is not about the exact value but rather the trend the company is showing. Therefore, predicting the exact ROA change proves less useful than predicting the tendency that is being displayed. The numerical field is therefore transformed to a categorical one by allocating the values to their corresponding bin that captures significant variations in the data. The bins and the reason for their values:

1. **Significant decrease:** ROA Change < -0.10 – Captures significant negative changes well beyond one standard deviation from the mean, showing a substantial decrease in performance.

2. **Moderate decrease:** $-0.10 \leq$ ROA Change < -0.03 – Captures a fall in ROA that is less severe but still below the 25th percentile, showing a fall that is still significant.

3. **Slight Decrease:** $-0.03 \leq$ ROA Change < 0 – Captures a fall in ROA that is no cause for concern. It is still close to the lower quartile, but not indicative of major performance issues.

4. **Stable:** $0 \leq$ ROA Change < 0.01 – Captures an ROA change that reflects stability. It encompasses changes around the median and these changes are negligible. Thus it can be concluded that there is no real change in the ROA.

5. **Slight Increase:** $0.01 \leq$ ROA Change < 0.02 – Captures an improvement in ROA that is between the median and the 75th percentile. This therefore indicates a minor improvement in ROA.

6. **Moderate Increase:** $0.02 \leq$ ROA Change < 0.10 – Captures positive changes in ROA that are significant but not extreme. Therefore it indicates a performance improvement that shows a good trend for the company.

7. **Significant Increase:** ROA Change ≥ 0.10 – Captures the most substantial positive changes in ROA that only very few companies will reach. It is well beyond one standard deviation from the mean, indicating a significant performance improvement.

The rationale regarding the way the bins are defined is that the boundaries, as explained, are based on the desire to represent the data spread. The negative bins are set more aggressively towards the lower end because of a long left tail in the data. The most positive bin is set at 0.10 slightly arbitrarily, but it still ensures that outliers on the positive side are captured.

3.3.5 Final Table

Given that all data is already well-cleaned and prepped, joining all of them together is very straightforward. ceo_v0 and div_v0 are joined on the ticker creating full_v0. Then board_v0 is joined onto full_v0, and finally, roa_v0 can be joined as well.

Chapter 4

Experiments & Results

4.1 Different Types of Models

When looking at the papers mentioned in Chapter 2 it is clear that most papers that predict the ROA do regression rather than a classification like this project. Therefore for this project, it was important to start by identifying different types of models and their performances before diving into the hyperparameter tuning for each. In this section we will therefore discuss the 6 model types that were experimented with: Random Forests, Gradient Boosting Machines, Support Vector Machines, Extreme Gradient Boosting, Logistic Regressions, and Convolutional Neural Networks.

Random Forest (rf) is an ensemble learning method that uses lots of decision trees in its training phase and outputs the mode of the classification of the individual trees. The goal of using random forests over decision trees is to correct decision trees' tendency to overfit predictions on the training set. This method is appreciated due to its ability to handle categorical and numerical data, not overfit and provide estimates of feature importance. However, its drawbacks are based on its high complexity leading to more computational resources and a challenging interpretability. Random forests' positives outweigh its negatives in the context of this project. It was chosen as the data being handled is diverse (boolean, categorical, and numerical) and this variation usually leads to overfitting.

Gradient Boosting Machines (GBM) is a technique that builds models sequentially. Each model tries to correct errors made by previous ones. In the end the model that is used for each prediction is selected through a decision tree. It is praised for its ability to be highly effective on datasets with non-linear relationships, and for supporting many loss functions, allowing for customization for each task. On the other hand, it has a tendency to overfit if not tuned properly as well as being computationally expensive for the training phase given the sequential models being made. It was selected for this classification given that the task required predicting categories from various data types, which leads to non-linear relationships.

Support Vector Machines (SVMs) are a supervised machine learning algorithm that is used for both classifications and regressions. This method performs classifications by finding the multidimensional plane that best divides a dataset into its classes. It is

effective in high dimensional spaces as well as datasets with clear margins of separation but is unsuitable for very large datasets because of its computational complexity. It also performs poorly in very noisy datasets. Considering the relatively small dataset being used, and its high dimensional feature space, SVMs seem like a good candidate for classifying the ROA changes, especially if there are clear margins of separations.

Extreme Gradient Boosting (XGBoost) implements gradient-boosted decision trees to enhance speed and performance. It is more advanced than GBMs having better efficiency, and flexibility. XGBoost offers regularization options which limit overfitting, are highly scalable and allow for cross-validation. Nonetheless, it is easy to overfit with XGBoost if hyperparameters aren't properly tuned, and therefore require a lot of hyperparameter tuning. Given its efficiency and versatility, it seems like an ideal candidate to run experiments on.

Logistic Regression (LogReg) is a statistical method used to predict binary outcomes by linearly combining predictor variables. While it traditionally handles binary outcomes, this method can be extended to multi-class classification through techniques like one-vs-rest (OvR). In OvR, separate LogRegs are trained for each class against all others, allowing the algorithm to handle multiple categories by breaking down the problem into various binary decisions. LogRegs are easy to implement and understand given that they return output probabilities, which provide insight into the confidence level of each prediction. However, it has significant drawbacks due to its assuming linearity between dependent and independent variables leading it to perform poorly in highly complex relationships in data. Although this seems like a method that will perform worse than the others discussed, it could prove as a baseline comparison against more complex models, giving insight into whether it is worth it to use complicated models on this data.

Convolutional Neural Networks (CNN) are deep learning algorithms primarily used in classifications, clustering, and image processing. CNNs automatically detect and learn spatial hierarchies of features through their building blocks, and convolutional layers. These layers apply convolutional operations to the input passing the result to the next layer and so on. Then, a filter or kernel is slid over the input data to make a feature map that emphasizes certain aspects of the data. CNNs will therefore go through the layers passing one's output to the next one's input, and then, will use backpropagation and gradient descent algorithms to learn the optimal parameters of the filters, making it improve its accuracy as it trains. These models are praised for their exceptional ability to capture dependencies in the data through the application of relevant and custom filters, allowing them to detect important features without human supervision. However, they require large datasets to train effectively, as well as acting like a black box limiting interpretability. While the dataset being used is not large, and it is not traditional for CNNs to be used on tabular data, their ability to detect intricate patterns and dependencies in data could offer unique insights and in turn performance.

4.2 Models Tested

4.2.1 Random Forests

3 random forests have been made with different values for the hyperparameters regarding the depth and number of trees. However, they all have the same class weight and random state. Class weight is set as balanced to ensure that imbalances in the occurrence of each ROA change category do not impact the model. This will adjust weights inversely proportionally to class frequency. The random state is set at 42 so that the reproducibility of results is ensured.

rf_model_1 utilizes 200 trees in the forest in an attempt to maximize model accuracy by reducing the variance in predictions given that the outcome is the mode across more trees. However, it increases complexity, which given the large number of features, offers a robust foundation to capture patterns.

rf_model_2 has 100 trees, which is less than rf_model_1. This value aims at balancing computational efficiency and model accuracy in the hope that the reduction in model complexity will not significantly decrease performance while significantly reducing complexity. Setting a max depth of 10 means no tree will have more than 10 levels. This limitation also aims at reducing complexity and overfitting, making sure trees are generalizing and not creating all cases to fit all training data.

rf_model_3 maintains efficiency of training like rf_model_2, having 100 trees. It adds another parameter of a minimum sample split of 5. This means that each node must have at least 5 samples to consider a split. This is done to control tree growth and limit overfitting, as once more, it will prevent trees from perfectly fitting the data.

When comparing these three models rf_model_1 is the most complex to train due to it having the highest number of estimators, whilst both rf_model_2 and rf_model_3 have half the trees, with a stronger focus on targeting overfitting. The choice of hyperparameters for the last 2 models aims at refining the learning process. It controls model complexity and makes them potentially more suited for datasets like this one, where overfitting is an issue given that financial data evolves quickly and shows different patterns in very close periods.

4.2.2 Gradient Boosting Machines

Again, there are 3 different models made, and all 3 use a random state of 42 to ensure reproducibility of results. Similarly to RFs, GBMs use trees but use them sequentially rather than individually to then get the mode.

gbm_model_1 uses 200 trees, allowing for a thorough learning process that can truly correct lots of errors made by previous trees. As with all models, this increased complexity aims at increasing accuracy but will also increase computational time. For this model, the learning rate is set at 0.05. This smaller learning rate makes each tree contribute a little to the overall prediction for each ticker. As a consequence, it ensures a model robust

to overfitting ensuring complex relationships are captured through the multitude of trees.

gbm_model_2 uses fewer trees than gbm_model_1 (100), aiming at a balance between learning efficiency and computational complexity. Given that the model uses 100 trees, this model does not require a low learning rate, and instead subsampling is set to 80%. This introduces a stochastic gradient boosting using 80% of the data randomly selected for each tree. Ultimately as the data is not used for each, it reduces both overfitting and training time.

gbm_model_3 also has 100 trees, but each of them will have a maximum of 5 levels. This is another attempt at preventing overfitting.

Overall, these three models, like the ones before are structured as one being a "base" with the highest complexity, while the rest aim at reducing this complexity while still maintaining good accuracies. gbm_model_1 aims at capturing intricate patterns in the data at the expense of computational resources while the two others focus on efficiency. Models 2 and 3 take different approaches to overfitting control making them more suitable for datasets with high risks of this issue. The tricky part about the dataset being used is that while it is not large in terms of rows, it does have many features of different types, and finding relationships between these features is what presents complexity. Therefore gbm_model_1's slow learning rate allows for a more granular model over many iterations whilst the other two take an approach that aims at making the model more future-proof.

4.2.3 Support Vector Machines

The SVMs' class weight and random state are set at balanced and 42 respectively for the same reasons as the previous models, and the hyperparameter that will differentiate the models the most is the kernel type.

svm_model_1 uses a linear kernel in an attempt to separate the data using a linear threshold. In the case that the data can be well separated through a straight line or hyperplane, the model would perform well.

svm_model_2 uses a radial basis function (RBF) as a kernel allowing for non-linear boundaries. Given the many features and complex relationships between them, this kernel type could be a powerful model. However, this model is more complex and usually leads to overfitting. To prevent this the regularization parameter C is set to 0.5. This value indicates a balanced penalization for errors in both the training and testing data, while a smaller C would have led to more penalization for errors on the training set.

svm_model_3 also uses an RBF kernel but instead uses the other main hyperparameter SVMs to tackle overfitting. Gamma, which controls the influence of a training example on the decision boundary (C) is set to 0.01, indicating that the model will generalize better for unseen data by not focusing too much on the training set.

Over the 3 models, the kernel type chosen dictated the tuning of hyperparameters based on its tendency to over or underfit. Both svm_model_2 and svm_model_3 required regularization as seen by setting the C parameter at 0.5 and gamma at 0.1. Additionally, all three use balanced class weights to address the data imbalance that is present even though the ROA change buckets were set to start targeting the issue.

4.2.4 Extreme Gradient Boost

The XGBoost models being tested aim to further the testing made on GBMs. Given XGB's more advanced algorithms, the goal is to use a more complex model while still being careful to not overfit. For all three models, some hyperparameters are maintained such as the scaling applied to tackle class imbalance, the loss function, and the random state. Scaling is set to 1, indicating that none is being applied. Regarding the loss function, as is common in multiclass classification problems such as this one, the models use multinomial logarithmic loss (mlogloss) which calculates the logloss for each of the m classes.
In order to avoid overfitting xgb_model_1 has a maximum depth of 8 as it is deep enough to understand complex patterns but not too much to learn the whole dataset. Just like for gbm_model_1, in an aim at reducing overfitting, xgb_model_2 is set to have a learning rate of 0.05, slowing down learning by keeping adjustments made to weights more conservative. Lastly, xgb_model_3 takes another approach to overfitting. Minimum child weight controls the minimum sum of instance weight needed for a node to exist, preventing nodes from corresponding to only one example. In this case, it is set at 5.

4.2.5 Logistic Regressions

The last machine learning model used is logistic regression. This is the most conventional base model used for classification. All three of the models tested have the same random state and a maximum of 1000 iterations. This means that through these iterations the model will converge and reach optimal coefficients for a balance between performance and overfitting.

logreg_model_1 uses C = 0.5 for regularization, indicating a moderate level of penalization in an attempt to strike a balance between bias and variance. logreg_model_2, on the other hand, uses a liblinear solver. This is optimal on small datasets like this one, being praised for its efficiency and performance on linear models. While the first two models have a balanced weight for each class, logreg_model_3 manually assigns weights to each class, which is something that can be done as we know the representation of each class.

4.2.6 Convolutional Neural Networks

As mentioned, CNNs use layers to apply convolutional operations to their input, passing the result to the next layer, therefore the type of input and the operation being performed greatly impacts the performance of the model. The models built use 4 different kinds of layers: input, dense, dropout, and batch normalization. Every model starts with an input

layer as it is the entry point for data into a neural network specifying the shape of the input it will receive. Dense layers, also called fully connected, are layers in which each neuron is connected to all neurons from the previous layers. Each neuron will compute the weighted sum of values outputted by previous neurons, adding a bias, and optionally applying an activation function. This process is summarized by equation 4.1 below:

$$output = activation(dot(input, weights) + bias) \tag{4.1}$$

Dropout layers are used to prevent overfitting and do so by acting like a colander. During training, dropout layers are inserted to "drop out" some of the layers' output by setting them to zero. By dropping out some of this data, the model is sure to overfit less given that it randomly eliminates the outputs of neurons in the previous layer. Lastly, batch normalization layers are used to standardize the inputs of a layer by subtracting the batch mean and dividing by the batch's standard deviation. Given that smaller batches of data are being used, training is done much faster and the model is less sensitive to the network's initialization.

Other than layers, models have activation functions, optimizers, and loss functions. Optimizers are algorithms used to change the attributes of NNs such as the weights of neurons and the learning rate in an attempt to improve performance. All CNNs used in this classification will use the Adam optimizer which uses a fixed learning rate for updating weights while having varying learning rates for each parameter. This leads to an optimizer that combines the benefits from the AdaGrad and RMSProp algorithms in an efficient manner. All models will also be using categorical cross entropy for the loss function. Lastly, the activation functions used in the dense layers will be one of 3: Rectified Linear Unit (ReLU), Hyperbolic Tangent (Tanh), and Softmax. ReLU is a function that outputs its input if it is positive, otherwise it outputs 0. This creates a nonlinear component into the linear network seen in equation 3. Tanh transforms its input into a value between -1 and 1, per equation 4.2 below which details the calculation of the Tanh activation function. Like ReLU it introduces non-linearity through its hyperbolic shape, being 0 centered.

$$Tanh(x) = \frac{e^x - e^- x}{e^x + e^- x} \tag{4.2}$$

The last function used in the models is Softmax, which is used in the output layer for multi-class classification problems. It converts the raw predictions of the model into probabilities by taking the exponential of each output and normalizing the values.

cnn_model_1 is a relatively simple model, using 3 consecutive blocks of dense followed by a dropout layer to prevent overfitting. This pattern attempts to capture complex patterns in the data while maintaining generalization. For the three dense layers, it uses the ReLU algorithm, and after the final dropout uses a Softmax layer to categorize ROA changes to the specific classes.

cnn_model_2 replaces the dropout layers from the first model for batch normalization aiming at stabilizing and accelerating the training process. Additionally, it uses a mix

of ReLU and tanh as activation functions in an attempt to perform well given the complexity of the relationship between features.

cnn_model_3 is the simplest model of all consisting of only dense layers with ReLU activations, followed by a softmax output. The goal of this model is to increase interpretability, although it has a chance of underperforming compared to the other 4.

cnn_model_4 scales up the complexity of the models by using more neurons (256 vs 128 in previous models) as well as having higher dropout rates (40% vs 30%) aiming to capture more detailed patterns. Just like cnn_model_2, it uses ReLU and tanh as activation functions. Overall this model is configured to perform well for classifications in which slight nuances will mark the difference between categories.

cnn_model_5 is the most complex of all having even more neurons (512), and incorporating batch normalization with dropout layers. It strives to balance the benefits of deep learning's ability to learn complex hierarchical patterns while maintaining generalization and preventing overfitting.

Overall these models were built to each perform well in distinct situations. Given that the relationships between the features are too complex for a human to understand, the goal was to have models that should perform drastically differently with at least one being adapted to this dataset. Model 3 is the simplest one while 4 and 5 are much more complicated. Leaving 1 and 2 as a middle ground. While 4 and 5 are more complex they also have more measures to prevent overfitting such as their reduced learning rates.

4.3 Evaluation Methodology

Given the importance of selecting a model that not only fits the training data well but also generalizes effectively to unseen data, we employed a robust evaluation strategy that involves multiple runs of resampling and cross-validation.

To remove the random factor coming from the splitting of training and testing of the data, we implemented a resampling strategy that involves 1000 iterations for each model. In each iteration, the dataset was randomly split into training and testing subsets, simulating a k-fold cross-validation with the addition of random splits for each iteration. This approach helps in assessing the model's stability and generalizability across different data samples.

Given that accuracy measures all correct classifications (both true positives and true negatives) in the total number of cases examined, it stands as the most adapted performance metric from accuracy, precision, recall and f1 for this classification. This metric is particularly relevant for classification tasks where the objective is a multiclass prediction of a categorical outcome as opposed to binary classifications. It provides a straightforward measure of how often the model makes the correct prediction, irrespective of the class distribution.

4.4 Statistical Tests

Once the accuracies for each of the 20 models across the 1000 iterations are recorded, we rigorously compared the models using statistical tests to identify the best performer in the most objective way possible. This selection was done in stages as described below:

The first stage is to conduct a t-test for each model's mean accuracy against a baseline accuracy that represents random guessing. This baseline is critical as it sets a minimum threshold for model performance, ensuring that the selected model has predictive capabilities significantly better than random. In this case, given there are 7 classes, random selection would give $\tilde{1}4\%$ accuracy. Therefore, taking the null hypothesis (H0) as no model performs better than random and the alternative hypothesis (H1) for each model as it performs better than random. The t-test consists of calculating the α as per Equation 4.3 below:

$$\alpha = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \tag{4.3}$$

where $\bar{x}$ is the mean accuracy for a model, μ is the accuracy for a random model (1/7), σ is the standard deviation of the accuracies of the model, and n is the number of times the model is run (1000). If α is inferior to 0.10, then the model in question provides ROA predictions that are significantly superior to random. This step was used as a primary filter to exclude models not showing a statistically significant improvement over the baseline from further consideration.

If multiple models demonstrated statistically significant improvements over the baseline, further comparisons were made to identify the best among these. The selection process among significant models was carried out in two steps:

T-test for Difference in Mean Accuracy Between Top Models: To do this the models were ranked based on their accuracy, and a one-sided t-test was conducted to compare the mean accuracy of the best model against the second-best. This would ensure that there is a model that is statistically the best, allowing for its selection as the preferred model.

Analysis of Variance (ANOVA): If the difference in mean accuracy between the top two models was not statistically significant, we proceeded with an ANOVA. This stage aimed to detect any significant differences in mean accuracies among all top-performing models, providing a basis for selecting the model with the highest mean accuracy.

For all three of these tests, the threshold to reject H0 was set to 0.10.

4.5 Results

When analyzing Table 4.1 1 which contains the results of the experiments, it becomes apparent that all 20 models significantly outperformed the baseline accuracy threshold of 14%. Given that each model's mean accuracy is well above the baseline threshold and that the alpha for all of them is below 0.10 we will need to conduct further

comparative analyses to find the best one.

When performing the secondary t-test, we start by comparing the best-performing model to the second-best-performing one: cnn_model_4 (mean accuracy of 25.60%) vs cnn_model_1 (mean accuracy of 25.18%). When performing the t-test we get a p-value of 0.0935, which is lower than 0.10. This result indicated statistical significance, rendering further tests such as ANOVA unnecessary. Consequently, cnn_model_4 was selected as the model for future predictions.

Model Name	Mean Accuracy (%)	Standard Deviation (%)	Alpha
Random Forest Models			
rf_model_1	24.47	∼0.00	0.0
rf_model_2	18.09	∼0.00	0.0
rf_model_3	23.40	∼0.00	0.0
Gradient Boosting Models			
gbm_model_1	22.70	∼0.00	0.0
gbm_model_2	22.34	0.0	0.0
gbm_model_3	20.21	∼0.00	0.0
SVM Models			
svm_model_1	14.18	0.0	0.0
svm_model_2	14.89	∼0.00	0.0
svm_model_3	15.60	0.0	0.0
XGBoost Models			
xgb_model_1	18.79	0.0	0.0
xgb_model_2	21.99	∼0.00	0.0
xgb_model_3	20.57	∼0.00	0.0
Logistic Regression Models			
logreg_model_1	11.70	∼0.00	0.0
logreg_model_2	14.18	0.0	0.0
logreg_model_3	17.02	∼0.00	0.0
CNN Models			
cnn_model_1	25.18	2.33	1.32e-69
cnn_model_2	20.51	1.89	1.87e-55
cnn_model_3	24.96	2.44	8.45e-67
cnn_model_4	**25.60**	**2.18**	**5.50e-74**
cnn_model_5	23.18	2.24	6.74e-63

Table 4.1: Model performance summary

It is worth noting that all models except those in the CNN category exhibited near-zero variance in their performance metrics. This is due to the deterministic nature of the methods used for these models when applied with a fixed random state. In contrast, CNNs are subject to variability due to their inherent stochastic nature coming from weight initialization, optimization during training, and dropout layers which randomly eliminate a subset of neurons. Such randomness introduces variability in the model's performance across different runs, explaining the non-zero standard deviation values

observed.

The design choices in the neural network architectures also contribute to this variability. For instance, cnn_model_4 includes layers with a high number of neurons and dropout layers with a dropout rate of 0.4. While its architecture introduces more variance, it appears to contribute to higher mean accuracy, suggesting a beneficial trade-off between the model's complexity and its ability to generalize.

In summary, the results indicate that cnn_model_4 not only offers a significant improvement over random guessing but also stands out amongst its competitors. Its architectural features, while introducing some variance, have proven to be advantageous in achieving the highest mean accuracy, making it the preferred model for this classification problem. Appendix C contains the data dictionary of the final table generated by this entire process with the predicted ROA change.

Chapter 5

Automation

In an attempt to streamline the data processing workflow for the project, we have used Apache Airflow, a software that allows people to author, schedule, and monitor workflows.

The pipeline has been broken down into 7 distinct tasks, which are executed in a sequence ensuring data dependencies are respected while running as many tasks simultaneously as possible to ensure efficiency. The breakdown is as follows:

1. Data pull:
   (a) Company pull
   (b) CEO information pull
   (c) Diversity information pull
   (d) Board information pull
   (e) ROA information pull
2. Combine all subtables to form full_v0
3. Train model and record predictions to create final.csv, the final table of this process.

The tasks are encapsulated into a direct acyclic graph (DAG), ensuring they are executed in the proper sequence. Airflow's scheduler manages the triggering of each task ensuring the order for the graph is respected. By having the DAG defined as in Figure 5.1, and running all 4 data pulls together, time is being saved. Running the DAG in such a way takes about 20 minutes, as opposed to 30+ minutes, doing all tasks one after the next.

Finally, given how often the datasets are refreshed, and that the ROA gets updated quarterly, the pipeline is set to run every 3 months, ensuring the most up-to-date data.

Each task in the DAG is idempotent, meaning it can be re-executed without causing duplicate data entries or side effects, which is crucial for the integrity of the process. On top of that, Airflow's error handling system ensures that if, for some reason, an error occurs, administrators are alerted and tasks can be retried or skipped.

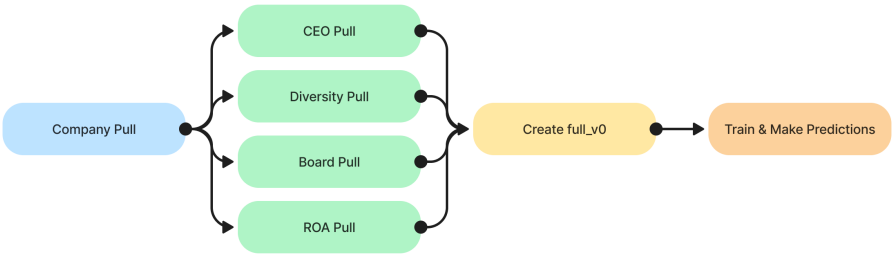

Figure 5.1: Airflow DAG of pipeline

Chapter 6

Discussion

In order to better understand how the CNN behaves and the weights it associates with the different governance metrics, I used the Local Interpretable Model-agnostic Explanation (LIME). This provides transparency to how the governance metrics may influence ROA predictions made by cnn_model_4. LIME explains the classification of one ticker by displaying 3 tables. On the left-hand side, is the probability of the ticker falling into the different classifications, in the middle it displays the thresholds that explain why the feature values lead to the classification outputted in decreasing order of importance. Lastly, on the right is the weight contributed to the classification for each feature. Figures 6.1 to 6.4 show the interpretations for 4 different data points, in which the model behaved differently.

Figure 6.1 shows a LIME explanation that is common throughout the dataset. Given that there is a 57% chance that the ROA will not change or only slightly decrease (between a 0% and a 3% change). The decision for this classification is mostly based on the average number of meetings per board committee, whether there are co-CEOs, the Shannon index, the existence of a diversity-dedicated board, and the representation of women directors. Figure 6.2 shows a less common example given that the classification probabilities are more balanced. However, the prediction is still based on the same top 5 features as in Figure 6.1. Figure 6.3 shows the opposite. It displays a case where the classification was much more decisive. There is about a 40% chance of a slight decrease in ROA, with the next most probable class at 19%.

Again, the prediction is mostly based on the same features. Figure 6.4 shows the last case that is commonly seen for the classification by the CNN in which the decision is balanced between 2 main classes. Even though there are 2 main classes, it seems like the model finds that stable is more probable than a slight decrease.

Overall, although LIME shows how different individual classifications are done, what is most interesting is the patterns that emerge when examining all of them together. Clearly, some features are the most prevalent in classifications. As explained in the previous paragraph these features are the existence of a co-CEO, the existence of a diversity board, the Shannon index, the average number of meetings held by committees by year, and the female representation on these committees.

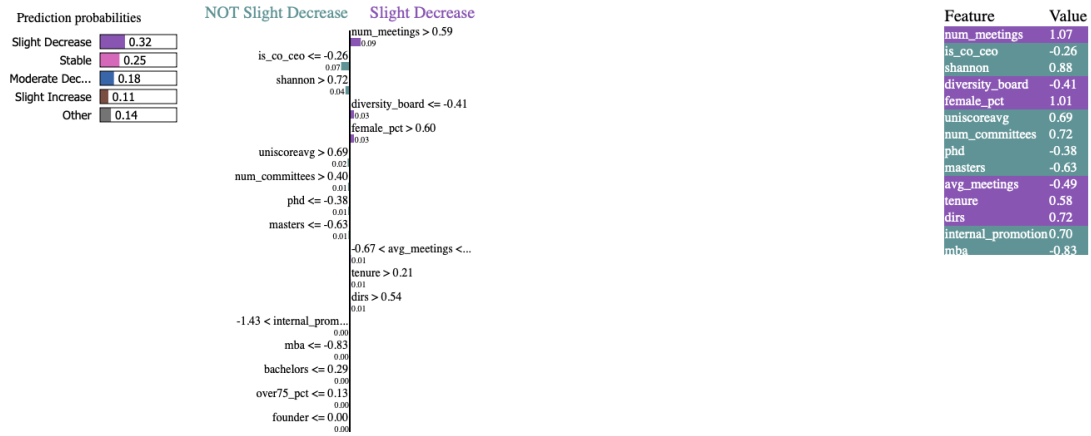

Figure 6.1: LIME Model Interpreter

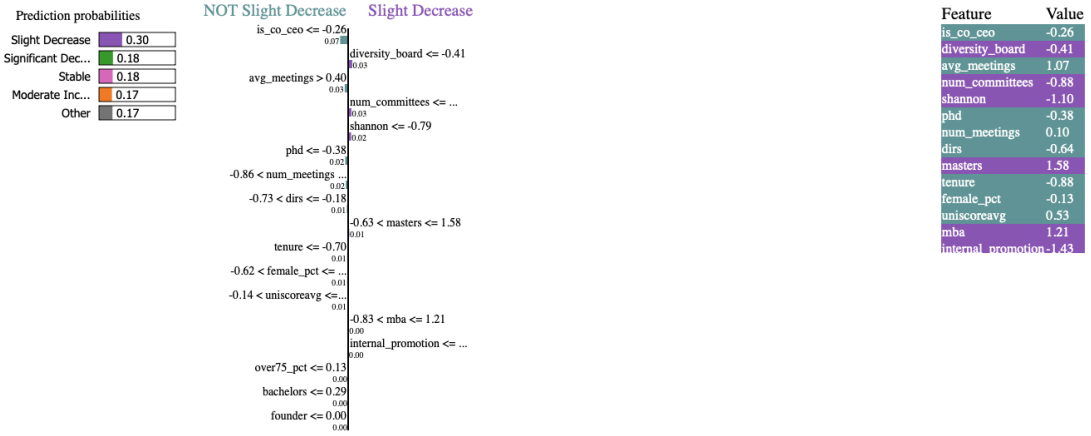

Figure 6.2: LIME Model Interpreter with Balanced Probabilities

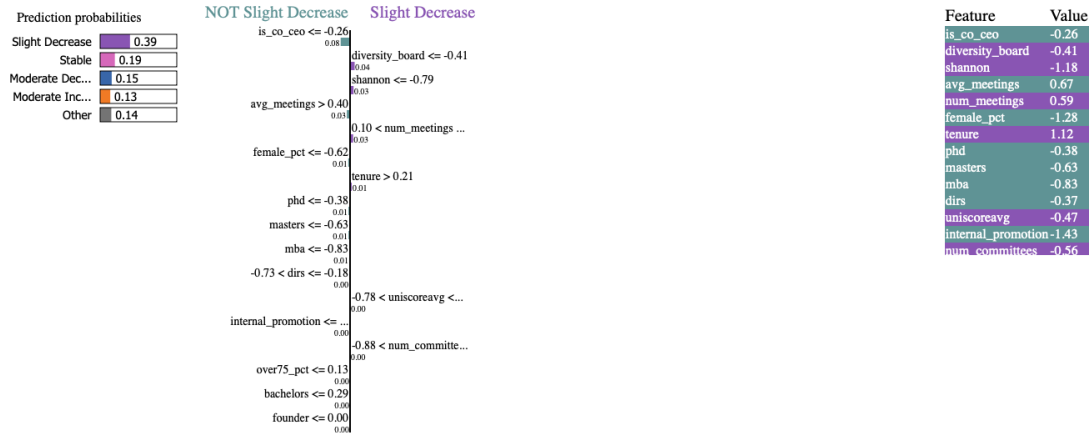

Figure 6.3: LIME Model Interpreter with Decisive Probabilities

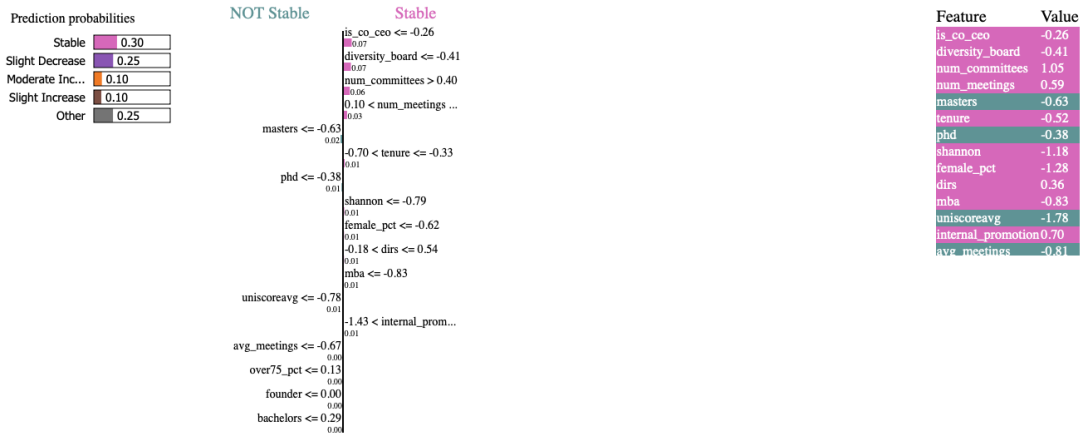

Figure 6.4: LIME Model Interpreter 2 Main Classes

When looking at the overall representation of each class, we can see in Table 6.1 that the model follows the same patterns as the original data. However it seems like the model has exaggerated the data imbalances as seen in the vast over-representation of "Slight Decrease" and "Stable", but the under-representation of all other classes, even eliminating "Significant Increase" altogether.

Class	ROA Change - Original	ROA Change - Predicted
Significant Decrease	7.89%	2.90%
Moderate Decrease	14.96%	9.96%
Slight Decrease	27.70%	44.77%
Stable	19.67%	24.01%
Slight Increase	9.28%	2.54%
Moderate Increase	17.31%	15.82%
Significant Increase	3.19%	-

Table 6.1: Change in class representation in original vs. predicted table

Keeping in mind that the goal of this project is to understand the impact of governance factors on the ROA, we can use the findings from LIME to conclude the relative importance of the features used. This can be done by finding the average position of each field in LIME explanations. When doing so, we can see that there are four main groups formed:

1. Most important: is_co_ceo (average position 1.1), diversity_board (average position 2.5), num_meetings (average position 3.3)
2. Second most important: shannon (average position 5.9), phd (average position 6.2), masters (average position 7.7), avg_meetings (average position 7.8), female_pct (average position 7.9)
3. Third most important: num_committees (average position 9.5), uniscoreavg (average position 9.7), dirs (average position 9.8), tenure (average position 10), mba (average position 10.6)
4. Least important: internal_promotion (average position 13), over75_pct (average position 15.3), bachelors (average position 16.3), founder (average position 16.4)

These results indicate the existence of co-CEOs, a diversity-dedicated board and the number of meetings per board committee are good indicators to predict firm performance. However, internal promotions, attendance, having a bachelors and being a founder are not significant either because the value is usually the same (i.e. most CEOs have a bachelor's degree or companies that perform well enough to be in the S&P 1500 have good attendance from directors), or because the different values have no relationship with the change in ROA.

While the creation of the final table and model may seem rather straightforward, lots of decisions had to be made along the way.

An area that is often considered when examining corporate governance is compensation. I've omitted compensation from this study due to its endogenous relationship with return on assets (ROA); directors often receive higher pay based on strong ROA figures, largely because their earnings are tied to bonuses and stock options, which are dependent on ROA. However, playing devil's advocate it's arguable that higher compensation could boost motivation and thereby improve company performance. This situation presents an endogeneity issue, making it challenging to determine the causality between compensation and ROA.

Another significant decision taken regarded the model and the predictions. The original idea was to not join the subtables and have a model for each to predict the change in ROA, getting the final prediction by getting the mode of the three. In the end given the models trained, I decided to join the dataset as described because, especially for CNNs, the more data used in training the better performance you are likely to get.

Chapter 7

Conclusions

7.1 Summary of Results

In this project, we've gone through the creation of a data pipeline to collect and clean information on the corporate governance of S&P 1500 companies to finally use this data to predict their YoY change in ROA. With this project, we have examined the state of the current industry, seeing how research has been done on the relevance of corporate governance to firm performance. Yet, data is still scattered and often incomplete, needing the extra work of combining heterogeneous sources in a variety of ways and then processing the data to reach a comprehensive dataset.

We have also tested 6 different types of machine learning and deep learning models, running experiments for each of those to better understand the kind of model that is best adapted to this data. The conclusion was that our 4th CNN, having a high complexity but also a higher dropout rate to balance overfitting was the best performing. Through statistical tests, it was concluded that although this model is not close to being perfect, it can have significantly better results than random, and more importantly is the best-performing model by a statistically significant margin, as shown through the t-test conducted.

In the end, the predictions made proved to follow the overall representation of each class, while exaggerating the imbalance present in the dataset. It is apparent through the over-representation of the most popular classes like "Slight Decrease" and "Stable", and the under-representation of less popular ones such as the elimination of the "Significant Increase" class.

Through the use of LIME, a library that provides a view into the black boxes that are ML models, we saw that although for each row, the weights change for each feature, the same features appeared as being the most important: having a co-CEO, a diversity board, and the value of the Shannon index. Alternately, some features were highlighted as having little to no importance relative to the rest: the CEO being internally promoted, attendance at board meetings, having a bachelor's degree, and the CEO being the founder of a company.

7.2 Future Work

Future steps could include building on this pipeline by adding more features and areas of focus to build a larger table.

Another option would be to improve that data richness by broadening the company list by making it include companies on the NYSE or NASDAQ. This would allow for the dataset to include lots of companies whose data is still relatively readily available and therefore have more data to better train the models and therefore get better results.

Lastly, given that the ROA is a measure of the short-term performance of a company, using corporate governance to see its impact on long-term goals could be interesting since fundamental analysis is based on long-term trading. Measures such as the Altman Z-Score (percentage prediction of a company going out of business in the next 24 months) or the Tobin Q (measure expressing the relationship between market valuation and the intrinsic value of the company) are both good examples of values that could be predicted using the features collected and therefore evaluate a company's long term performance.

Bibliography

Melsa Ararat, Mine H. Aksu, and Ayse Tansel Cetin. The impact of board diversity on boards' monitoring intensity and firm performance: Evidence from the istanbul stock exchange, 2010. Posted: 24 Mar 2010, Last revised: 14 Mar 2015.

Erhan Beyaz, Firat Tekiner, Xiao-jun Zeng, and John Keane. Comparing technical and fundamental indicators in stock price forecasting. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1607–1613, 2018. doi: 10.1109/HPCC/SmartCity/DSS.2018.00262.

Stephen J. Brown, C. Veld, and Y. Veld-Merkoulova. Why do individuals not participate in the stock market? *Cognitive Social Science eJournal*, 2018. doi: 10.2139/ssrn.2822094.

Irene Cherono, Tobias Olweny, and Tabitha Nasieku. Investor behavior biases and stock market reaction in kenya. *Journal of Applied Finance Banking*, 9(1):147–180, 2019. ISSN 1792-6580 (print version), 1792-6599 (online).

Jiwana Christian, Bambang Juanda, and Bayu Bandono. Good corporate governance on stock prices of companies listed in the kompas 100 index 2014-2018. *International Research Journal of Business Studies*, 2020. URL `https://api.semanticscholar.org/CorpusID:234579182`.

Jerilyn W. Coles, Victoria B. McWilliams, and Nilanjan Sen. An examination of the relationship of governance mechanisms to performance. *Journal of Management*, 2001. doi: 10.1016/S0149-2063(00)00085-4.

Martijn Cremers and Allen Ferrell. Thirty years of corporate governance: Firm valuation & stock returns, 2009. Revised: 08 Nov 2009.

Veliota Drakopoulou. A review of fundamental and technical stock analysis techniques. *Finance Educator: Courses*, 2015. doi: 10.4172/2168-9458.1000163.

Rüdiger Fahlenbrach. Founder-ceos, investment decisions, and stock market performance. *Journal of Financial and Quantitative Analysis*, 44(2):439–466, 2009. doi: 10.1017/S0022109009090139.

Zhenyu Gao, Haohan Ren, and Bohui Zhang. Googling investor sentiment around the world. *Journal of Financial and Quantitative Analysis*, 55(2):549–580, 2020. doi: 10.1017/S0022109019000061.

Hugh Grove and Lisa Victoravich. Corporate governance implications from the 2008 financial crisis. *Journal of Governance and Regulation*, 1, 03 2012. doi: 10.22495/jgr_v1_i1_p7.

D. Hendry and G. Mizon. Unpredictability in economic analysis, econometric modeling and forecasting. *Journal of Econometrics*, 182:186–195, 2014. doi: 10.1016/j.jeconom.2014.04.017.

Liena Kano, Sean Simoes, and Alain Verbeke. Governance failure and firm-level crises: the case of the volkswagen emissions scandal. In Anthony Goerzen, editor, *Research Handbook on International Corporate Social Responsibility*, chapter 12, pages 168–186. Edward Elgar Publishing, 2023. doi: 10.4337/9781802207040.00018. URL https://doi.org/10.4337/9781802207040.00018.

Ying-Fen Lin, Yaying Mary Chou Yeh, and Feng ming Yang. Supervisory quality of board and firm performance: a perspective of board meeting attendance. *Total Quality Management & Business Excellence*, 25(3-4):264–279, 2014. doi: 10.1080/14783363.2012.756751.

Paul W. MacAvoy and Ira M. Millstein. The active board of directors and its effect on the performance of the large publicly traded corporation. *Journal of Applied Corporate Finance*, 11(4), 1999. doi: https://doi.org/10.1111/j.1745-6622.1999.tb00510.x.

Grant McQueen and V. Vance Roley. Stock prices, news, and business conditions. *The Review of Financial Studies*, 6(3):683–707, 1993. doi: 10.1093/rfs/5.3.683. URL https://doi.org/10.1093/rfs/5.3.683.

Uche Modum, Robinson Onuora Ugwoke, and Edith Ogoegbunam Onyeanu. Content analysis of effect of board size, composition, frequency of meetings and regulrity in attendance at meetings on financial performance of quoted companies on the nigerian stock exchange 2006-2012, 2013. ISSN 2222-1905 (Paper) ISSN 2222-2839 (Online). URL https://core.ac.uk/download/pdf/234625056.pdf.

Dodi Prastuti and Pristina Hermastuti Setianingrum. Company performance and macroeconomics variables influence on stock price. In *Proceedings of the 5th Annual International Conference on Management Research (AICMaR 2018)*, pages 32–35. Atlantis Press, 2019/02. ISBN 978-94-6252-672-3. doi: 10.2991/aicmar-18.2019.8. URL https://doi.org/10.2991/aicmar-18.2019.8.

Chiara Refaeuter. Microsoft's renewal: The effect a new ceo can have on strategic change and firm performance. Working papers, Yale School of Management, 2019. Revised: 08 Nov 2009.

Sani Saidu. Ceo characteristics and firm performance: focus on origin, education, and ownership. *Journal of Global Entrepreneurship Research*, 9(29), 2019. doi: https://doi.org/10.1186/s40497-019-0153-7. URL https://link.springer.com/article/10.1186/s40497-019-0153-7.

Yukari Shirota, Kenji Yamaguchi, Akane Murakami, and Michiya Morita. An analysis of political turmoil effects on stock prices: a case study of us-china trade friction. In *Proceedings of the First ACM International Conference on AI in Finance*, ICAIF

'20. Association for Computing Machinery, 2021. ISBN 9781450375849. doi: 10.1145/3383455.3422558. URL https://doi.org/10.1145/3383455.3422558.

L.N. Simionescu, Ş.C. Gherghina, H. Tawil, et al. Does board gender diversity affect firm performance? empirical evidence from standard & poor's 500 information technology sector. *Finance Innovation*, 7(52), 2021. doi: 10.1186/s40854-021-00265-x. URL https://doi.org/10.1186/s40854-021-00265-x. Published 01 July 2021.

N. Tamimi and Rose Sebastianelli. Transparency among sp 500 companies: an analysis of esg disclosure scores. *Management Decision*, 55:1660–1680, 2017. doi: 10.1108/MD-01-2017-0018.

The Chartered Governance Institute. What is corporate governance?, 2021. URL https://www.cgi.org.uk/about-us/policy/what-is-corporate-governance. Accessed: 2024-03-14.

Andrew Urquhart and Hanxiong Zhang. Phd ceos and firm performance. *European Financial Management*, 28(2):433–481, 2022. doi: https://doi.org/10.1111/eufm.12316. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/eufm.12316.

Piotr Zielonka. Technical analysis as the representation of typical cognitive biases. *International Review of Financial Analysis*, 13(2):217–225, 2004. ISSN 1057-5219. doi: https://doi.org/10.1016/j.irfa.2004.02.007. URL https://www.sciencedirect.com/science/article/pii/S1057521904000158.

Appendix A

Data Dictionary for Source Tables

Any appendices, including any required ethics information, should be included after the references.

Column	Description
Directors US	
ticker	Ticker
director_id	Dataset owned director id
attend_less75_pct	Attended 75% of meetings
meeting_date	Date of meeting
ethnicity	Ethnicity of director
female	Director is a female
Governance US	
ticker	Ticker
spindex	S&P Index of ticker (400,500,or 600)
Individual Profile Education	
companyname	University name
qualification	Degree obtained
Board and Director Committees	
directorid	Dataset owned director id
directorname	Full name of director
rolename	Role of director
datasetartrole	Date employment started at this role
dateendrole	Date employment ended at this role
Organization Summary Analytics	
ticker	Ticker
Financial Ratios	
ticker	Ticker
qdate	Quarter financial ratios were released
roa	Return on assets
THE World University Ranking	
name	University name
overall_score	University score

Table A.1: Data Dictionary for Model Features

Appendix B

Tables from Reviewed Papers

Results of linear regression made between governance factors and ROA in [Lin et al. (2014)]

Variable	**Expected sign**	**Coefficient**	**t-Value**	**Tolerance**	**VIF**
BDATTEND	+	0.05***	4.12	0.90	1.12
ROA_{t-1}	+	0.70***	57.86	0.94	1.06
YEAR2006		0.27***	19.27	0.72	1.39
YEAR2007		0.23***	16.84	0.72	1.39
No. of observations		3244			
F-value		984.13			
Adjusted R^2		0.548			

Table B.1: Regression Results [Lin et al. (2014)]
***$p < 0.001$
Notes: Two-tailed tests
BDATTEND: Estimated board attendance rate in [Lin et al. (2014)]

Appendix C

Data Dictionary for Final Table

Column	Type	Description
ticker	object	Ticker
uniscoreavg	float64	Average score of all universities attended by CEO
bachelors	bool	CEO has a bachelor's degree
masters	bool	CEO has a master's degree
mba	bool	CEO has an MBA
phd	bool	CEO has a PhD
internal_promotion	bool	CEO was internally promoted to CEO role
tenure	float64	Tenure as CEO
founder	bool	CEO is the founder of the company
is_co_ceo	bool	The company is run by co-CEOs
female_pct	float64	Percentage of female directors
diversity_board	bool	The company has a board committee solely dedicated to diversity
shannon	float64	Shannon index of the company
spindex	object	S&P Index of the company
num_committees	float64	Number of board committees
dirs	float64	Number of directors
over75_pct	float64	Percentage of directors with an attendance of over 75% at official board meetings
num_meetings	float64	Total number of meetings held at the company
avg_meetings	float64	Average number of meetings per committee per year
roapredicted	object	Predicted ROA change category

Table C.1: Data Description