AMERICAN UNIVERSITY OF BEIRUT

TOWARD IDENTIFYING A TOP EMPLOYER USING MACHINE LEARNING

by

LAMIS JALALEDDINE

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Science in Business Analytics
of the Olayan School of Business
at the American University of Beirut

Beirut, Lebanon
June 2020

AMERICAN UNIVERSITY OF BEIRUT

TOWARD IDENTIFYING A TOP EMPLOYER USING MACHINE LEARNING

by

LAMIS JALALEDDINE

Approved by:

Dr. Hoda Daou, Assistant Professor   Advisor
Olayan School of Business   Hoda Daou

Dr. Lama Moussawi, Associate Professor   Member of Committee
Olayan School of Business   on her behalf
Hoda Daou

Date of thesis defense: June 19, 2020

AMERICAN UNIVERSITY OF BEIRUT

THESIS, DISSERTATION, PROJECT RELEASE FORM

Student Name: Jalaleddine Lamis Mohammad Faek

Last First Middle

● Master's Thesis ○ Master's Project ○ Doctoral Dissertation

☒ I authorize the American University of Beirut to: (a) reproduce hard or electronic copies of my thesis, dissertation, or project; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes.

□ I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of it; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes after: **One ___ year from the date of submission of my thesis, dissertation or project.**
Two ___ years from the date of submission of my thesis , dissertation or project.
Three ___ years from the date of submission of my thesis , dissertation or project.

01/07/2020

Signature Date

This form is signed when submitting the thesis, dissertation, or project to the University Libraries

Acknowledgements

I would first like to express my sincere gratitude to my thesis advisor Dr. Hoda Daou for the dedicated support and guidance throughout this study, for her motivation, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my thesis.

I would also like to acknowledge Dr. Lama Moussawi as the second reader of this thesis, and I am gratefully indebted to her for her insightful comments on this research work.

This accomplishment would not have been possible without them. Thank you.

An Abstract of the Thesis of

Lamis Jalaleddine for Master of Business Analytics
Major: Business Analytics

Title: Toward Identifying A Top Employer Using Machine Learning

A firm's reputation, a measure of employee satisfaction, is valuable because, among other benefits, it can attract and motivate good employees, and this in turn, can create labor resource efficiency advantages. There were many efforts to quantify corporate reputation, with fewer attempts to identify the drivers of a firm's reputation, and none used machine learning algorithms. This research work investigates the drivers, in both human and operational practices, that determine performance and increase a firm's reputation as a top employer. Using top employers from Fortune 100 Best Companies to Work For survey, we examine whether a company's financial and operational data in Compustat can foretell its corporate reputation by testing several machine learning algorithms. We provide mathematical evidence that increased spending on R&D and employees' benefits, such as salaries, retirement plans, and insurance packages, does help companies achieve this reputation. Thus, we provide insights on the indicators that help classify a company as a top employer and build recommendations both for general and industry-specific to follow towards becoming a top employer. This research presents a methodological framework to guide employers into becoming a top employer by advancing their internal operational processes and adopting the recommended strategies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the past decade, a firm's reputation has become a focus of attention to both researchers and practitioners (Fombrun and Shanley, 1990; Barnett et al., 2006). A firm's reputation is 'the global perception of the extent to which an organization is held in high esteem or regard' (Weiss et al., 1999). As such, a firm's reputation represents the image associated with how an individual perceives the firm based on past actions and future prospects (Fombrun and Shanley, 1990; Fombrun, 1996).This perceptual representation influences the consumer's decision on the products to buy and the services to purchase (Love and Singh, 2011; Ponzi et al., 2011), allowing to attract new customers and to retain existing ones (Shkolnikov, 2004). Moreover, a firm's reputation serves as a reflection of what the firm stands for, and its relationship with its employees, hence qualified employees are tempted to pursue a career with such companies (Grant, 1991; Lado and Wilson, 1994; Lau and May, 1998; Roberts and Dowling, 2002). Indeed, over the years, there has been wide-ranging research that established the importance of having a good reputation due to its positive impact on the market and financial performance (Fulmer et al., 2003; Lau and May, 1998; Barnett et al., 2006). Evidence from these studies was consistent with the findings that a firm's reputation is an intangible asset (Weigelt and Camerer, 1988; Dierickx and Cool, 1989), which leads to better financial outcomes (Fombrun and Shanley, 1990; Grant, 1991).

Moreover, the focus on the corporate reputation has become significant enough by the media, to publish yearly lists of top employers ranked based on the best work conditions (examples include *Forbes Best Employers* and *Fortune 100 Best Companies to Work For*). Academics considered these lists as a measure of reputation and used these lists for studies investigating a firm's reputation, with the most familiar one being Fortune's survey data (Roberts and Dowling, 2002; Fulmer et al., 2003; Ballou et al., 2003; Fombrun and Shanley, 1990). Fortune's list of top employers, released annually by the magazine since 1982, associated the companies who made it to the list with having a good corporate reputation (Fombrun and Shanley, 1990). Therefore, executives rushed to have their companies rank on these lists for the value and the competitive advantage it would generate (Shrum and Wuthnow, 1988). The benefits of being ranked on Fortune's list align with the benefits of having a good reputation. As such, stakeholders are more willing to exchange resources with that company (Hall, 1992), qualified and loyal talents are attracted to work for these firms (Roberts and Dowling, 2002), and consumers are drawn to purchase their services and products (Ponzi et al., 2011).

Since getting on this list can be viewed as akin to having a high reputation, companies strive to become top employers and make it on Fortune's list. Therefore, it becomes crucial to understand *how* a company becomes a top employer. Several studies have investigated the relationship between corporate reputation and financial revenue, where they confirmed the financial benefits associated with having a good reputation (Fombrun and Shanley, 1990). However, there have been sparse studies on how to build a good reputation, very few looked at whether financial measures played a role and none examined whether a firm's operational performance influenced its reputation. Thus, there is an apparent lack of empirical research on the measures and strategies to adopt toward attaining a good reputation and, more specifically, on how to become a top employer. This research fills this gap by demystifying what it truly takes to become a top employer by examining the relationship between a top employer and their financial and operational measures to identify the areas to invest in as a means to become a top employer. The following section discusses in detail the objectives and methodology of the research work.

1.1 Research Objectives and Methodology

This study examines whether a company's organization and operational decisions would help foretell whether it represents a top employer. The goal is to identify important features of a top employer in terms of top employer to build recommendation on how a company becomes a top employer. In order to achieve this purpose, we propose the following methodology displayed in Figure 1.1.

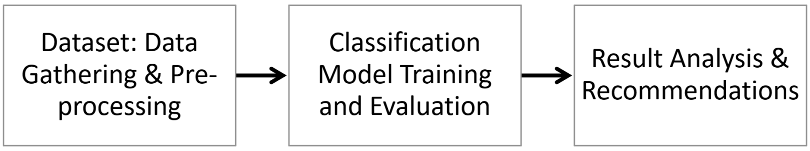
Dataset: Data Gathering & Pre-processing
Classification Model Training and Evaluation
Result Analysis & Recommendations

Figure 1.1: A summary of the overall methodology used in this research work

The first step is to build and prepare the data-set by gathering the list of firms and joining them to their data in Compustat. The next step is to build and train multiple classification models, and then to evaluate their performance towards choosing the optimal classifier that can be used to identify a top employer. Finally, the classifier is analyzed to identify the rules and the associations between the different variables, thus determining the important features that characterize a top employer. Areas of improvement that could increase a firm's reputation are identified to provide firms with recommendations on how to become a top employer.

The remainder of this paper is organized as follows. Chapter 2 summarizes the previous literature on corporate reputation, its relationship with financial performance and the machine learning applications. Chapter 3 discusses gathering the sample with firms from Fortune's list and variables from Computsat accounting and operating data. Moreover, it presents the processing applied on the data to ensure a consistent representation of the observations. Chapter 4 discusses several machine algorithms, such as decision tree, random forest, and logistic regression classifiers. Models are trained and evaluated on the processed data-set and the best model is selected. Moreover, it presents different tuning techniques applied on the classifier. Chapter 5 reports the best classifier and discusses the important features that characterize a top employer. Ultimately, recommendations are developed proposing strategies of investments that would help firms become top employers.

Chapter 2

Literature Review

As a first step in this research, an in depth literature review is conducted to understand what previous research has been done in this scope and identify gaps for the purpose of defining our contribution. This section starts by highlighting the research that has been done in the pursuit of a better understanding of a firm's reputation. It moves on to discuss the previous application of machine learning techniques in different aspects of firms such as analyzing firms performance.

Firm's Reputation

One of the earliest definitions of a firm's reputation is the Fomburn definition. Fombrun defines a firm's reputation as 'A perceptual representation of a company's past actions and future prospects that describes the firm's overall appeal to all of its key constituents when compared with other leading rivals' (Fombrun, 1996). His research suggests that a firm's reputation is the image of which an individual perceives a firm as desirable.

Even though understanding firm reputation has driven considerable academic work in recent literature, there has not been one commonly agreed-upon definition to conceptualize the corporate reputation (Barnett et al., 2006). In an attempt to consolidate the definitions that are attributed to the corporate reputation, Barnett et al. (2006) identifies three distinct clusters of corporate reputation definitions. The first cluster considers the reputation as an awareness of how consumers and stakeholders perceive a firm without making any judgment about it (Fombrun and Shanley, 1990). The second cluster positions the reputation as an assessment of how attractive the firm is based on the judgments and opinions conceived around it (Roberts and Dowling, 2002; Fombrun, 1996; Weigelt and Camerer, 1988). The final cluster characterizes the reputation as an economic, financial , and intangible asset. Indeed, many experts agree that corporate reputation is an intangible and strategic asset (Dierickx and Cool, 1989; Hall, 1993) that results from years of valuing and maintaining strong relationships not only with customers but with internal and external stakeholders (Fombrun, 1996). This affects the strategic choice by generating future rents (Weigelt and

Camerer, 1988) and giving a significant competitive advantage to the firms that possess them (Grant, 1991; Lado and Wilson, 1994).

As such, reputation is found to have significant benefits on the firm that possesses it. More specifically, a good reputation can factor in attracting and retaining a competent quality workforce (Grant, 1991; Lado and Wilson, 1994; Lau and May, 1998; Roberts and Dowling, 2002); understandably, high profile and skilled professionals are attracted and thus seek to work at companies with a good reputation. This is because they are influenced by a company's image, the actions it takes, and the social actions it stands for (Gatewood et al., 1993). Moreover, a good quality of work-life (QWL) that promotes favorable work conditions and a supportive environment would nurture loyal, productive, and motivated employees. As such, a firm's reputation is an essential recruitment, engagement, and retaining tool; a matter recognized by senior executives who are increasingly aware that their most important company assets are, in fact, their human resources (Greening and Turban, 2000).

Moreover, the corporate reputation is found to be associated with a firm's financial performance. This relation has generated attention in various empirical work (Fombrun and Shanley, 1990; Roberts and Dowling, 2002; Dierickx and Cool, 1989; Stuebs and Sun, 2010), which confirmed the existence of a positive relationship between the firm's reputation and several measurements of firm performance. As such, a good business reputation increases the firm's labor efficiency, productivity, and performance (Stuebs and Sun, 2010)). This is possible because even at comparable compensation levels with other market players, a firm with a better reputation attracts better quality employees, who in turn are intrinsically more productive. Consequentially, more productive employees eventually lead to higher labor efficiencies. Moreover, a company's financial history is found to affect its current reputation, which in turn contributes to the persistence of a company's financial success (Roberts and Dowling, 2002).

Furthermore, companies with a good corporate social responsibility (CSR) are considered to have a good reputation. This is because they are driven by environmentally and socially responsible practices that create an image of a good working environment and socially conscious management (Greening and Turban, 2000). This improves a company's perception among its staff and potential recruits. Indeed, CSR programs are credited with giving insight into the company's inner working conditions and its management style, thus molding the general perceptions of how it would be like to work for this firm (Greening and Turban, 2000). The relationship between CSR practices and their financial impact has been thoroughly investigated in recent years. Research suggests that companies doing CSR have better financial performance (Sun, 2012). Also, companies that implement CSR on their chain management activities, such as driving environmentally and socially responsible practices from material sources to customer service, show to have stronger financial performance (Huatuco et al., 2013). These companies compete more effectively and lead to higher growth in sales and profitability (Lau and May, 1998).

Thus, it becomes essential for firms to acquire a good reputation, which makes it crucial to understand how to build a good reputation. Surprisingly, few previous attempts investigated and analyzed what drives a firm's reputation. From the available literature, it appears that the public construes a firm's reputation based on a few indicators, chiefly historical performance, and non-economic cues. Indeed, some variables are found to be positively correlated with the reputation, including a firm's economic performance, as indicated by its prior accounting profitability and the current annual sales (Fombrun and Shanley, 1990). Prior financial performance is found to be a predictor of a firm's future reputation as a socially responsible institution (Hammond and Slocum, 1996). Furthermore, exposure to information through advertisements and media coverage, as well as the use of products and services, also influences a firm's image (Gatewood et al., 1993). Accordingly, a firm's risk-return profiles, responsiveness to social and environmental concerns through corporate responsibility, resource allocations ownership, media exposure (articles and media representation), and corporate diversification (business segment) are found to be drivers for its reputation (Fombrun and Shanley, 1990; Hammond and Slocum, 1996; Gatewood et al., 1993). Further research needs to be done to shed more light on the drivers of a reputation, by increasing the number of potential indicators to be considered as drivers and by examining their association with a firm's reputation. Investigation needs to be made to identify whether operational performance indicators are drivers of a firm's reputation. Hence, this study offers additional research efforts that serve to clearly identify what drives a reputation by examining operational and accounting measures, and investigating their relationship with corporate reputation.

Machine Learning

So far, research has extensively used machine learning (ML) to predict a company's financial performance and to handle financial decision-making problems. Machine learning refers to a set of algorithms and statistical models that learn from historical data using underlying patterns in the data to predict on new data (Langley, 1996). Machine learning algorithms are used in various applications, from healthcare and image recognition to financial market analysis. Some studies investigated the use of neural network algorithms (which is a ML algorithm) to predict a firm's financial performance using the firm's financial indicators (Lam, 2004). Other studies tested multiple models to compare the prediction performance and accordingly select the best performing model. Artificial Neural Network (ANN), Support Vector Machine (SVM), random forest and naive-Bayes were applied to predict the direction of movement of stock and stock price index for Indian stock markets (Random Forest had the best performance for this prediction) (Patel et al., 2015), whereas other models were studied to predict a corporate's bankruptcy (Aziz and Dar, 2006). A machine learning technique, the classification models, are learning models used to identify to which set of categories a new observation belongs to. Among the most popular classification algorithms are Decision Trees, Logistic Regression, and Random Forests. Several of these algorithms were developed for application on financial data (Patel et al., 2015; Aziz and Dar, 2006) and very few tackled the corporate reputation. Despite some sparse research looking at the correlation of a firm's operating and accounting variables with its reputation, there has not been any study to date investigating the use of machine learning models to classify a firm's reputation as a top employer. In this study, the machine learning algorithms are investigated, and more specifically the classification models, in an attempt to classify a firm as top employer or not top employer using the firm's financial and operational data.

The purpose of this research study is to fill the gap in the corporate reputation literature by examining the drivers, in both financial and operational practices, as different predictors of a firm's reputation using machine learning classification models. In an attempt to focus on a more direct measure of employees' satisfaction, we use workplace branding and analyze top employers, identified through employee surveys that aim to measure workplace branding and examine common best practices. For this purpose, we select firms from Fortune top employers' surveys, and then, gather their financial and operational data from Compustat. Then, we test different machine learning models to select the best model, as per specific criteria, that would allow classifying a firm as a top employer. Moreover, we analyze the rules and associations behind this classification to determine which activities contribute to increasing a firm's reputation. Finally, we provide a methodological framework to guide firms into becoming top employers by adopting recommended strategies to help them advance their internal operational processes.

Chapter 3

Data Preparation

3.1 Measurement of Firm's Reputation

For this study, firms included on Fortune's 100 Best Companies to Work For (Fortune, 2020) are assumed to have a good reputation. Fortune has been publishing this list since 1998 around the months of January and February of each year for companies across more than 32 industries. Examples of such industries include real estate, manufacturing, mining, retail trade and services among others. To be considered as a top employer, Fortune in partnership with the Great Place to Work Institute (GPWI) considers applications of companies with more than 1,000 employees and then calculates scores for these firms based on employees' responses to a sixty question survey. This survey mainly focuses on the employee's experiences of trust and their ability to reach their full human potential in the respective company they work in. It is important to note that a quarter of Fortune's criterion is actually based on the HR firm's data programs and policies submitted by the company itself.

To test that Fortune's list is indeed a measure of a firm's reputation that classifies employers as best companies to work for, Fulmer, Gerhart et al. (2003) compared responses from firms on Fortune top 100 with responses of employees from clients of other firms (Hewitts Associates). Employees were asked whether they would like to keep working for the same company in the coming year. Based on the answers that they received, employees seems to prefer working for firms listed on Fortune's list and are more willing to stay working for them than for others not on the list. Fortune's list is extensively used as a measure of the firm's reputation in studies examining the corporate reputation and investigating the relationship between a firm's reputation and its financial performance (Hammond and Slocum, 1996; Wang and Smith, 2008; Blazovich et al., 2013; Fulmer et al., 2003; Fombrun and Shanley, 1990; Roberts and Dowling, 2002). Sufficient proof is provided to assume that businesses with a good reputation lead to better financial capability (Roberts and Dowling, 2002) and to conclude that the market value of the 100 best companies exceeded those of other firms to enjoy performance advantages over the broad market (Fulmer et al., 2003). As such for this study, companies listed on Fortune's 100 Best Companies to Work For survey represents the reputation measure to identify top employers.

3.2 Sample Selection

Financial and Operational data

To investigate what drives the corporate reputation, past financial and operational measures of a firm are examined to determine what helps to classify a firm as a top employer. As such, each company listed on Fortune's survey, representing a top employers, is matched to its data in Compustat. Accounting, market-based and operational measures are obtained from Standard & Poor's Compustat database for the publicly traded top employers. Compustat's data includes various indicators that belong to the balance sheet, the income statement data and the company identifiers (Example the standardized industrial classification (SIC) which characterizes the business segment of a firm) among others (Fombrun and Shanley, 1990; Capkun et al., 2009; Huatuco et al., 2013).

Matching Year Value to Firm Data

The general methodology used by Fortune to place a company on their list has a time delay between the list publishing date and the firm's data and evaluation date. This is due to the fact that Fortune gathers the firms' required data from annual reports dating to two years preceding the list publishing date (Blazovich et al., 2013). For example, Fortune published the list for the 2020 top employer's in February 2020, however, Fortune examined these firms during the previous year (in 2019) and they collected the firms' data for two years preceding the list's date (in 2018). Therefore, the time lag between the year of the firm's data on which the evaluation is done and the year the list is published, is between 15 and 24 months. As such, in this study, the data gathered from Compustat is shifted by two years, and Fortune's top employers are matched with their data in Compustat dating two years before making it on Fortune (Example, a top employer that listed in 2020 is matched with its data in Compustat for the 2018 fiscal year) (Blazovich et al., 2013).

Fortune's 100 Best Companies to Work For is a list which is published annually as such, a company's ranking as a top employer on Fortune's list can change throughout the years. Moreover, the data of these companies retrieved from Compustat represents the fiscal year statement which is also declared on an annual basis. As such, a top employer's classification is defined by the firm itself as well as the year it was listed. This suggests that the observations used in this study for classification of firms is represented as a firm-year observation (Roberts and Dowling, 2002; Capkun et al., 2009; Sun, 2012; Huatuco et al., 2013).

3.2.1 Building the Data-set

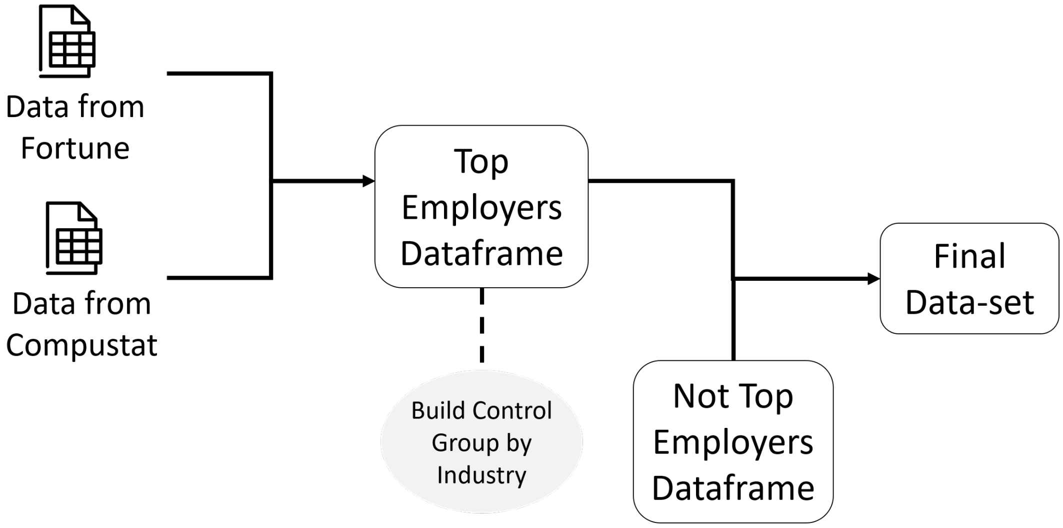

Figure 3.1: Block diagram that summarizes the main steps used in building the data-set

Top Employers Selection

To form the data-set for this study, the first step is to build the top employer's data-set. As per Figure 3.1, the top employers data (Data from Fortune) is matched with their Compustat data (Data from Compustat) using the firm's ticker symbol (a unique identifier of the firm) and the ranking year (taking into account the time lag discussed earlier in this section). Because Fortune's list includes private and public firms, the data for the private top employers is not possible to obtain from Compustat and these companies are discarded. after matching with Compustat, the top employer data-set has 600 firm-year observations.

Control Group Selection

The next step is to form a control group for the top employers. First, the industry and the listing year of the top employers are selected from the top employers data-set as per Figure 3.1. Then, this criteria is used to select companies in Compustat that have the same industries of the top employers at the year of the listing. This operation is performed using the first two digits of the SIC code (a variable in Compustat) to match on the industry and the year of the listing taking into account the two years lag effect. Using the result of this match, further refinement is performed to exclude firm-year observations that are listed on Fortune to avoid including any top employer in the control group. Moreover, firm-year observations that don't have their complete data available in Compustat are discarded. Finally, to ensure a balanced data-set in this study, the number of firm-year observations in the control group is reduced to an equal number of the ones in the top employer data-set. As a result, the control group has 600 firms-year observation labeled as not top employers.

Final Data-set

The sample data-set has 1200 firm-year observations distributed as an equal number of 600 top employers and 600 not top employers. The yearly data from 2006 to 2020 has an equal number of associated top employers and not top employers. Table 3.1 reports the distribution of the number of firms (top employers and not top employers) over the years and the percentage of firms each year out of the total number of observations. It is clear that the number firms is quasi equally distributed over the years with a 5.5% to 8.1% of representation per year.

	# of Firms	% of Total
	Per Year	Observations
2006	64	5.33%
2007	70	5.83%
2008	74	6.17%
2009	74	6.17%
2010	76	6.33%
2011	80	6.67%
2012	92	7.67%
2013	84	7.00%
2014	86	7.17%
2015	82	6.83%
2016	76	6.33%
2017	88	7.33%
2018	96	8.00%
2019	84	7.00%
2020	74	6.17%
Total	1200	100%

Table 3.1: Table showing the number of firms and percentage of observations per year out of total number of observations ($\frac{\#ofobservationsPerYear}{\#ofObservationsforAllYears}$) in the final data-set

3.3 Data Pre-processing

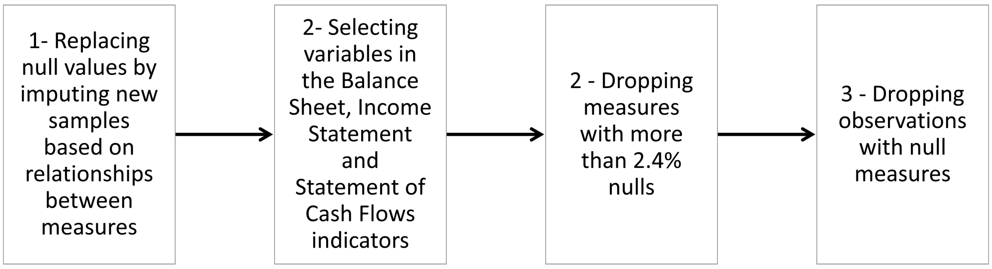
1- Replacing null values by imputing new samples based on relationships between measures
2- Selecting variables in the Balance Sheet, Income Statement and Statement of Cash Flows indicators
2 - Dropping measures with more than 2.4% nulls
3 - Dropping observations with null measures

Figure 3.2: Block diagram describing the different steps applied in pre-processing the data. This includes cleaning the data, imputing new samples and dropping observations based on well defined relationships and assumptions.

The initial data-set is composed of 578 variables across 1200 observations. However, some variables in Compustat were not reported for the firm-year observations which resulted in a 34% of missing values in the data-set. These missing values need to be handled to ensure a complete data representation of each observation. This is done by first, imputing missing values based on formula recalculation. Then, the measures are reduced to include only those that are of interest to this study. Finally, with no other choice to manage residual nulls, any firm-year observation that is mapped to at least one null measure is discarded.

Imputation by Recalculating Formulas

The variables retrieved from Compustat are part of the balance sheet, income statement and the statement of cash flow. The definitions of each of these variables is examined to check how companies are reporting them. This info is captured from the variables' definition in Compustat (Standard&Poor's, 2003). As such, the formula of each variable is retrieved to identify the items that compose it or whether it is derived from another variable. The mapping of each variable is done to form a list with the equations between the variables in the data-set. Then, the nulls in each equation are flagged to attempt imputing the missing values by recalculating the associated formula. This process is possible in case there is at most one null variable in a a certain formula. In this case, the formula is recalculated allowing to identify the missing value and then to assign it to the variable accordingly.

After iterating on the variables to validate their equations and impute whenever it is applicable, the percentage of null values in the data-set decreased from 34% to 17.1%. To note that measures with more than 90% of null values were dropped. Figure 3.3 reports the variables with the highest percentage of nulls and compares the initial percentage of null to the percentage of nulls after imputation by formula calculation. Some of the variables have a decrease of more than 35%

in their missing values.

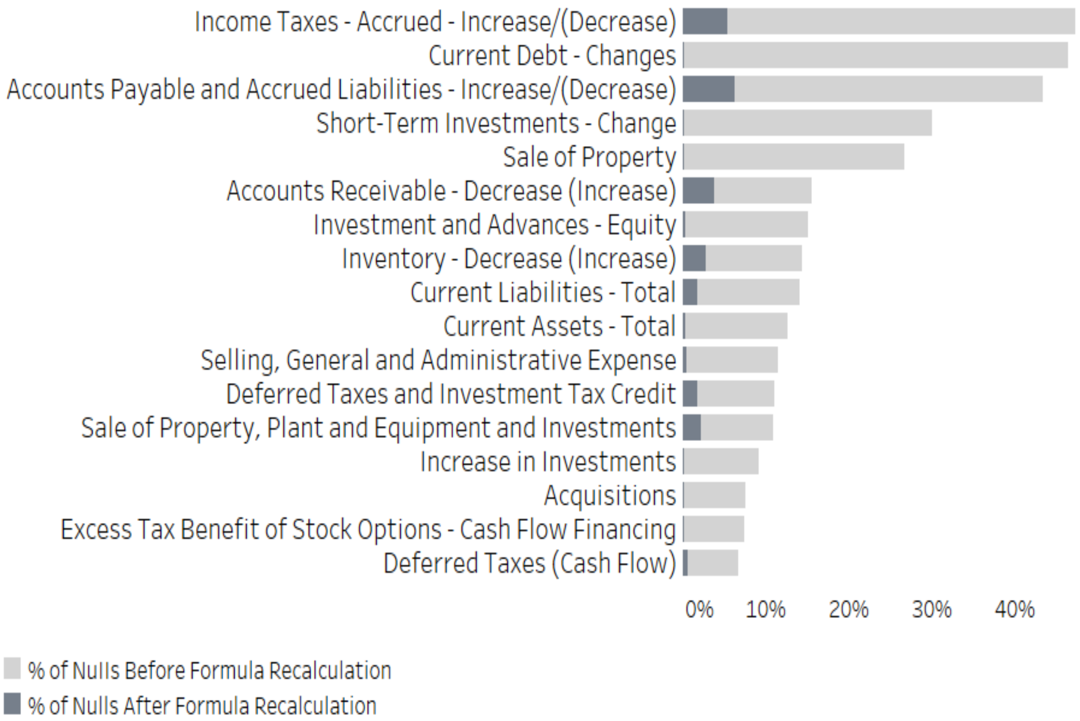

Figure 3.3: Bar chart showing the percentage of nulls decrease in some variables after recalculating their formulas

Refining to Balance Sheet, Income Statement and Statement of Cash Flows

Companies report their data in Compustat concerning many aspects of their companies. Some of the reported measures represent identical type of information but with different measurements methods or units (for example same measures are reported as the fiscal year and the calendar date). Other reported measures are identifiers and descriptors of the company that don't serve this study (examples include multiple address variables and website of the firm). Furthermore, some variables are stratification of others that goes into more than three levels of detail. Analyzing the resulting dataframe showed high correlation between variables that are related in terms of, for example, (1) same metric measured differently across firms and (2) linear equations used to derive different measures. To tackle this, the basic measures from the balance sheet (BS), income statement (IS) and statement of cash flows (CF) are kept to only include possible predictors of a firm's reputation, and thus reduced the number of variables from 375 to 86 variables.

Discarding Columns and Observations with Nulls

After imputing by formula recalculation as well as selecting the relevant measures in the data-set, nine columns still presented more than 2.4% of missing values. Since there is not a possible imputation to be performed on these columns, it was decided to drop them. As such, the overall percentage of missing values in the data-set dropped to 0.3% null values. Furthermore, 120 observations had at least one missing variable and had to be discarded to ensure a complete representation of each observation in the data-set.

	% of Nulls	# of Variables	# of Observations
— Initial Data	34.0%	578	1200
1 - Recalculating formulas	17.1%	375	1200
2 - Refining to BS, IS and CF	1.9%	86	1200
3 - Dropping measures > 2.4% nulls	0.3%	77	1200
3 - Dropping observations with nulls	0%	77	1080

Table 3.2: Table showing the percentage of null values, number of variables and number of observations for all firm-year samples. The values are shown after each pre-processing step explained in the block diagram of Figure 3.2

Table 3.2 reports the process of imputation starting with the formula recalculation that dropped the nulls to 17.1%, the restricting to essential measures dropping the percentage to 1.9% and the number of variables from 375 to 86. Finally, dropping observations and measures having missing values reduced the size of the data-set to 77 variables and 1080 observations.

After processing the data, Table 3.3 reports the firms yearly distribution by top employers and not top employers. The data-set is still representing a balance between the number of top employers and not top employers.

Industries

The sample data-set includes firms that belong to several industries. Industries are identified using the 2 digits code of the standardized industrial classification (SIC) in Compustat which characterizes the business segment of a firm (Wang and Smith, 2008; Fulmer et al., 2003; Blazovich et al., 2013; Stuebs and Sun, 2010). Table 3.4 represents the distribution per industry of the firm-year observations. There are nine industries in the data-set with the manufacturing and the services industries presenting more than 65% of total top employers.

Year	Pre-processing		Post-processing	
	Not Top Emp	Top Emp	Not Top Emp	Top Emp
2006	40	40	32	27
2007	42	42	34	29
2008	43	43	37	32
2009	45	45	35	30
2010	42	42	36	33
2011	45	45	38	33
2012	51	51	43	38
2013	46	46	37	35
2014	48	48	37	36
2015	45	45	36	40
2016	44	44	37	35
2017	49	49	43	41
2018	54	54	45	42
2019	48	48	38	35
2020	41	41	35	31

Table 3.3: Table showing the class distribution of firms between top employers and not employers to compare the balance between these classes before and after processing the data

	# of Observations	% of Totals
Manufacturing	384	35.6%
Services	312	28.9%
Retail Trade	158	14.6%
Finance, Insurance and Real Estate	103	9.5%
Mining	63	5.8%
Transportation, Communications, Electric, Gas	54	5.0%
Construction	4	0.4%
Wholesale Trade	1	0.1%
Agriculture, Forestry and Fishing	1	0.1%

Table 3.4: Table showing the number of firm-year observations that belong to each industry and the percentages of observations in each industry out of the overall observations

Variables

This study uses several performance variables from Compustat statements to investigate the classifiers of a firm's reputation. Variables are indicators of the firm's accounting and operational performance. Relating to previous research studies on the relationship between a firm's indicators and its reputation, additional variables are calculated and included in the data-set. Accounting indicators are calculated using Compustat variables, the return on assets (ROA) to measure how efficiently a firm uses its assets (Roberts and Dowling, 2002; Fulmer et al., 2003; Huatuco et al., 2013), the return on equity (ROE) to measure corporate performance to generate returns for their equity investors (Huatuco et al., 2013) and the return on income (ROI) to measure the firm's asset productivity (Stuebs and Sun, 2010). Moreover, the market to book ratio (mtb), a measure widely used in studies of reputation and performance, is also included to represent the proportion of a firm's asset base that is in intangible form and to capture the market's expectations of anticipated future economic returns (Roberts and Dowling, 2002; Fulmer et al., 2003).

Compustat variables are scaled by the assets and the assets variable is represented by its natural logarithm. Table 3.5 lists the category of each variable whether it represents an operational or an accounting indicator.

Accounting Measures
Assets - Total
Liabilities - Total
Long-Term Debt - Total
Property, Plant and Equipment - Total (Net)
Stockholders Equity - Parent
Net Income Adjusted for Common/Ordinary Stock (Capital) Equivalents
Capital Expenditures
Investing Activities - Net Cash Flow
Sales/Turnover (Net)
Financing Activities - Net Cash Flow
Market to Book Ratio
Return on Equity
Return on Investment
Return on Assets
Net Income (Loss)

Operating Measures
Cost of Goods Sold
Selling, General and Administrative Expense
Operating Income Before Depreciation
Operating Activities - Net Cash Flow

Table 3.5: Table showing the stratification of indicators by accounting or operating measures

Furthermore, two industry control variables are included. The first one is the industry code identified by the SIC two digits codes to control the inter industry differences in the sample. The second one is the industry turbulence variable to control for deviation in industry sales in the past three years (Sridhar et al., 2014) Finally, the size of the firm is measured by the total sales indicator in Compustat (Roberts and Dowling, 2002; Sridhar et al., 2014).

	min	max	mean	std
Assets - Total	0.004	1,119,796	31,511	104,865
Liabilities - Total	0.028	1,069,731	22,845	93,319
Long-Term Debt - Total	0	236,027	5,782	21,248
Property, Plant and Equipment - Total	0	247,101	5,296	19,142
Stockholders Equity - Parent	-8,446	191,794	8,559	20,031
Sales/Turnover (Net)	0	420,714	12,209	29,048
Cost of Goods Sold	0	339,228	6,964	21,419
Selling, General and Admin. Expense	0	32,576	2,233	4,372
Operating Income Before Depreciation	-7,236	65,769	3,012	6,835
Net Income (Loss)	-6,132	44,880	1,461	3,566
Net Income Adj for Common Stock	-6,173	44,880	1,452	3,553
Capital Expenditures	0	37,985	1,073	3,382
Investing Activities - Net Cash Flow	-50,854	15,324	-1,718	4,671
Operating Activities - Net Cash Flow	-68,197	56,170	2,314	6,186
Financing Activities - Net Cash Flow	-41,078	73,390	-392	4,875
Market to Book Ratio	-942.87	49.78	2.58	30.26
Return on Equity	-7.37	3.35	-0.001	0.38
Return on Investment	-29.80	13.16	0.06	1.15
Return on Assets	-397.43	390.72	-0.58	18.89

Table 3.6: Summary statistics that show the minimum, maximum, standard deviation and mean values of the samples calculated across all firm-year observations

Table 3.6 presents a statistics summary for the key variables. Firms included in the dataframe have a mean ROA score of -0.58, mean ROE sore of -0.001 and mean Market to Book Ratio score of 2.58. The mean of total assets is 31,511 with a standard deviation of 104,865 and the mean sales is 12,209 with a standard deviation of 29,048.

Table 3.7 reports the list of variables along with the definition of each one.

Balance Sheet	
Assets - Total	Total assets
Liabilities - Total	Total liabilities
Long-Term Debt - Total	Debt obligations due more than one year from the company's balance sheet date
Property, Plant and Equipment - Total (Net)	Cost of tangible fixed property used in the production
Stockholders Equity	Common equity
Income Statement	
Sales/Turnover (Net)	Total sales
Cost of Goods Sold	Costs directly allocated to production, such as material, labor and overhead
Selling, General and Administrative Expense	Commercial expenses of operation
Operating Income Before Depreciation	Total income from normal business operations
Net Income (Loss)	Income or loss after subtracting expenses and losses
Net Income Adjusted for Common/Ordinary Stock (Capital) Equivalents	
Capital Expenditures	Funds expenditure
Statement of Cash Flows	
Investing Activities	Cash received or paid from investing activities
Operating Activities	Net change in cash from Operating Activities
Financing Activities	Cash paid or received from Financing Activities
Calculated Fields	
Market to Book Ratio	Price per share x Shares outstanding / total shareholder's equity
Return on Equity	Net income / Common equity
Return on Investment	Net income / Invested Capital
Return on Assets	Net income / Total assets
Industry Turbulence	Coefficient of variation of industry sales
Industry	
Industry Code	2 digits code of the standardized industrial classification (SIC)
Industry turbulence	Standard deviation of annual sales of all firms in an industry for the previous three years

Table 3.7: Table showing the definition of each variable present in the resulting dataframe. It should be noted that these are the formal definitions used in Compustat (Standard&Poor's, 2003)

Chapter 4

Building the Classification Model

4.1 Classification Models

The first goal of this study is to classify whether a firm is as a top employer or not top employer (i.e. identify top employers) using the firm's Compustat data. This research work employs a method that was not explored yet, machine learning classification algorithms, to perform this task. Classification models extract features from a given data to allocate a firm to one of these two classes, top employer or not top employer. First, the classifier needs to learn from labeled data by approximating the mapping between the variables and the label (the training process). Then, the model is evaluated on new data-sets to judge the performance of this model and whether it correctly classifies new firm-year data samples. A second goal of this study is to understand how a firm can reach a status of top employer (*how to become a top employer*) by extracting the rules which lead to the top employer classification, and hence, to understand the conditions that help a firm to attain such a status.

In this study three classifiers are evaluated, decision tree, random forest and logistic regression. These classifiers are well known to provide a good classification performance as well as to produce an interpretable classification result. As a first step, these classifiers run on the data-set to choose the best performing model, then, this model can be applied to classify the firms as top employers. Furthermore, the rules are extracted from this classification to understand these associations and to extract important features in the classification of a top employer.

Decision Tree

Decision trees are a set of "if-then" statements that are used to predict a given quantity or classify a record (Loh, 2011). Decision trees are built by algorithms that identify different ways of splitting the data, as such a set of rules sequentially evaluated up until reaching the result and a record is assigned to the most likely category. Decision trees are represented graphically as a tree, constructed by many nodes and branches. The rules derived from the branches of a tree makes the decision trees easy to interpret and facilitates understanding how the prediction was made (Syed et al., 2019).

Random Forest

Random forest is a classifier that combines a set of decision tree classifiers. Using random observation and features to build these trees, each tree is an independent and equally distributed classifier that casts a vote to choose a classification. Majority voting between the outputs of each tree classifier is then used to determine the final predicted value of the random forest. This behavior leverages the power of multiple decision trees leading to an improved model performance and better accuracy (Breiman, 2001).

Logistic Regression

Logistic regression is one of the most widely used traditional statistical algorithm method for binary classification using a logistic function. In producing the LR equation, the maximum-likelihood ratio determines the statistical significance of the variables (Hosmer and Lemeshow, 2000). This model has been proven to be very effective especially for solving relatively less complex problems.

Performance Measurements of Classification Models

		Classified	
		Not Top Employer	Top Employer
Actual	Not Top Employer	True Negative (TN)	False Positive (FP)
	Top Employer	False Negative (FN)	True Positive (TP)

Table 4.1: Confusion Matrix

The performance of the classification is evaluated using a confusion matrix represented in Table 4.1. The matrix is used to compare the result of the classification with the actual class of the observations.

There are several performance criteria usually used to evaluate the performance of the classification models out of which the accuracy, the recall and the specificity. The confusion matrix and these indicators are used in this study to evaluate the classification of the decision tree. False positives can be tolerated in this study whereas the focus is to seek a high number of true positives, because this would mean correctly identifying a large number of firms, thus serving the purpose of the study to solidify the recommendations to be provided.

- The accuracy measure is defined as the percentage of the correctly classified observations out of the total number of observation as per equation (4.1).

It measures the correctly classified observations.

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

- The recall (sensitivity) is defined as the percentage top employers correctly classified out of the the total actual top employers as per equation (4.2). It measures the proportion of top employers who are correctly classified.

$$\frac{TP}{TP + FN} \tag{4.2}$$

- The specificity is defined as the percentage of not top employers correctly classified out of the total actual not top employers as per equation (4.3). It measures the proportion of not top employers correctly identified.

$$\frac{TN}{TN + FP} \tag{4.3}$$

4.2 Building the Models

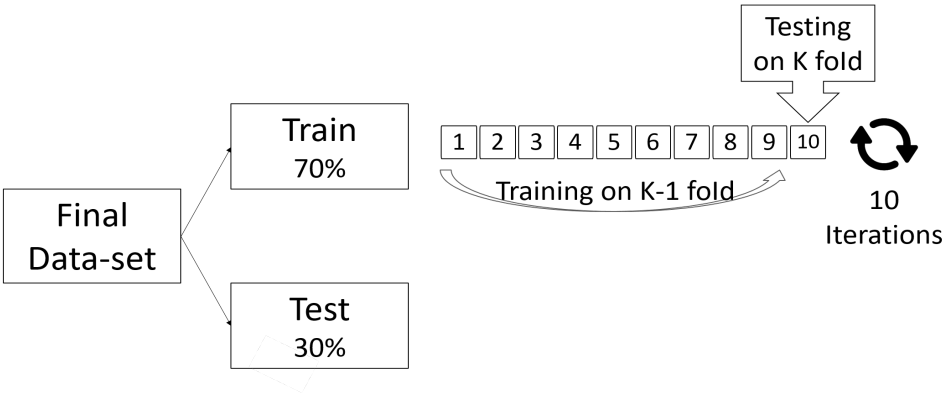

Figure 4.1: Diagram that summarizes the main steps used in building and training the models

The same process is followed to build the three classifiers, decision tree, random forest and logistic regression, as displayed in Figure 4.1. First, the data is split in 30% testing (324 observations) and 70% training (756 observations). Then, the classifiers run on a 10 fold cross validation to generalize well on unseen data and to avoid over-fitting. As such, the training data is split into 10 equal sized and random subsets. One subset is retained as a validation and the other 9 are used for training. The accuracy of the models is evaluated on the test fold, and this process is repeated until each subset serves as the test fold and the score is calculated on that fold. The scores of all 10 subsets are averaged to get the final cross validation score.

The classification results of the three models are presented in Table 4.2 and Table 4.3. The random forest classifier has the highest cross-validation accuracy of 90.21% followed by the decision tree classifier with 81.5% and the logistic regression classifier with 75.14%. Furthermore, the models are tested on the testing data to evaluate the accuracy, the recall (sensitivity) and the specificity of the model. The random forest prediction has the highest accuracy and recall on the test data with 88.58% and 90% respectively, followed by the decision tree with 83.33% and 79.33% respectively, and the logistic regression with 75.3% and 76% respectively.

	Accuracy Cross-Validation	Accuracy on Test
Random Forest	90.21%	88.58%
Decision Tree	81.50%	83.33%
Logistic Regression	75.14%	75.31%

Table 4.2: Table showing the accuracy obtained from cross-validation and from running the decision tree, random forest and logistic regression on the test set

	Recall	Precision	Specificity	F1-Score	TP	FN	FP	TN
Random Forest	90.00%	85.99%	87.36%	87.95%	135	15	22	152
Decision Tree	79.33%	83.80%	86.78%	81.51%	119	31	23	151
Logistic Regression	76.00%	72.15%	74.71%	74.03%	114	36	44	130

Table 4.3: Table showing the classification performance of decision tree, random forest and logistic regression on the test set

Although the random forest classifier has higher accuracy than the other two models, it is not the sole criteria to consider for deciding on which model to adopt:

1. The recall (sensitivity) of the model is a major criteria to consider having a higher significance over the specificity and the accuracy of the model. The first goal of this research is to classify firms according to their reputations, hence it is important to identify a high number of top employers correctly, which is why it is essential to have a good recall even at the cost of having few wrongly classified top employers.

2. The classifier should be interpretable. The second goal of this study is to understand the rules that defines a firm becoming a top employer, this is why the classifier should be decomposable and from which the associations between the variables and the top employers can be extracted.

Figure 4.2 reports the comparison between the performance of the three classifiers using the cross-validation accuracy, the accuracy on the test data and the sensitivity. Unsurprisingly, random forest classifier, known to outperform other classifiers in predictions, has the higher sensitivity and accuracy. However, since the random forest combines multiple decision trees, with each tree trained on a random selection of features, the interpretation of the prediction result and the association between the features becomes extremely challenging if not impossible. The decision tree classifier has the second highest accuracy and sensitivity. Moreover, the decision tree classifiers are known to be effective because they deliver good accuracy and provide human-readable rules of classification (Mantovani et al., 2018). As such, the decision tree is chosen to classify firms as top employers.

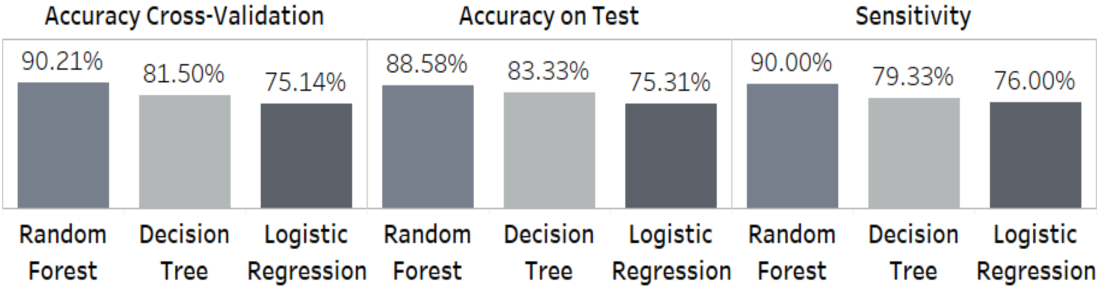

Figure 4.2: Bar chart comparing the cross-validation accuracy, test accuracy and the sensitivity of the three models

4.3 Tuning the Model

The decision tree classifier can be further optimized by tuning the hyper-parameters of the model. This is achieved by searching a set of values for these parameters that could optimize the model architecture and hence its performance.

The ultimate goal of this study is to analyze the classification results of the model to understand how a top employer is classified. For this purpose, a visual representation of the tree needs to be generated to allow examining the rules behind the classification and the association between the features. Therefore, the illustrated decision tree should be interpretable and comprehensible while maintaining a good classification performance. One technique widely used tackles decreasing the complexity of the tree by reducing the size of the tree and removing sections. This is labeled as pruning the tree by tuning some of the hyper-parameters. Pruning is usually performed to improve the accuracy of the classification, as well as to simplify the structure of the tree allowing a better visual interpretation. The maximum depth and the minimum leaf sample number are the decision tree hyper-parameters which control the size of the tree (Mantovani et al., 2018). The repercussion of modifying the values of these parameters is evaluated using the classification accuracy, the sensitivity and the F1-score. As such, choosing the values for the two hyper-parameters relies on whether tuning the tree parameters would produce a simplified tree to ease the interpretation without compromising its performance.

Maximum Depth

The maximum depth of a decision tree denotes how deep a tree can grow. With this parameter left unspecified, all nodes will expand until all leaves are pure. However, the deeper a tree grows, the more complex it becomes by splitting on more data. Since growing the tree is performed on the training data, the tree risks to become a perfect fit for the training data and thus could not be generalized on the test data. One way to resolve the overfitting of the tree and to simplify the tree, is to tune the hyper-parameters by reducing the maximum depth parameter of the tree.

To decide on which parameter value to use, the classification accuracy is investigated when the maximum depth parameter varies between a depth of 2 and 30. These accuracy values are benchmarked against the 83% accuracy taken from the initial model. The results are displayed in Figure 4.3 shows that when the maximum depth increases, the accuracy increases as well until it reaches a maximum of 83.64% associated with a depth value of 9. This accuracy surpasses the initial model's accuracy with a simpler tree. Moreover, starting a maximum depth of 15 the accuracy of the tree reaches a constant value of 83.33%.

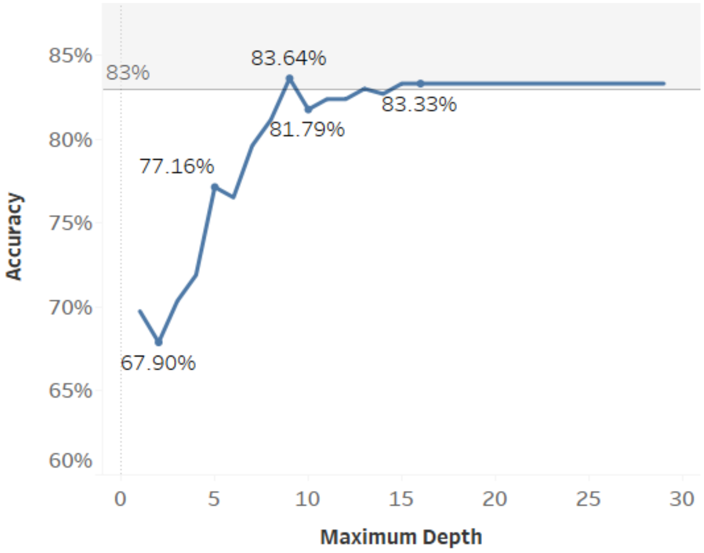

Figure 4.3: Line chart displaying the variation of the cross-validation accuracy for different values of the maximum depth parameter

Minimum Leaf Sample

The minimum leaf sample parameter is the minimum number of sample required to have at a leaf node. Splitting the nodes is performed only when the sample number in the leaf is greater than the minimum leaf sample parameter. The ideal number of sample in a leaf tends to be between 1 to 20 (Mantovani et al., 2018), but in this study the range is increased to a minimum leaf sample of 70. This is because the goal of the study is to examine the classifications' rules which are applied on the top employer leaves. As the number of samples in the leaf increases, the more generalized the recommendation becomes. Thus the purpose is to increase this parameter as possible without compromising the performance of the tree. The results are displayed in Figure 4.4. When the minimum leaf sample increases, the accuracy decreases until it reaches a minimum of 67% associated with a sample value of 50 and above. Moreover, for a low minimum leaf sample value range between 3 and 12, the accuracy is between 81% and 82%, but then, the accuracy quickly decreases once the minimum leaf sample surpasses a value of 12.

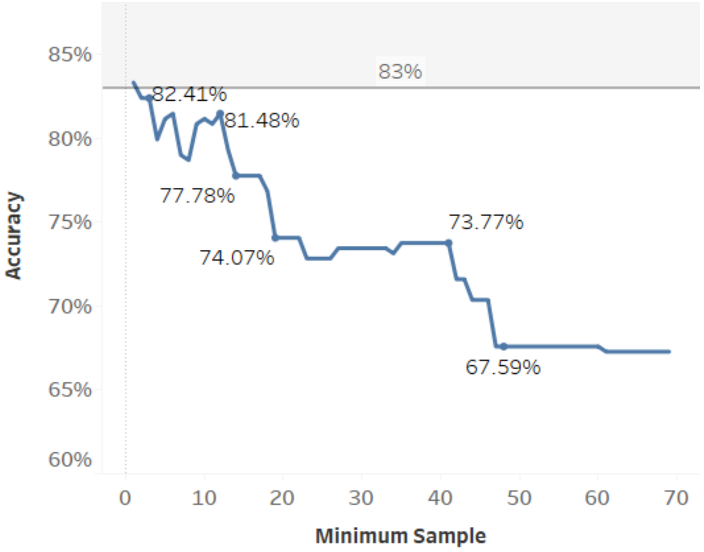

Figure 4.4: Line chart displaying the variation of the cross-validation accuracy for different values of the minimum leaf sample parameter

Grid-Search

After examining the maximum depth and minimum leaf sample individually, the results of the optimal parameters obtained are 9 for the maximum depth and 12 for the minimum leaf sample. However, the tree performance and complexity should be evaluated using a combination of these two parameters. The grid-search is a technique to evaluate combinations of several parameters' values to select the best tree within the set of parameters' values provided. As a first run, the grid-search is performed using the maximum depth and the minimum leaf sample parameters by specifying a large set of values for these parameters, between 1 and 17 for the maximum depth and between 1 and 30 for the min sample values. The result of this grid search is a tree with a cross validation accuracy of 83.22% which is better than the initial cross validation accuracy obtained (81.5%). However, the resulting tree has 107 nodes with a depth of 8 which still represents a complex tree which prevents visual interpretation.

In an attempt to simplify the tree, a new grid-search is performed using restricted parameter values. The values are determined by looking at the previous section's individual interpretation of the parameters to decrease the upper limit of the maximum depth and increase the lower limit of the minimum sample thus ensuring a simpler tree structure. As such, the grid search is executed with a combination of maximum depth values lower than 9 and minimum sample values greater than 11. The result is a tree with a maximum depth of 5 and a minimum leaf sample of 15 which has a cross-validation accuracy of 76.86%. Despite the drop in the accuracy value, better judgment can be made by looking at the sensitivity which returned a value of 75.33% dropping by 4% from the initial tree's accuracy. Furthermore, the tree resulting from the second grid-search has an extensive drop in the number of nodes from 145 initially to 39 nodes.

Minimal Cost-Complexity

A different method to prune the tree is the cost-complexity technique. Contrarily to the previous tuning method, this technique prunes the tree after it is built, whereas, the minimum leaf sample and the maximum tree depth parameters prune the tree before it is built. The cost-complexity technique assigns to each node a value, alpha, to indicate the weakness of the link. It then recursively finds the nodes with the smallest alpha to prune. By default alpha is set to zero which indicates that no pruning is done. When alpha increases, more of the tree is pruned and thus the number of nodes as well as the tree depth decreases. To find the optimal value for alpha to use, the performance of the tree is evaluated as the alpha value changes. This technique is applied on the tree that resulted from the second grid-search tuning with a maximum depth of 5 and minimum leaf sample of 15.

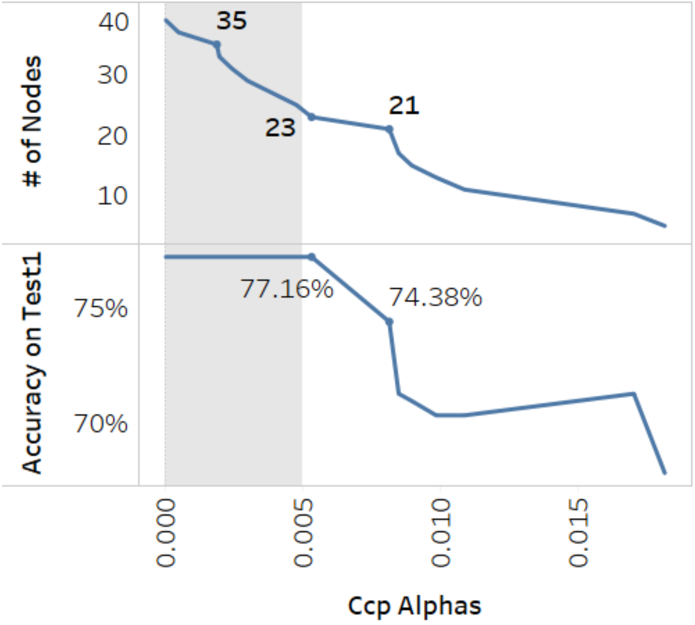

Figure 4.5: Line charts displaying the variation of the cross-validation accuracy and the number of nodes for different values of the minimal cost-complexity parameter

Figure 4.5 shows that as the value of alpha increases, the number of nodes in the tree decreases sharply, whereas the accuracy decreases in a slower manner. To choose the value of alpha to use, the accuracy on test is examined to select the highest value of alpha that preserves the accuracy of three at 77.16% previously obtained. Using 0.005 as a value for the alpha parameter, the accuracy of the tree remains at 77.16%, while the number of nodes decreases from 39 to 25 nodes.

Performance Comparison of Tuning Methods

After tuning the tree using several methods with grid search for the maximum depth and the minimum leaf sample, as well as the post pruning using the alpha parameter, the performance of each resulting tree is displayed in Table 4.4. Despite the initial model and the optimal grid-search model having high-performance indicators with a cross-validation accuracy of 81.5% and 83.22% respectively and a recall of 79.33% and 77.33% respectively, they show a sophisticated structure with more than 100 nodes. Since the second goal of the study is to interpret the rules and the association between the variables of a top employer, it seems challenging to interpret these two trees considering their complexity. Moreover, the third pruned tree is done using refined grid-search parameters' values that are more likely to produce a simpler tree. This has an impact on the performance indicators where the cross-validation accuracy drops by 6.36% and the recall by 2%, whereas, as expected, it simplifies the tree where the number of nodes decreases from 107 to 39 nodes. Finally, the last pruning technique is to set a complexity cost to the tree using the parameter alpha with a value of 0.005. The accuracy and recall of the last tree remain equal to those of the previous one, but the number of nodes decreases further to 25 nodes.

	Before Tuning	Grid-Search with Large Set of Parameters' Values	Grid-Search Refined *(min samp $\geq$ 11 & max depth$\leq$ 9)*	Grid-Search Refined with alpha=0.005 *(min samp=15 & max depth=5)*
Max Depth	15	8	5	5
Min Sample	1	1	15	15
Accuracy CV	81.50%	83.22%	76.86%	75.66%
Accuracy Test	83.33%	81.17%	77.16%	77.16%
Recall	79.33%	77.33%	75.33%	75.33%
Specificity	86.78%	84.48%	78.74%	78.74%
Precision	83.80%	81.12%	75.33%	75.33%
F1-Score	81.51%	79.18%	75.33%	75.33%
TP	119	116	113	113
FN	31	27	37	37
FP	23	147	137	137
TN	151	34	37	37
nodes	145	107	39	25

Table 4.4: Table showing the decision tree cross-validation accuracy and performance on the test set for each parameter tuning iteration

4.4 Visualizing the Decision Tree

The different tuning techniques in section 4.3 are performed with a purpose to simplify the tree, thus easing the interpretation of the rules of classification and the associations between the features of a top employer. The pruning that is performed decreased the complexity of the trees, as shown in Figure 4.6 to Figure 4.9 with the last pruned tree in Figure 4.9 representing a structure with the fewer nodes. However, it is important to note that all four trees have shown to grow with a similar upper section i.e. identical nodes and splitting rules. This leads to concluding that pruning the tree has not modified the rules of classification of the tree but has removed additional nodes that were most probably overfitting the model.

Visualization of decision trees at each step of the tuning process

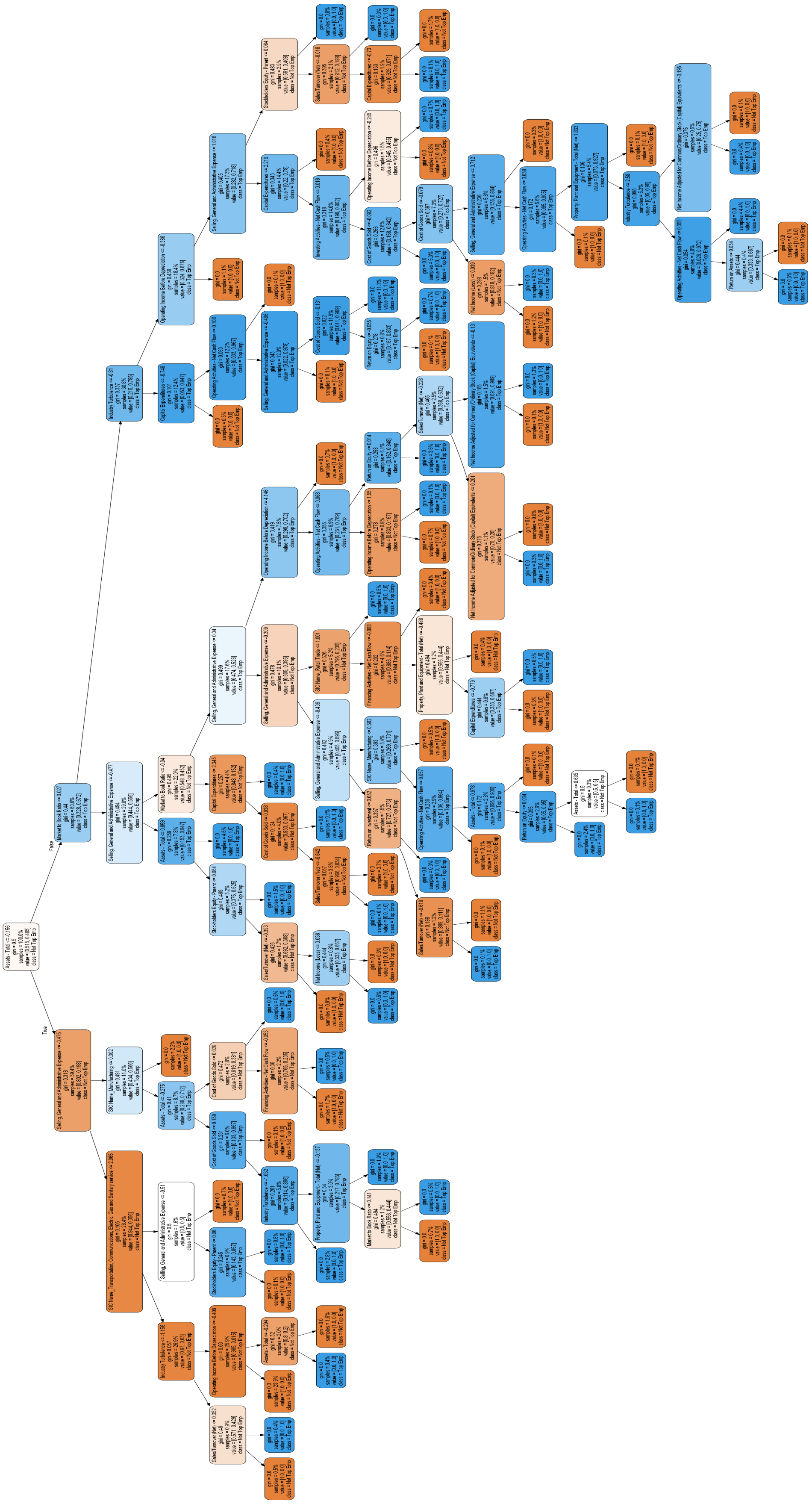

Figure 4.6: Visualization of decision tree before applying any tuning

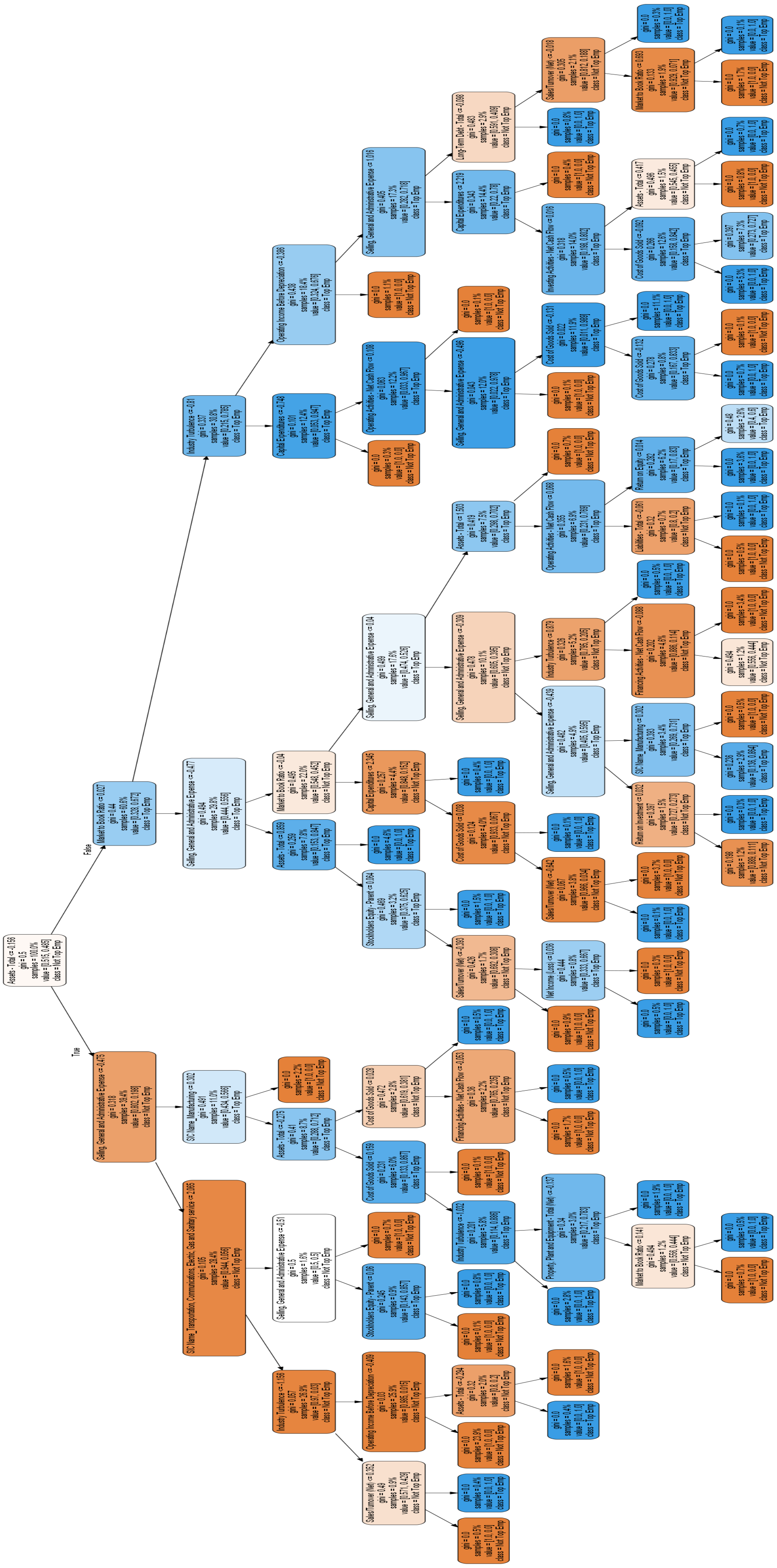

Figure 4.7: Visualization of decision tree after tuning using grid-search with large set of parameters' values

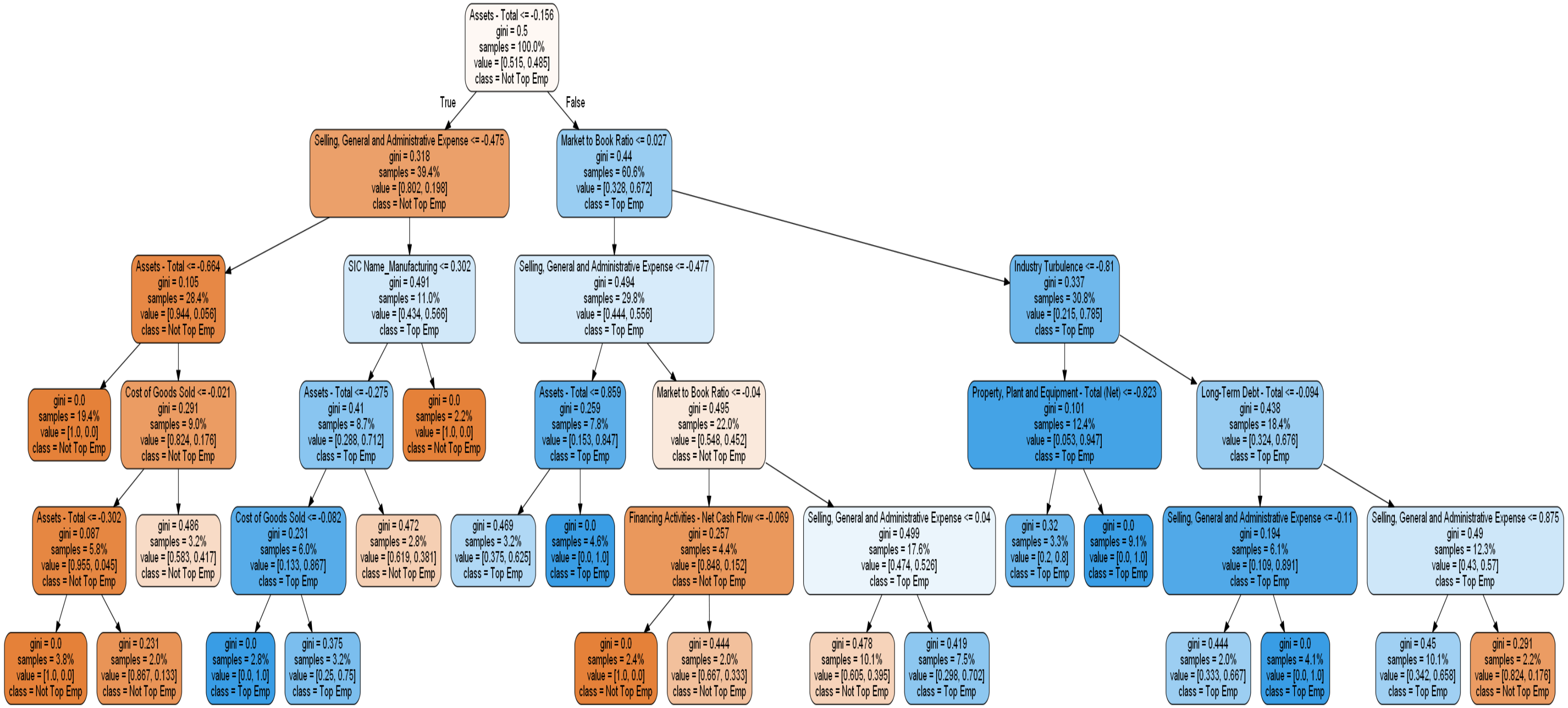

Figure 4.8: Visualization of decision tree after tuning using grid-search. Tree is tuned using the parameters (minimum sample 11 and maximum depth 9)

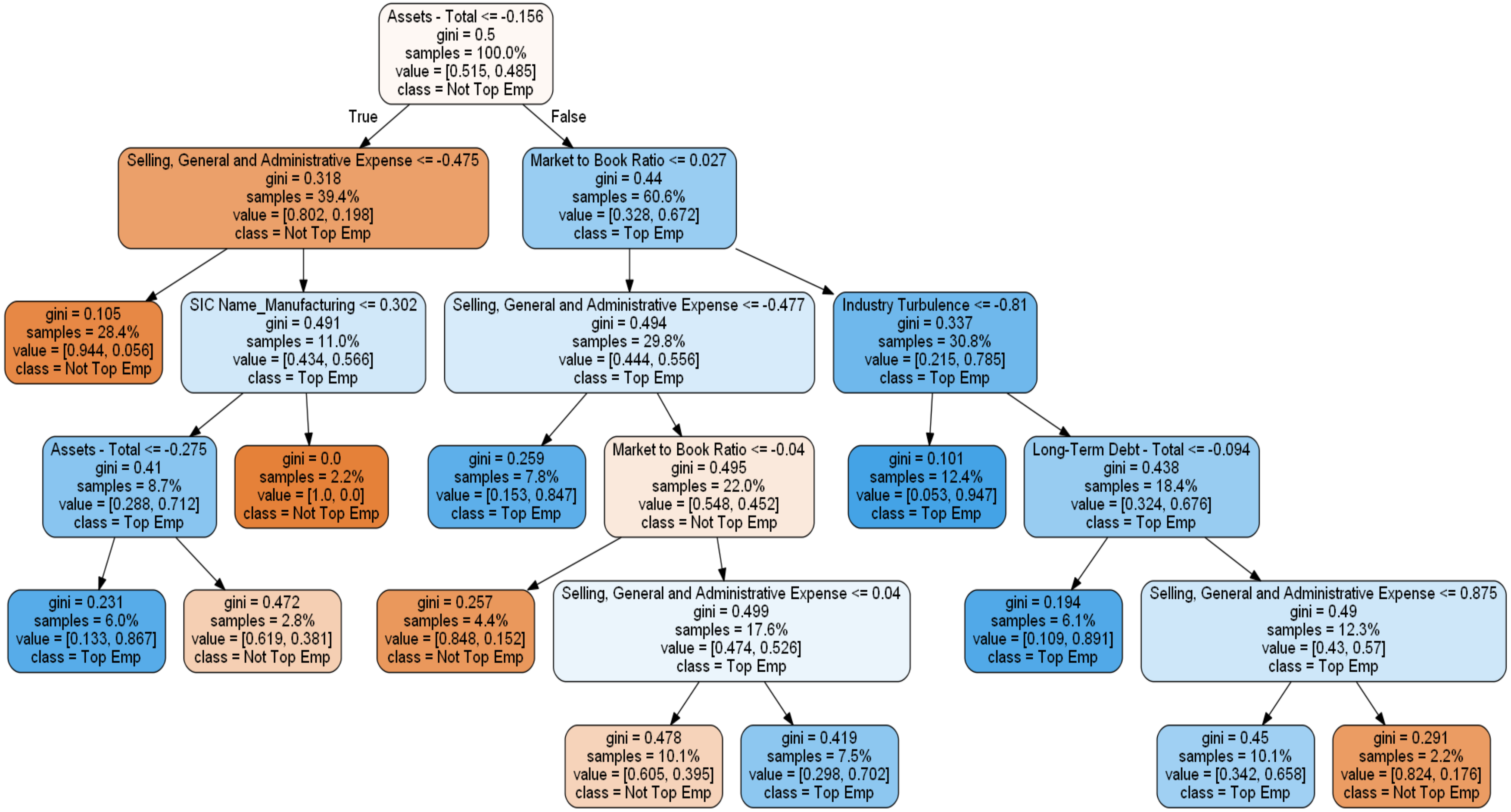

Figure 4.9: Visualization of decision tree after tuning using grid-search. Tree is tuned using the parameters alpha $= 0.005$ and (minimum sample $=15$ and maximum depth $= 5$)

For this reason, it is reliable to consider the upper nodes of the tree as represented in Figure 4.10, because the nodes are consistent throughout the tuning process. Furthermore, these nodes identify that the most important features in the top employer classification are the assets, the market to book ratio, the selling, general and administrative expenses, and whether the firm belongs to the manufacturing industry or not.

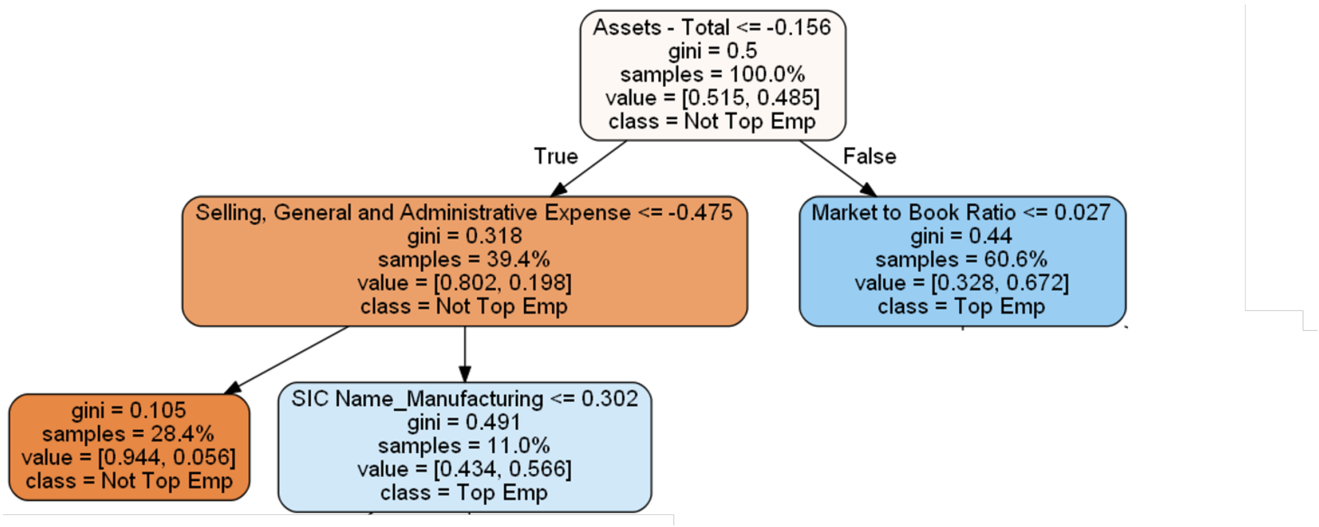

Figure 4.10: A selection of the upper nodes that are common between all the decision trees

The Rules and the Nodes of the Decision Tree

The first node of the tree splits on the total assets leading to a node with 67% of the sample classified as a top employer for firms with assets higher than 7.5 million$. In contrast, total assets lower than 7.5 million$ leads to a not top employer node with 80% of the node sample classified as not top employer. Furthermore, the next split is done on the selling, general, and administrative expense node as per the amount of 156.73 million$. Having a lower amount leads to a leaf with 94% classified as not top employers, and a higher amount leads to a node of top employers with 57% of the node sample classified as top employers. In-depth analysis of these rules are presented in chapter 5.

4.4.1 Decision Trees Per Industry

After building the non-industry specific tree, a more in-depth investigation is done to identify whether the features of a top employer differ from one industry to another. This is achieved by applying the decision tree on observations stratified by industry. For this purpose and to generalize the results per industry, a fair number of observations need to be available for each industry. Thus, the manufacturing industry is chosen for having the most significant number of firm-year observations in the data-set with 384 observations of manufacturing firms.

Building the Manufacturing Decision Tree

Examining the number of firms between the two classes suggests that they are not equally represented with 247 not top employers and 137 top employers. For this reason, the sample is downsized to 274 manufacturing observations to ensure an equal number of observations per class, and thus a balanced sample. Since this sample is relatively small, splitting the sample in test and train samples is not performed; instead, all samples are used to generate the model, suggesting there is overfitting of the values. The overfitting is tolerable in such a scenario because the purpose is not to build an optimal model but to have an idea of the important features in the manufacturing tree.

Visualizing the Manufacturing Decision Tree

The Manufacturing decision tree structure is displayed in Figure 4.11 showing an apparent distinction with the generic tree in Figure 4.6 to 4.10, with similarity in some features. Unlike the overall industries' tree, the manufacturing decision tree identifies the operating income before depreciation as the most important feature for their top employer's classification.

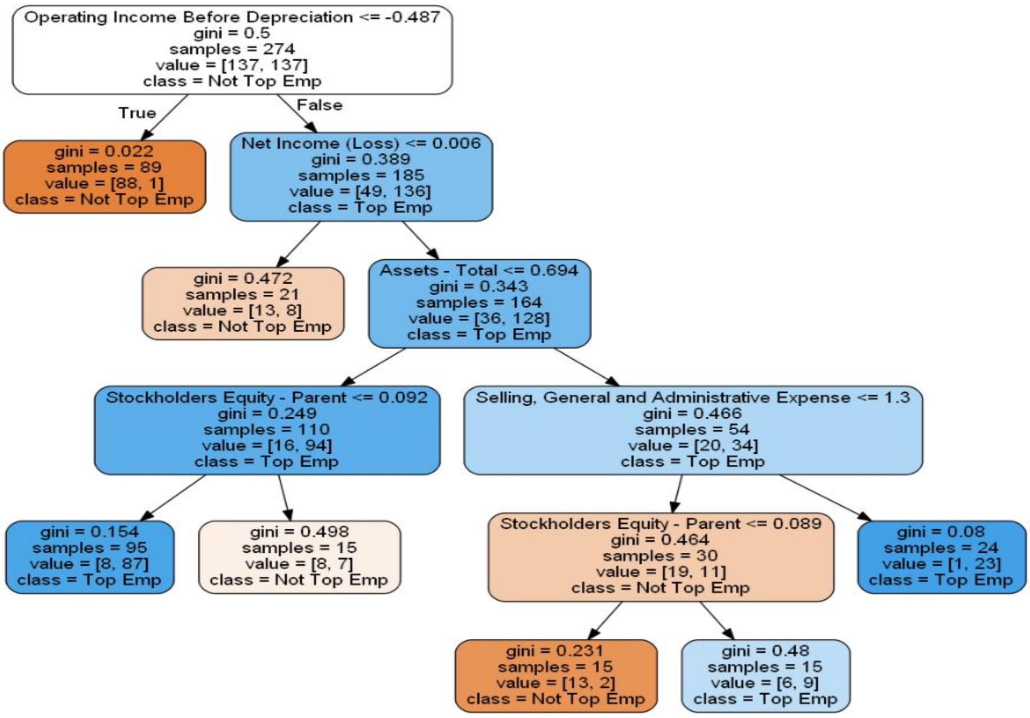

Figure 4.11: Visual of decision tree structure applied for observations that belong to the manufacturing industry

This decision tree is built using the generic tree's optimal parameters. Additional iterations are done to examine the output of the tree with different parameter values, using grid search, and also, using a train and test sets instead of one sample as test and train. The purpose of these iterations is to examine how the tree's features and rules evolve with each iteration while maintaining a good classification performance. The performance of each tree is displayed in Table 4.5, showing the accuracy CV is above 75% for all iterations.

	Parameters as the Generic Tree with data split on train and test sets	Parameters as the Generic Tree with train and test on same set	Parameters from Grid Search with data split on train and test sets	Parameters from Grid Search with train and test on same set
Max Depth	5	5	2	7
Min Sample	15	15	1	8
Accuracy CV	77%	75%	84%	82%
Accuracy on Test	84%	NA	82%	NA
Recall	100%	87%	100%	96%
Specificity	70%	89%	65%	91%
F1-Score	86%	88%	84%	94%
TP	40	119	40	131
FN	0	18	0	6
FP	13	15	15	12
TN	30	122	28	125
# Nodes	13	17	7	29

Table 4.5: Table showing the manufacturing decision tree cross-validation accuracy and performance using the generic tree parameter values or using grid search parameter values and compared when using one set or two sets to train and test the model

Similar to the generic tree, the important features remain present throughout the different tree iterations, however, the importance of these features, as well as the associated rules, change throughout these trees except for one feature, which is the operating income before depreciation.

Unsurprisingly, this behavior is expected since the tree is built using a small sample, which does not provide a robust tree structure. However, it does highlight a feature as the most important in driving the reputation of a manufacturing firm. As such, the feature which is stable with the same importance across the different manufacturing trees is the operating income before depreciation, and hence, is confirmed as a driver for the reputation of manufacturing firms.

Chapter 5

Analysis & Recommendations

The first finding in Chapter 4 recognizes the built decision tree model as a classifier to identify whether a firm is a top employer or not top employer using the firm's Compustat data. Nowadays, firms await Fortune to publish the yearly top employer's list to determine if they are selected as a top employer. This list represents a two years lag from the date the firm's data is reported and evaluated until the list is published. So if the firm does not make it on the list, it would have already lost two years where potential investments could have been made to improve its reputation. This study offers a new tool, a model that any firm can use to identify whether they represent a top employer or not. As such, this classification can be performed at any point in time solely by processing the firm's Compustat data. If the firm is not classified as a top employer, improvements can be taken to increase the firm's chance to become a top employer and therefore benefit from the fact that such reputation has a positive impact on its financial performance.

5.1 Analysis & Recommendations : Non-Industry Specific

Features of a Top Employer

The second finding of the study highlights the important features in the model's classification: The most important features in the classification of a firm's reputation are as below:

- Total assets: This item represents current assets plus net property, plant, and equipment plus other non-current assets (including intangible assets, deferred charges, and investments and advances) (Standard&Poor's, 2003). Table 5.1 reports that top employers have a minimum assets of 158 million$

with a maximum of 1,119,796 million $, whereas not top employers have a minimum assets of 0.004 million $ with a maximum of 346,196 million$.

- Market to book ratio: This item represents the proportion of a firm's asset base that is in intangible form and to capture the market's expectations of anticipated future economic returns (Roberts and Dowling, 2002; Fulmer et al., 2003).Table 5.1 reports that top employers have market to book ratio values between -5,354 and 3,021, whereas not top employers have values between -94,288 and 4,978.
- Selling, general and administrative expenses: This item represents all commercial expenses of operation (such as, expenses not directly related to product production) incurred in the regular course of business pertaining to the securing of operating income (Standard&Poor's, 2003). Table 5.1 reports that top employers have a minimum selling, general and administrative expenses of 0 million$ and a maximum of 28,237 million $, whereas not top employers have a minimum of 0 million$ and a maximum of 32,576 million$.

		Not Top Emp	Top Emp	All Firms
Assets - Total	Min	0.004	158	0.004
	Median	1,307	9,071	4,294
	Avg	13,180	51,472	31,511
	Max	346,196	1,119,796	1,119,796
Market to Book Ratio	Min	-942.88	-53.54	-942.88
	Median	2.36	3.77	3.05
	Avg	0.36	5.01	2.59
	Max	49.78	30.21	49.78
Selling, General & Admin Expense	Min	0	0	0
	Median	141	1,244	337
	Avg	1,329	3,217	2,233
	Max	32,576	28,237	32,576

Table 5.1: Table showing the minimum, maximum, median and average of the top features in the classification

More specifically, the focus is to identify the drivers of a top employer and the features that characterize it. By investigating the common upper tree section as displayed in Figure 4.10, two features seems to be involved in the classification of a top employer, the total assets and selling, general and administrative expenses.

An interpretation for the firm's total assets being a dominant predictor is that the assets of a company is an indication of the size of the company, suggesting that big firms have a better reputation than smaller ones (Mak and Kusnadi, 2005). Furthermore, the selling, general and administrative expenses represent the commercial expenses of operation that are not directly related to product production. Examples of these expenses include expenditure on marketing and advertising, directors' fees and remuneration, labor and related expenses, and company sponsored research and development (R&D) expenses, among others. High expenditure on labor reflects a firms' willingness to invest more in employees' benefits such as salaries and pension, retirement plans, bonus, and employee insurance. The benefits package is a fundamental criterion that an employee considers to stay working for a firm or to seek another opportunity with a better package. The more competitive the packages are, the more competent job-seekers seek to pursue a career in such a firm and less likely existing employees leave the firm. Employees in such firms are characterized as loyal and content, holding a positive perception toward their employer. As such, they are more likely to provide positive feedback about their firms (for example, responding positively to Fortune's top employer's survey). Furthermore, investing in R&D is an indicator of how innovative a firm is and how it pursues better quality and competence, which in turn influence the stakeholders' perception favorably. Moreover, Firms that allocate more spending in R&D encourages employees to apply new innovative solutions. This can reflect positively on employees' self-development and therefore well being and attitude towards the company.

Rules

The third finding shows that the different tree tested have a common set of rules in their upper tree section as displayed in Figure 4.10. This highlights the stability of these rules, among different trained models, and therefore suggest solid rules that can be identified. Out of these rules, two rules control a firm's classification as a top employer.

The first rule states that firms owning assets greater than 1810.9 million$ are classified as a top employer. To further investigate this rule, the distribution of assets for each type of employers is examined in Figure 5.1 and shows that around 75% of the number of top employer observations have assets greater than this threshold whereas more than 75% of the number of not top employer observations have assets below this threshold. This distribution aligns with the first classification rule, where the rule states that firms with assets greater than threshold are classified as top employers and reflected in the distribution where the value of the rule threshold is at the limit of the first quartile, indicating that most top employers have assets greater than this value.

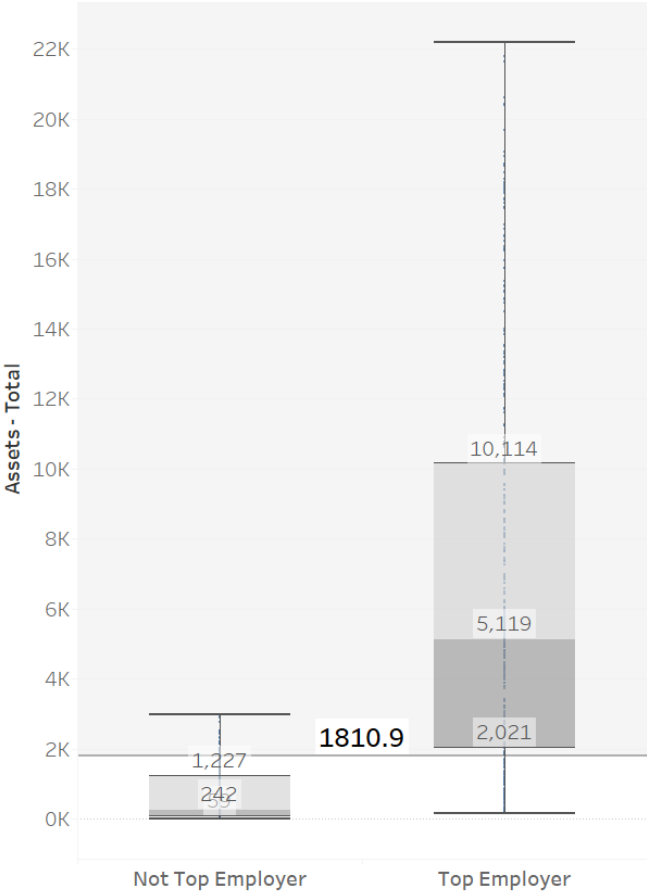

Figure 5.1: Distribution of the total assets feature by four quartiles and stratified by top employer and not top employer (outliers are excluded from the box-plot for visual clarity purpose)

The second rule states that firms with total assets lower than 1810.9 million$ and with selling, general and administrative expenses greater than 156.73 million$ are classified as top employers. To note that the selling, general and administrative expenses threshold value is normalized by the assets of each firm, thus the rule threshold is affected by the size of the firm. This means that firms with total assets lower than 1810.9 million$ can still be a top employer if they spend more than 156.73 million$ multiplied by the firm's assets, in the selling, general, and administrative area. Figure 5.2 examines the distribution of selling, general and administrative expenses for each type of employers and shows that around 75% of the number of top employer observations have selling, general and administrative expenses greater than the rule threshold, whereas more than 75% of the number of not top employer observations have values lower than this threshold.

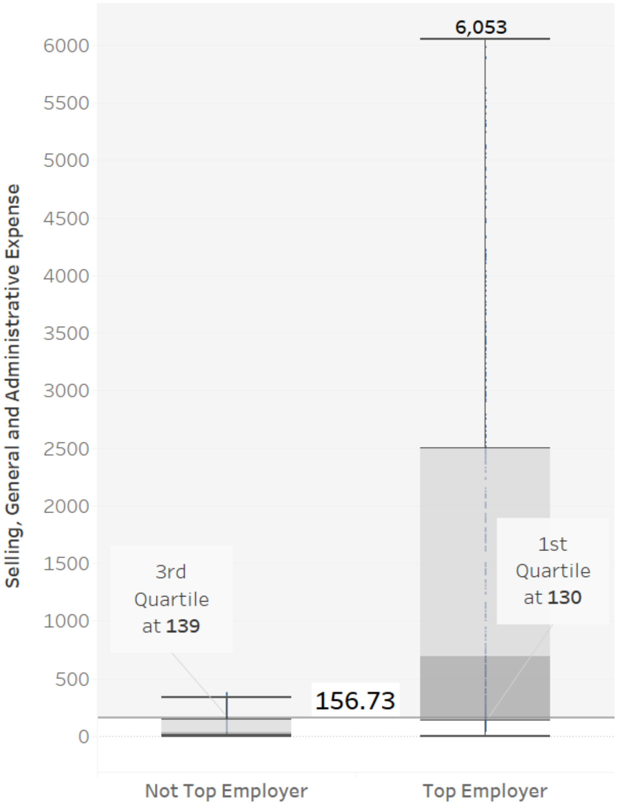

Figure 5.2: Distribution of the selling, general and administrative feature using four quartiles and stratified by top employer and not top employer (outliers are excluded from the box-plot for visual clarity purpose)

Recommendations

With this study's findings, a firm can increase its chance to become a top employer by investing in specific areas identified as drivers of its reputation. Although the assets are an important feature as indicated by the decision tree, it cannot be controlled by the firm, and it was included in the study to control the size of the firms. However, the selling, general and administrative expenses feature can be controlled by the firms by directing their spending. Values that constitute the breakdown of this feature are not accessible in this study, still, analysis is done based on the different components of this feature and how they can impact reputation. As such, this study recommends firms to increase spending on R&D and on employees' benefits, such as their salaries, retirement plans, and insurance packages, and identifies the threshold of these spending to be above 1810.9 and 156.73 million$ respectively. Adopting the model's recommendation ensures happier employees, which helps the firm become a top employer and improve its reputation, which in turn leads to better financial performance as well as more competent talents' hiring and retention. The model developed in this study can give a mathematical proof that such spending and investment does help companies achieve this reputation. Moreover, this study provides recommendations on how to direct the spending by specifying which features have more weight and what spending values will start making a difference in enhancing its reputation. Furthermore, the rule thresholds identified by the model can be further de-normalized to provide firm-specific spending values.

Firms from different industries and size can use this recommendation. Two examples of firms are further inspected in Figure 5.3, where Accenture is ranked as a top employer in Fortune for year 2017, while Astronics is not listed and considered as a not top employer. Accenture has assets and selling, general and administrative expenses greater than 1810 and 156 million$ respectively, so accordingly aligns with the top employer feature values. However, Astronics has assets and selling, general and administrative expenses lower than 1810million$ and 156 million$ multiplied by the company's assets respectively, so accordingly aligns with the findings that they are not top employers and should follow the recommendations and make investments in these areas to improve their reputation and become a top employer.

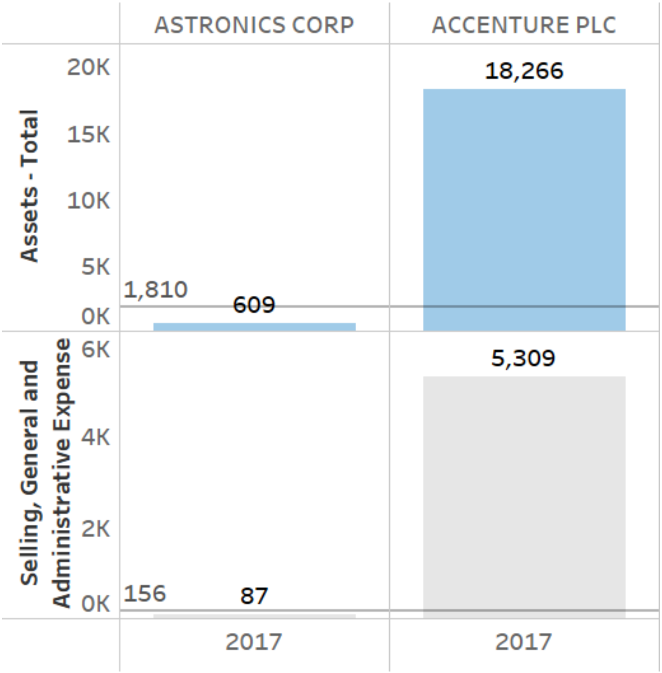

Figure 5.3: Bar Chart showing two firms examples, Accenture and Astronics identified in 2017 as a top employer and not top employer respectively with the assets and selling, general and administrative expenses compared to the rules' thresholds of 1810.9 and 156.73 million$ respectively.

Moreover, Figure 5.4 shows the yearly assets that the firm Acuity reported in Compustat between the years 2000 and 2020 compared to the rule threshold of assets equal to 1810. Over the years, Acuity shows an increase in assets to reach the rule threshold in 2013 and continues increasing, and two years later, it ranks as a top employer in 2015, 2016 and 2017.

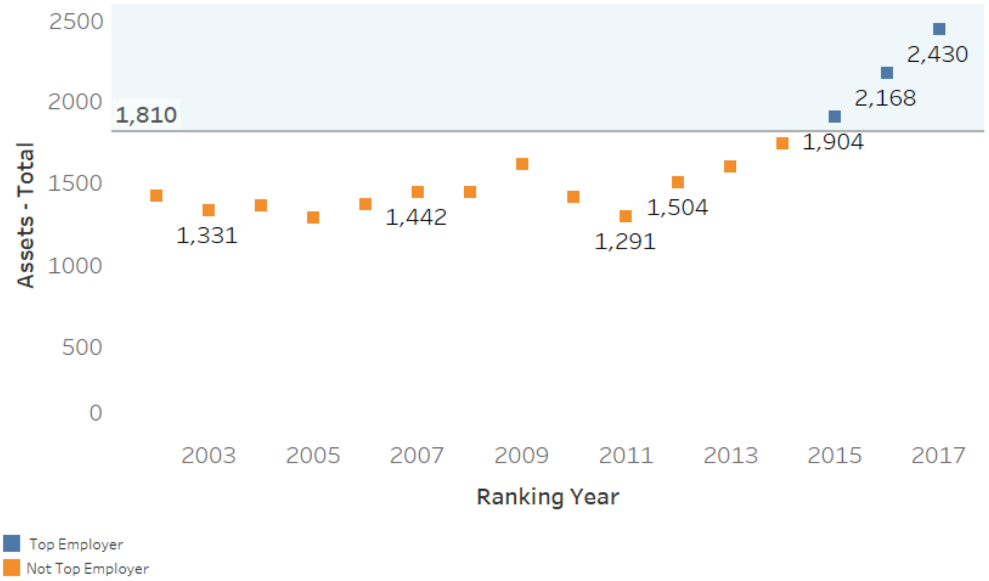

Figure 5.4: Chart showing the trend of assets per year for Acuity Brands, while highlighting the years 2015, 2016 and 2017 when it was a top employer

5.2 Analysis & Recommendations : Manufacturing Industry

In an attempt to focus on a certain industry scenario, the model is applied on manufacturing firms to analyze industry-specific reputation drivers and recommendations. By investigating the manufacturing tree as displayed in Figure 4.11, the main feature involved in the classification is the operating income before depreciation:

- Operating income before depreciation: This item represents Sales (Net) minus Cost of Goods Sold (Amortization of software costs, Motion picture and entertainment companies' amortization of film costs, Labor and related expenses reported above a gross profit figure, Rent and royalty expense, Taxes other than income taxes) and Selling, General, and Administrative expenses (Corporate expense, Parent company charges for administrative service, Research and development expense) before deducting Depreciation, Depletion and Amortization (Standard&Poor's, 2003). Table 5.2 reports that top employers have a minimum Operating income before depreciation of 112 million$ and a maximum of 23,996 million$, whereas not top employers have a minimum of -222 million$ and a maximum of 65,769 million$.

		Not Top Emp	Top Emp	All Firms
Operating Income	Min	-222	112	-222
Before Depreciation	Median	129	2,604	616
	Avg	3,221	4,761	3,771
	Max	65,769	23,996	65,769

Table 5.2: Table showing the minimum, maximum, median and average of the top feature in the classification of Manufacturing firms

This feature suggests that the manufacturing firm's operating income affects whether the firm is classified as a top employer or not. Furthermore, the main rule to classify a top employer states that firms with operating income before depreciation greater than 281.53 million$ multiplied by the firm's assets are classified as a top employer.

Comparing the manufacturing and the non-industry specific trees

Table 5.3 reports that the manufacturing and the non-industry specific tree have different top features, with the assets, the market to book ratio and the selling, general and administrative expenses identified as classifiers for the non-industry specific tree, while the manufacturing decision tree has the operating income before depreciation as a top feature. This suggests that manufacturing firms have particular drivers for their reputation different from those identified by the non-industry specific tree.

Tree	**Feature**
Non-Industry Specific Tree	Assets Market to Book Ratio Selling, General and Administrative Expense
Manufacturing Decision Tree	Operating Income Before Depreciation

Table 5.3: Table showing the features selected for the non-industry specific tree and the manufacturing decision tree, from running multiple iterations with different parameters

Despite that the industry and the non-industry specific trees have different structure, further examining Figure 4.11 shows that the assets and the selling, general and administrative expenses (non-industry specific features) are also represented in the manufacturing tree as nodes in the lower tree structure. This suggests that the manufacturing tree has features identical to the generic tree's features, in addition to its own specific features. Thus, manufacturing firms' reputation is driven by non-industry-specific indicators as well as industry-specific ones.

This study could not focus on firms from other than the manufacturing industry, because the number of observations stratified by industry is low. Moreover, the relatively low number of observations that belong to the manufacturing firms prevents this study from having a conclusive stable tree for the manufacturing firms. Furthermore, industry-specific analysis shows that rules and important features could change, and thus it is recommended to carry an extension to this work to capture a more significant number of samples that are industry-specific, and apply the same methodology to come up with solid recommendations.

Chapter 6

Conclusion

This study is perhaps one of the first attempts to investigate the drivers of corporate reputation using machine learning classification models. Our work builds on research that hypothesis a positive relationship between corporate reputation and the firm's financial performance. Surprisingly, published empirical studies investigating the drivers of a good corporate reputation are sparse. This research is the first to examine a firm's financial and operational data to identify what characterizes a top employer in an attempt to build general and industry-specific recommendations toward becoming a top employer.

Using top employers from Fortune and data from Compustat, we find that a decision tree classifier can be used to classify a top employer using the firm's financial and operational data. Moreover, we find that spending on employees' benefits and R&D helps to improve a firm's reputation. The practical implication of these findings is that firms can use our model at any point in time to evaluate their reputation status. As such, a firm classified by the model as not top employer can take advantage of the early detection, and make changes to enhance its reputation without the need to wait for Fortune's list. Furthermore, realizing the reputation implication of investing in R&D and labor expenses, our recommendation is beneficial to firms with different size and industries that are interested in enhancing their reputation. As such, the recommendation provided in this study can be adopted by top employers to ensure preserving their status or by not top employers to help improve their reputation and become a top employer.

Additionally, future work can continue to develop our understanding of the features of a firm's reputation. This study only investigates same year ranking-Compustat data as a driver for the reputation. A firm's historical data can also play a role in defining its reputation. Moreover, with the limited industry data that we have, industry-specific recommendations are not possible for every industry because of the lack of top employers in such industries, and the availability of only public firms' data in Compustat.

Overall, given the importance of corporate reputation and the proven financial performance returns, more research on the drivers of reputation is needed. By providing this new framework to firms from all industries, using Compustat data only, firms can evaluate their reputation as a top employer or not, and make the best investments to enhance their reputation by following the provided recommendations. This study helps set the stage for more and more robust research in this field.

Appendix A

Abbreviations

Top Emp	Top Employer
Not Top Emp	Not Top Employer
MTB	Market to Book Ratio
ROE	Return On Equity
ROA	Return On Assets
ROI	Return On Income
TN	True Negative
FP	False Positive
FN	False Negative
TP	True Positive
CV	Cross-validation
BS	Balance Sheet
IS	Income Statement
CF	Cash Flows

Bibliography

Aziz, M. A. and Dar, H. A. (2006). Predicting corporate bankruptcy: where we stand? *Corporate Governance: The international journal of business in society*. Type: Journal Article.

Ballou, B., Godwin, N. H., and Shortridge, R. T. (2003). Firm value and employee attitudes on workplace quality. *Accounting Horizons*, 17(4):329–341. Type: Journal Article.

Barnett, M. L., Jermier, J. M., and Lafferty, B. A. (2006). Corporate reputation: The definitional landscape. *Corporate reputation review*, 9(1):26–38.

Blazovich, J. L., Smith, K. T., and Smith, M. (2013). Employee-friendly companies and work-life balance: is there an impact on financial performance and risk level? *Journal of Organizational Culture, Communications and Conflict, Forthcoming*. Type: Journal Article.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32. Publisher: Springer.

Capkun, V., Hameri, A., and Weiss, L. A. (2009). On the relationship between inventory and financial performance in manufacturing companies. *International Journal of Operations & Production Management*. Type: Journal Article.

Dierickx, I. and Cool, K. (1989). Asset stock accumulation and sustainability of competitive advantage. *Management science*, 35(12):1504–1511. Type: Journal Article.

Fombrun, C. (1996). J.(1996)'Reputation: Realizing Value from the Corporate Image'. *Harvard Business School Press, Boston, MA*. Type: Journal Article.

Fombrun, C. and Shanley, M. (1990). What's in a name? Reputation building and corporate strategy. *Academy of management Journal*, 33(2):233–258. Type: Journal Article.

Fortune (2020). 100 Best Companies to Work For.

Fulmer, I. S., Gerhart, B., and Scott, K. S. (2003). Are the 100 best better? An empirical investigation of the relationship between being a “great place to work” and firm performance. *Personnel Psychology*, 56(4):965–993. Type: Journal Article.

Gatewood, R. D., Gowan, M. A., and Lautenschlager, G. J. (1993). Corporate image, recruitment image and initial job choice decisions. *Academy of Management journal*, 36(2):414–427. Type: Journal Article.

Grant, R. M. (1991). The resource-based theory of competitive advantage: implications for strategy formulation. *California management review*, 33(3):114–135. Type: Journal Article.

Greening, D. W. and Turban, D. B. (2000). Corporate Social Performance As a Competitive Advantage in Attracting a Quality Workforce. *Business & Society*, 39(3):254–280. Type: Journal Article.

Hall, R. (1992). The strategic analysis of intangible resources. *Strategic management journal*, 13(2):135–144. ISBN: 0143-2095 Publisher: Wiley Online Library.

Hall, R. (1993). A framework linking intangible resources and capabiliites to sustainable competitive advantage. *Strategic management journal*, 14(8):607–618. Type: Journal Article.

Hammond, S. A. and Slocum, J. W. (1996). The impact of prior firm financial performance on subsequent corporate reputation. *Journal of Business Ethics*, 15(2):159–165. Type: Journal Article.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied logistic regression*. Wiley New York.

Huatuco, L. D. H., Montoya-Torres, J. R., Shaw, N., Calinescu, A., Wang, Z., and Sarkis, J. (2013). Investigating the relationship of sustainable supply chain management with corporate financial performance. *International Journal of Productivity and Performance Management*. Type: Journal Article.

Lado, A. A. and Wilson, M. C. (1994). Human resource systems and sustained competitive advantage: A competency-based perspective. *Academy of management review*, 19(4):699–727. Type: Journal Article.

Lam, M. (2004). Neural network techniques for financial performance prediction: integrating fundamental and technical analysis. *Decision support systems*, 37(4):567–581. Type: Journal Article.

Langley, P. (1996). *Elements of Machine Learning*. Morgan Kaufmann. Google-Books-ID: TNg5qVoqRtUC.

Lau, R. and May, B. E. (1998). A win-win paradigm for quality of work life and business performance. *Human Resource Development Quarterly*, 9(3):211–226. Type: Journal Article.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23. Publisher: Wiley Online Library.

Love, L. and Singh, P. (2011). Workplace Branding: Leveraging Human Resources Management Practices for Competitive Advantage Through "Best Employer" Surveys. *Journal of Business and Psychology*, 26:175–181. Type: Journal Article.

Mak, Y. T. and Kusnadi, Y. (2005). Size really matters: Further evidence on the negative relationship between board size and firm value. *Pacific-Basin Finance Journal*, 13(3):301–318.

Mantovani, R. G., Horváth, T., Cerri, R., Junior, S. B., Vanschoren, J., de Carvalho, A. C. P. d., and Ferreira, L. (2018). An empirical study on hyperparameter tuning of decision trees. *arXiv preprint arXiv:1812.02207*.

Patel, J., Shah, S., Thakkar, P., and Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert systems with applications*, 42(1):259–268. Type: Journal Article.

Ponzi, L. J., Fombrun, C. J., and Gardberg, N. A. (2011). RepTrak$^{\text{TM}}$ pulse: Conceptualizing and validating a short-form measure of corporate reputation. *Corporate Reputation Review*, 14(1):15–35.

Roberts, P. W. and Dowling, G. R. (2002). Corporate reputation and sustained superior financial performance. *Strategic management journal*, 23(12):1077–1093. Type: Journal Article.

Shkolnikov, R. (2004). The business case for corporate citizenship.

Shrum, W. and Wuthnow, R. (1988). Reputational Status of Organizations in Technical Systems. *American Journal of Sociology*, 93(4):882–912. Publisher: University of Chicago Press.

Sridhar, S., Narayanan, S., and Srinivasan, R. (2014). Dynamic relationships among R&D, advertising, inventory and firm performance. *Journal of the Academy of Marketing Science*, 42(3):277–290. Type: Journal Article.

Standard&Poor's (2003). *Standard & Poor's Compustat® User's Guide*.

Stuebs, M. and Sun, L. (2010). Business Reputation and Labor Efficiency, Productivity, and Cost. *Journal of Business Ethics*, 96:265–283. Type: Journal Article.

Sun, L. (2012). Further evidence on the association between corporate social responsibility and financial performance. *International journal of law and management*. Type: Journal Article.

Syed, N. S. H., Ismail, S., and Yap, B. W. (2019). Personal bankruptcy prediction using decision tree model. *Journal of Economics, Finance and Administrative Science*, 24(47):157–170. Publisher: Emerald Publishing Limited.

Wang, K. and Smith, M. (2008). Does corporate reputation translate into higher market value?

Weigelt, K. and Camerer, C. (1988). Reputation and corporate strategy: A review of recent theory and applications. *Strategic management journal*, 9(5):443–454. Type: Journal Article.

Weiss, A. M., Anderson, E., and Maclnnis, D. J. (1999). Reputation management as a motivation for sales structure decisions. *Journal of Marketing*, 63(4):74–89. Publisher: SAGE Publications Sage CA: Los Angeles, CA.