ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Business & Economics: Data Science & Marketing Analytics

The Costs of Gender-Based Discrimination on Dutch Public Firms

Job Lagewaard

476702jl

Supervisor: Prof. Dr. A.C.D. Donkers

Second assessor: Prof. Dr. P.J.F. Groenen

25-10-2022

Abstract

This research gains insights on the cost of gender-based discrimination. Firms that build their workforce with a gender bias are not fully utilizing their potential because half of the national workforce is treated as a less valuable recourse, which could make the firm less efficient. To quantify this loss in efficiency, multiple categorizing XGBoost models are developed with the help of the 'Anonymous Application Procedure'. The outputs of these models are used to quantify gender biases using a 'Gender Bias Ratio' (GBR). This ratio finds evidence for the existence of gender-based discrimination for most Dutch public firms. The loss in efficiency is calculated using the GBR and the stock performances of each firm in a Kendall's ranked correlation test. This test suggests there are no significant costs arising from gender-based discrimination in a firm.

Table of Contents

Introduction

For Dutch LinkedIn members, the year 2022 started with many women changing their username to 'Peter' as a statement against the lack of diversity in the top management of Dutch companies (thedrum.com, 2022). This name was not chosen at random. The lacking diversity in Dutch publicly traded firms is characterized by the fact that there are more CEOs named Peter (five) than there are female CEOs (four). These four CEOs, or rather the missing 24 female CEOs which would be needed to meet the minimum requirements set by the Dutch government, are a mere symptom of a wider problem in the corporate ladders of Dutch firms (Senden, 2018).

The small number of female CEOs can partially be explained by a lower supply of proper candidates in the top layers of companies. This lack of proper candidates in the top management starts at the bottom of a company, as women are less likely to be invited for job interviews, less likely to be hired and earn less if they are equally or even better qualified then a male competitor (Åslund et al, 2012) (Carnevale, 2018).

To investigate if existing policies to reduce discrimination worked, as well as to derive to what extent women have a lower chance of being hired, Åslund et al. (2012) piloted the 'Anonymous Application Procedure' (AAP) in Sweden. In this procedure, hiring firms could not see the gender, name, or ethnicity of an applicant. After analyzing more than 3,500 applications they found female applicants indeed have a lower chance of being interviewed for a job under the traditional hiring processes. However, under AAP this gender gap completely disappears (Åslund et al, 2012). Further, the importance of measures negatively correlated with being a woman, such as working experience and years of education, increased a lot in the hiring decisions made under AAP. This could suggest managers are using these measures to discriminate on gender but given the fact the chances of being invited to the first interview round became equal for both women and men under AAP, it is expected that these measures are often underrepresented in the hiring process. These results suggest women are at a disadvantage in the hiring process (Åslund et al, 2012).

This discriminatory employee selection results in a workforce that undervalues certain perspectives and thus lacks the ability to gather and use all knowledge needed to have an optimal decision-making process (Hall & Davis, 2007). Such a decision-making process

ultimately creates a less efficient firm than a virtually identical firm that operates without gender-based (Bernile et al., 2018). The same paper states that being less efficient means a gender biased firm is likely to be underperforming and is wasting economic opportunities.

Literature

A lot of research has been done on the changing performance of a firm when a woman is employed as top manager. Chapple et al. (2014) find there is no relation between the diversity of a firm's board and its stock performance in most sectors (Chapple et al., 2014). Although a positive relation has not been found, there is evidence suggesting that diversity improves the informativeness of the stock prices, diverse boards are more transparent and have better corporate governance (Gul et al., 2014). Most researchers focus on the boards and CEOs of stock traded firms. This makes sense as it is a very tangible part of public firms, and the necessary data collection process can be done with relative ease.

These and many other researchers do not find a substantial difference in the performance of male and female lead firms. This lack of differences can be explained by two principles.

Firstly, CEOs and board members are chosen from a pool of experienced and well performing candidates. Publicly traded firms generally employ thousands of employees. This theoretically means the highest ranked candidates do not differ a lot when comparing their professional performance. This is because they all can be considered outliers within the firm. However, these employees are outliers in a skillset predominantly chosen by male managers, or managers working in a male dominated culture. Managers hire people with similar skills and cultures to their own (Rivera, 2012). Therefore, if a firm is male dominated, the men and women hired are chosen based on the skills and perceptions which male cultured managers deem important (Rivera, 2012). Thus, selecting only male employees from the best candidates should result in a team possessing an almost identical level of skills as a team with only female employees.

Secondly, the lack of differences can be explained by a lack of proper female candidates. As stated in the introduction, a firm needs to have a good hiring policy at the bottom of the organization that does not put women on the backfoot. This to have the best women in the top of the firm. Historically this has not been the case, creating a low supply of suitable women.

Extraordinarily little empirical research has been done on the relation of gender diversity and the general workforce's performance. One of the few extensive research projects on this matter investigated performances of investors and found that employing female investors is indeed positively related to an investing firm's performance (Barber & Odean, 2001). A further research paper named 'The Difference' proved that groups with a range of perspectives outperform a group of like-minded experts in any situation; ranging from workforces, school projects, democratic processes, or even scientific research (Page, 2007).

To gain more understanding on the financial effect of a gender-wise diverse workforce, this paper will be researching the monetary impact of gender biases in the workforce of publicly traded companies in the Netherlands. With this research, this paper takes a different and more innovative approach in measuring the effect of discrimination on the performance of a listed firm. This will help firms to gain insights into discrimination within their own firm and the market. This is of foremost importance as it will help firms to minimize economic and societal costs arising from missing information and perspectives. Also, finding insights on the quantification of the gender gap using machine learning can be used to create tools to optimize algorithms that help firms in the hiring process, as these algorithms are as biased as the people who created the input data over the years. The algorithms will therefore be putting women lower in their rankings without any rationale, besides biased inputs.

In the next section of this paper, the collection and use of the data will be explained. This is followed by the methodology with a presentation of the relevant models and measures. In the fourth section, the obtained results will be discussed and finally the conclusions will be presented.

Data

The data for this paper will mostly be obtained from LinkedIn.com. To investigate the differences between companies and the number of women they employ versus the number of men they employ, several thousand accounts in The Netherlands will be scraped using both an Rscript as well as an API in Python.

From all Dutch university faculties, alumni data can be scraped. By doing this it is ensured that not only the 1000 newest employees of a firm will be considered, which is the case when directly accessing data from firms. Another advantage of scraping by faculty rather than by firm lies in the fact that it creates a dataset where education directions can be easily obtained. When scraping alumni pages on LinkedIn, a firm specific selection can be made which ensures there are no profiles scraped from accounts that do not work for one of the listed firms.

Before scraping the profiles, it is important to define which firms can be included in this research.

Firm selection

It is important to create enough data points for each firm and to filter out listed firms that contain little useful information. These are SPACs as they have little to no business activities or firms that only registered in the Netherlands for tax reasons such as Universal Music Group. These firms all have exceedingly small (Dutch) employee counts, often up to a few hundred. Therefore, the minimum required number of employees is set to be 2000 employees based in the Netherlands.

In total, there are 75 firms publicly traded on the three major indices: AEX, AMX and ASCX. However, only 30 of those firms have 2000 or more employees in the Netherlands. This small number can be explained by the fact that many firms are registered in the Netherlands for taxation reasons although they do not actively hold office in the Netherlands. A brief overview of the included firms with their employee number and basic information is shown in table A1 in appendix A.

The 30 included firms employ close to 350,000 people in the Netherlands, with Ahold Delhaize's 100,911 employees as the main driver. Although one firm employs about 30% of all employees this is not expected to unbalance the data towards Ahold Delhaize as most of their employees have not attended university and only work in the supermarkets as a side

job, which often is not mentioned on LinkedIn. Ahold Delhaize does however show why it is important to only include Dutch employees as they employ 413,000 people worldwide which would make it extremely hard to create a balanced and representative dataset.

The employee counts are from 2021 and are obtained from DeGiro, in case DeGiro did not have recent employee numbers, annual reports were used (DeGiro.com, 2021).

In the introduction, the juxtaposition of female and peter-named CEOs was discussed, this can also be seen when looking at firms lead by women and Peters. Two of the included firms are led by a woman, totaling 24,279 employees, or 7.4% of the total included workforce. Two other firms have a CEO named Peter. These two firms account for 11.0% of the total workforce, or 36,114 employees.

Profile Selection

The initial profile selection is done by hand as using automated tools to search for people online is, as LinkedIn makes perfectly clear, illegal. This procedure simply consists of making a search inquiry on LinkedIn with all 30 firms in the 'current employer' filter after which the education filter allows to select people based on their faculty and educational background. The search inquiry is completed by adding the country filter to ensure only Dutch employees are considered.

This results in a list of LinkedIn members that fall within the criteria. As it is possible to preserve a log of visited webpages, the list of profiles is obtained by making sure each profile was shown on the screen i.e., by clicking on 'next' a few hundred times. This list was then saved into HAR files. A HAR file contains all information shown on all webpages visited during a specific session. To ensure that the results are not biased towards my own acquaintances, a new account has been created with only the minimum required number of connections to see other accounts (three). To further ensure a minimal selection bias, these connections are all newly created accounts with only each other as connections.

Using an Rscript, the person specific URLs are filtered from the HAR files and appended into a list consisting of a person's URL and educational direction. The educational direction is determined by the person's faculty. Five different directions are considered in this paper: Economics & Business (E&B), Technological (Tech), Biology and Chemistry (Bio/Chem), Medical (Medicine) and Law (Law). Table 1 shows the number of people per faculty spread

over thirteen universities included in the list of URLs. In total, after removing ~300 profiles that indicated to have studied at more than one faculty, 6640 profiles can be used as input for the API.

University	E & B	Tech	Bio/Chem	Medicine	Law	Total
Erasmus Uni. Rotterdam	687			24	182	**893**
Maastricht Uni.	265				11	**276**
Radboud Uni. Nijmegen	2	9				**11**
Rijksuni. Groningen	340	27		30	108	**505**
TU Delft		700				**700**
TU Eindhoven		263	262			**525**
Tilburg Uni.	300				72	**444**
Uni. Leiden		201	56	22	359	**638**
Uni. Twente		273				**273**
Uni. Utrecht	158	232	234	22	114	**760**
Uni. Van Amsterdam	400	274	145	24	200	**1,043**
VU Amsterdam			146	22	280	**448**
Wageningen Uni. & Res.			124			**124**
Total per direction	**2,152**	**2,051**	**967**	**144**	**1,326**	**6,640**

Table 1: Number of students per university and faculty in 2022. Any excluded University did not have a single relevant faculty or had overly broad faculties in which multiple educational directions were included.

Web scraping

LinkedIn is known for its aggressive stance against scrapers, this to ensure the privacy of its users.

This paper will employ an existing API to ensure all data is scraped in a legal and consistent manner. One of the few existing packages for scraping that is also capable of logging into LinkedIn without being immediately blocked is the Python package called “LinkedIn-API”. LinkedIn-API is capable of scraping all datapoints from a given LinkedIn user’s URL. The chosen data points only include ones that do not directly lead to a specific person and are presented on a public profile in order to be compliant with GDPR (GDPR, 2018)

With this API, in combination with an Rscript, it is possible to find all necessary information for profiles. The first step using the API is to assign some basic parameters, such as a username and password of the LinkedIn account to be used. The second step in the setup of the scraper is to set the search parameters. In the second step, the list of URLs is used to scrape all

datapoints from each profile. Each profile is then individually saved as a JSON file for later use, as is explained in the data cleaning section. To make sure that LinkedIn believes this script could be a human, a random number between zero and one is generated between scraping each profile. This number is multiplied by seven and used as a waiting time before going to the next profile. By doing this, the scraping process can continue a lot longer (up to 200 profiles per day) because LinkedIn is less likely to stop a program when it shows human-like behavior, such as not reading a complete profile within a fraction of a second.

Data cleaning and filtering variables

The use of this API generates a lot of data, which after cleaning and filtering the appropriate datapoints will be used in the model, as will be explained in the methodology section. Some profiles could not be scraped due to privacy settings. In total 5,282 profiles could be successfully scraped. This gives an equal number of JSON files which are all cleaned and appended into one dataset. This dataset consists of the following datapoints:

Gender: [categorical] As this research is about gender-based discrimination, the gender of the employees is needed. Firms and social media do not explicitly state the gender of employees. Because LinkedIn does not save genders, the first names of each LinkedIn member will be scraped and checked using the 'Nederlandse Voornamenbank' from the Meertens Instituut. Via a loop which creates the correct web address for each name and then scrapes the number of men and women with said name, the gender of a person can be determined. A name is categorized as female if more than 80% of its carriers is female, and vice versa.

Some genderless names such as Anne and Ezra will have to be filtered out, as well as the name 'Peter' because the recent protest made this name biased. All names with special characters are filtered out as these could not be parsed into a URL by R, as well as all names that are given to less than five people over the last 135 years because the Meertens Instituut does not show these for privacy related issues. This might raise a slight selection bias as it mostly affects foreign names, but it is not expected to be biased towards women or men. Foreign names might include information on performance differences caused by general diversity, but this paper only considers diversity through gender, making rare foreign names not useful for the direction of this paper if they are not related to gender. Any relevant selection bias can therefore be disregarded.

In total, 440 unisex, rare, Peter, and/or unreadable named profiles had to be removed. 1697 female as well as 3047 male profiles were kept. Giving a total of 4744 profiles.

Working Experience: [numerical] Experience is a crucial factor in the hiring process. Thus, working experience will be scraped. LinkedIn stores a starting and ending date for each job. Because current jobs do not have an end date, the 1st of July 2022 is used as an end date to calculate the experience of these people. An average employee in this dataset has been working for ~14.0 years. This variable also captures some differences between men and women; the average man has been working for 14.6 years, the average woman for only 12.2 years. To prevent outliers, the extremes of this variable are checked manually, resulting in not a single outlier.

It is possible for people to hold more than one position at the same time. In cases where multiple experiences overlap, the starting date of the second experience is changed to the end date of the first experience. By doing this, double counting of working experiences is prevented. Of all profiles, a principal scientist at Philips has gained the most experience, with a total of 50.49 years.

Educational direction: [categorical] The profiles will be scraped based on Dutch faculties. With this, the type of education someone followed can be used as a control variable. The inclusion of one's education makes it possible to account for any behavioral differences between men and women, as shown in table 2. E.g., almost 80% of all tech students are men, therefore a firm that only hires technical personnel cab be expected to hire more men than women.

Educational direction	Male (%)	Female (%)
E&B	64.6	35.4
Tech	78.9	21.1
Medicine	51.7	48.2
Law	38.7	61.3
BioChem	68.7	31.3

Table 2: Percentage of male and female employees with their respective educational direction.

Education Experience: [numerical] The number of years a person followed their study program will be scraped to account for any differences between men and women. This is important because there is a negative relation between being a woman and education, and because it is a key factor in the hiring process. The average man has 13.31 years of educational experience, the average woman's educational career is one year shorter, with a mean of 12.35 years. The reason behind this gap can be caused by many factors. For example, men may study longer due to their high presents in technical universities which more often have two-year master's degrees. On top of that, extracurricular activities might attract more men, or men need extra time as they enter universities with lower high school grades, although they leave them with higher grades on average than women (Francesconi and Parey, 2018). The exact rationale between the gap is beyond the scope of this research but it is likely that it will help the model to differentiate between men and women better.

While processing this variable, the educational experience of people who do not indicate any education on their LinkedIn profile are replaced by "NA's". When checking for outliers, 40 outliers were dropped from the dataset. These outliers came from people with 40 up to 91 years of educational experience. The cutoff of 40 years has been chosen because, without any delay, a Dutch person who finished a study in medicine has had 20 years of education. When this person obtains a PhD (five years) and one of the longer specializations (6 years), they study for 31 years. Given most people have some delays during their education, 40 years seems enough to finish. This has further been confirmed by the fact that there are hardly any observations above the cutoff.

Firm: [categorical] The firms will be scraped to investigate any differences in performance for firms that do employ women more, or less, often. Initially, all 30 selected firms will be included.

To categorize the firms, the data must be cleaned based on the numerous ways of spelling a firm's name. For example, each string containing 'boskalis', 'westminster' or any variation with capital(s), spaces and/or special characters is replaced by 'Boskalis'. The many ways of typing a firm's name are simply found by tabulating all unique names and sorting based on frequency of occurrence. In this process the possibility of a firm's name to occur within another firm's name has been considered, e.g., ING is the last firm to be filtered as its name

occurs in many strings of people working for a 'verzekeringsmaatschappij' (insurance company), such as ASR and NN. TKH group did not have a single observation and is therefore filtered out for any future model.

Another important thing to note is that LinkedIn does not save work experiences in chronological order. Therefore, all firms scraped from each profile were ordered based on the last day someone worked for the specific firm. This is to ensure the highest number of actual current employers are considered. LinkedIn offers the option to not show the latest employer on one's profile. However, these people still show up in search enquiries. Because of this, 121 did not contain the actual current firm. These observations are replaced by 'unknown'.

Experience at current firm: [numerical] It is important to take a person's experience with their current employer into account as a control variable. The participation rate of women in the overall working force has been increasing over the last decades, causing a negative relation between experience at the current firm and being a woman. However, this negative association only exists for older employees. For employees with less than 15 years of experience, there is a small difference in the years they have been working for their current firm; men have a mean of ~2.3 years while women have been working for their current employer for 2.1 years on average. For employees with more than 15 years of total experience, there is a gap of 0.8 years on average; these women work on average for 6.0 years at their current employer, while men have 6.8 years of experience.
Interestingly, the person with the highest working experience at the same firm is a woman, she has been working at Ahold Delhaize for more than 44 years.
In some rare cases, the experience at the current firm turned out to be negative. Fourteen negative outliers had to be removed because of this. These observations were negative as it is possible to add a starting date in the future on your LinkedIn profile. When doing this, there is a negative amount of time between the starting date and the end date, which is the 1st of July 2022 unless specified otherwise.

Number of firms: [numerical] Besides the number of years someone has worked, the number of different firms a person has worked for is included in the dataset. This is an important factor in the hiring process as it indicates someone's experience. This variable may therefore

be telling for a certain firm's workforce. To prevent multiple jobs at the same firm from being counted multiple times, this datapoint is measured as the number of unique firms one has worked for. Said variable ranges from zero to a staggering 36 unique firms for one of ING's board members. The profiles with the highest numbers of firms were checked to see if their numbers were not inflated due to small positions, e.g., advisory roles or board members. Only people with 'real' full time jobs were kept. This resulted in the removal of one outlier.

Highest obtained degree: [categorical] When working for a firm, having a degree is particularly important. It is possible to obtain a person's degree from LinkedIn data. For each person, the educational experiences are sorted based on their ending dates. After sorting, it is assumed that the last obtained degree is the highest degree a person has completed because most people are not very likely to e.g., follow a bachelor's degree after finishing a master's degree.

On LinkedIn, degrees are filled in as character strings. Therefore, hardly any profile has the same degree as someone else's. To make it suitable for the XGBoost models, it is necessary to subdivide the several thousands of different inputs into a few categories.

For this, five categories are chosen, and all data points are subdivided into these five categories based on several key words. The first category is 'lower education' and contains all studies below a university's bachelor's degree (e.g., 'high school', 'atheneum' and 'mbo'). This category does not contain many profiles because all data comes from people who state to be faculty alumni on their profile. A manual check shows people in this first category are people who dropped out of their study program or are currently following a study program; some part time in combination with their job and some full time. The second category contains all observations that include a keyword indicating a bachelor's degree (e.g., 'BSc', 'Bachelor' and 'WFT'). The third and fourth category consist of master's degrees (e.g., 'MSc', 'drs' and 'ir') and PhDs (e.g., 'dr', 'PDeng' and 'Ph.D.'). Lastly, some observations lacked clear information or were missing, these are categorized in a separate 'unknown' category.

Position: [categorical] When quantifying gender-based discrimination it is important to consider someone's position because firms that select on gender likely have less women in higher positions. The positions for each profile can be found on LinkedIn and need to be cleaned using several keywords. Every position is categorized in one of the eight following categories; a 'student' category containing all people who are still studying, secondly the 'graduate' category containing all positions with keywords such as 'junior' or 'traineeship'. The third 'medior' category is the biggest category and contains positions such as 'Data scientist' or 'analyst'. This is followed by the 'senior' category which contains amongst others 'sr. Data scientist' or 'senior'. The top two categories are 'management' and 'Top'. These two categories respectively contain positions such as 'manager' and 'eigenaar' (owner). Lastly, some positions could not be scraped or contained very vague descriptions[1]. These are combined into an 'unknown' category.

Number of Languages: [numerical] As some of the companies have a very international focus, the average number of languages their employees speak can be expected to be higher. The number of spoken languages will therefore be helpful in the classification models.

On average, people who reported their spoken languages speak ~3.0 languages, with more international firms such as ING having averages up to ~3.3. More Dutch focused firms tend to have averages around 2.7 spoken languages. People who have not stated their spoken languages are considered to speak one language. Interestingly, there seems to be no connection between men and women when it comes to the number of spoken languages.

Number of Certifications: [numerical] The chance of working in a certain position or a specific firm partially depends on the number of certifications a person has obtained. For this reason, the number of certifications reported on LinkedIn is included in the data. When checking for outliers, fourteen observations were dropped. These profiles contained between 21 and 45 certifications and tended to subdivide a single certification into many smaller ones. For example, these users were adding each level of a certain certification as a separate certification. When no certification is reported, the total number is set to zero. The number

[1] A problem that mainly recruiters seem to have.

of certifications is not a good indicator for gender; there is no relationship between gender and the reported number of certifications.

Number of Accomplishments: [numerical] In addition to the certifications, people can add their accomplishments on LinkedIn, which consist of prizes, grants, etc. After obtaining the number of unique accomplishments for each profile, some outliers between 33 and 65 reported accomplishments were deleted as these people reported accomplishments with extraordinarily little meaning, such as 'became third in high school run'. The number of accomplishments turns out to be a good indicator for genders. Men who report their accomplishments state an average of 6.36 accomplishments while reporting women have an average of 5.40 accomplishments. One might expect a difference in the number of men and women reporting their accomplishments. Only a 0.6% percentage point reporting different can be found in this dataset, suggesting men and women are almost equally likely to report their accomplishments. After checking the distributions for both genders, no significant difference besides the lower average for women can be found.

Volunteering: [Boolean] Some firms value volunteering experience, which makes it a good variable to add to the XGBoost models. It can also be considered a variable that correlates with gender. This can be derived from the fact that 16.6% of all men say they have experience volunteering, versus just over a quarter of all women (25.7%). The years and types are not considered in this research as it is near impossible to exactly determine without breaching LinkedIn's privacy standards. Therefore, this variable only states whether someone has done volunteering work in the past or not.

Exploratory analysis

After cleaning the data, 4,472 observations were left over all 29 firms, as shown in Table 4 below. A few interesting anomalies in the data can be seen in this table.

Smaller firms such as JDE Peets and SBM offshore have very few observations. Because this research uses a 5-fold XGBoost-model and therefore cannot cope with very small observation numbers for the firms, as it often selects zero employees for such firms. Firms with less than

50 observations are thus not included in the models. Ten firms are removed for this reason, resulting in a final dataset of 4216 employees over nineteen firms.

An interesting aspect of this data is the striking number of employees with a law background employed by ABN AMRO as compared to its competitors. This might be due to the increase in KYC-focus of banks in the last few years, which is further indicated by the other financial institutions who also state a relatively high number of personnel with a law background.

Firm	BioCh	E&B	Law	Med	Tech	NA	Male	Female	Total
ABN AMRO	75	253	283	4	44	14	55.27%	44.73%	**673**
ASML	96	89	7	1	324	11	80.87%	19.13%	**528**
ING	79	201	142	5	73	12	60.55%	39.45%	**512**
Unknown	95	173	93	5	139	7	65.23%	34.77%	**512**
Philips	147	125	9	15	78	8	62.04%	37.96%	**382**
Shell	51	82	21	2	75	3	69.66%	30.34%	**234**
NN	27	43	52	3	18	1	56.94%	43.06%	**144**
Unilever	24	95	13	0	8	1	43.97%	56.03%	**141**
PostNL	18	48	35	2	23	2	57.03%	42.97%	**128**
Adyen	10	75	12	0	20	2	66.39%	33.61%	**119**
Ahold Delh.	18	47	44	0	6	3	52.54%	47.46%	**118**
ASR	25	37	35	6	13	0	54.31%	45.69%	**116**
KLM	13	40	27	8	25	1	51.75%	48.25%	**114**
KPN	16	41	21	1	30	3	73.21%	26.79%	**112**
DSM	51	25	4	0	25	1	63.21%	36.79%	**106**
Heineken	8	53	12	0	24	0	71.13%	28.87%	**97**
AkzoNobel	39	17	9	1	7	1	67.57%	32.43%	**74**
Ordina	14	7	4	2	25	3	76.36%	23.64%	**55**
Boskalis	2	8	2	1	38	0	86.27%	13.73%	**51**
Total	**808**	**1459**	**825**	**56**	**995**	**73**	**63.91%**	**36.09%**	**4216**

Table 4: Number of employees per educational direction and percentage of women and men for each firm. Firms with less than 50 observations are left out, these are Aegon, Bam, Corbion, Heijmans, IMCD, InterTrust, JDE PEETS, Randstad, SBM offshore, Signify and Sligro.

Exploratory measures

Before building the models to investigate the research question, some variables are used to explore the dataset to gain initial insights and to get a better understanding of the data. These variables are:

Ratio male-to-female: [numerical] The number of male employees divided by female employees will be used as an exploratory measure for the diversification of each firm. A ratio of one is an -on paper- perfect diversification because there is a perfect split between genders. A ratio approaching zero indicates a lot of women are working for the company and

a remarkably high ratio indicates a lot of men are employed. An important note for this ratio is that the ratio itself gives a partial view of reality as it is not controlled for education, experience, and the other factors.

Ratio employed-to-available: [numerical] This is the percentage of female employees in each firm divided by the percentage of female employees available with the same educational background, as per table 5.

For example, if the hypothetical firm 'TechOnly' employs 200 people with a technical background, of which 150 are men and 50 women, the number of available women will be calculated by multiplying the number of employees with a technical background (200) by the percentage of female technical graduates (31.1%). This shows that TechOnly should have 62 women in its workforce given its workforce's educational background.

The ratio employed-to-available is simply the number of employed women (50) divided by the number of suitable women (62). This will be used as another exploratory measure of discrimination in the hiring process. The ratio indicates discrimination because a ratio approaching zero indicates a lot of women are not hired although they are available, and a ratio higher than one indicates a lot of women are hired as compared to the availability in the national workforce. Of course, there is more than only an employee's education when it comes to analyzing a firm's workforce so this ratio on its own is not enough to say something about discrimination, but it does give some initial insight into the dataset.

These exploratory variables are shown in table 5. As the firms with less than 50 observations hardly hold any statistical meaning for these ratios because they are not considered in any future model, they are omitted from the table. Overall, the Dutch publicly traded firms do not seem to discriminate based on gender when purely looking at the number of women with an applicable educational background they employ. For this table, the number of suitable women for each firm is calculated in the same manner as for the hypothetical 'TechOnly' firm above, however the various educational directions within each firm are considered as well.

A notable result is the Male-to-Female ratio of ASML. This firm employs 4.35 times more men than women. This can partially be explained by their high number of employees with a tech background. When controlling for this, ASML still hires 0.73 women for every woman

available, meaning its workforce has only 73% of the number of women that it should have when purely looking at educational backgrounds.

Interestingly, Unilever and Philips are an exception on the other side of the spectrum. Both firms hire more women than expected given their workforce's educational backgrounds. Philips is especially notable as they hire a relatively high number of technically trained employees, which often skews it a lot more to a lower ratio.

As a sidenote, the firms with the third and fifth highest employed-to-available ratios, DSM and PostNL, both have female CEOs. It would be interesting for future research to see if this holds any significant statistical meaning but given the sparse number of female-lead firms in this paper, it is hardly possible to investigate this properly.

Of course, all these ratios only use gender and educational background as measures, but they do give some initial insight into the gender-based division of the Dutch workforce. For a more elaborate analysis, the methods in the next section will be used.

Firm	Male	Female	Available women	Male/Female	Employed/ Available
ABN AMRO	55.27%	44.73%	45.01%	1.24	0.99
ASML	80.87%	19.13%	26.28%	4.23	0.73
ING	60.55%	39.45%	40.07%	1.53	0.98
Unknown	65.23%	34.77%	35.61%	1.88	0.98
Philips	62.04%	37.96%	32.04%	1.63	1.18
Shell	69.66%	30.34%	32.37%	2.30	0.94
NN	56.94%	43.06%	42.48%	1.32	1.01
Unilever	43.97%	56.03%	36.27%	0.78	1.54
PostNL	57.03%	42.97%	39.56%	1.33	1.09
Adyen	66.39%	33.61%	35.27%	1.98	0.95
Ahold Delh	52.54%	47.46%	43.75%	1.11	1.08
ASR	54.31%	45.69%	41.39%	1.19	1.10
KLM	51.75%	48.25%	38.84%	1.07	1.24
KPN	73.21%	26.79%	35.99%	2.73	0.74
DSM	63.21%	36.79%	31.05%	1.72	1.19
Heineken	71.13%	28.87%	34.72%	2.46	0.83
AkzoNobel	67.57%	32.43%	35.23%	2.08	0.92
Ordina	76.36%	23.64%	30.30%	3.23	0.78
Boskalis	86.27%	13.73%	25.85%	6.29	0.53
Total	**63.91%**	**36.09%**	**35.90%**	**2.11**	**1 (By default)**

Table 5: The percentage of women and men working for each firm, together with the ratio 'Male-to-female' (Male/Female), and the ratio 'employed-to-available' (Employed/Available).

Methodology

This research will investigate if firms that employ a more diverse workforce are more successful, while controlling for experience, education, and many other variables. To do this, it is vitally important to quantify the gender-based discrimination in the workforce of a specific firm. For this, a method comparable to the Anonymous Application Process (AAP) by Åslund et al. (2012) can be used.

The use of a random forest model will make it possible to simultaneously predict whether a specific person works for a certain firm, which features are important when making that prediction, and to quantify gender-based discrimination.

Quantifying gender-based discrimination using AAP

This paper uses an XGBoost model to find discrimination in the workforce, as will be elaborately discussed in the next paragraph. The XGBoost models will classify each observation into one of the nineteen firms. This model will output, among other things, the sensitivity of each firm. The sensitivity of a firm shows how many of the categorized people do indeed work for the specified firm, as compared to the total number of employees for the firm. More formally, formula 1 shows the calculation of a firm's sensitivity.

$$Sensitivity = \frac{CategorizedEmployees_{firm}}{CategorizedEmployees_{firm} + MiscategorizedEmployees_{firm}} \quad (1)$$

Where $CategorizedEmployees_{firm} \geq 0$, $MiscategorizedEmployees_{firm} \geq 0$, $sensitivity = [0,1]$

Here, $CategorizedEmployees_{firm}$ denotes the number of employees working for a firm that have been correctly categorized as an employee of said firm. $MiscategorizedEmployees_{firm}$ is the number of employees that have been categorized as not working for the firm although they do work for the specified firm. With this, it is possible to obtain the sensitivity as a decimal between 0 and 1.

The sensitivity on its own cannot be used to determine gender-based discrimination as it only entails a measure for model performance. Therefore, the AAP will be introduced to the model. Two different models will be trained; first a Standard XGBoost model (SXG), which will include gender as a variable and secondly an XGBoost model following the AAP principle of excluding gender from the dataset (AAPXG). For both models, the nineteen firms will be the

dependent variable and the sensitivity for each firm will be calculated. If a firm selects employees based on gender, albeit subconsciously, the sensitivity of the AAPXG will be significantly lower than the SXG.

The most important part when it comes to quantifying gender is to not include any discrimination imposed on an employee by someone else. For example, it is reasonable to employ someone with more experience but by doing this you also select based on gender as women have on average less experience. By including the aforementioned fourteen most important variables when it comes to building a workforce, the change in sensitivity will almost completely be caused by gender-based discrimination imposed by the current employer.

To capture the change in sensitivity for each firm, the *Gender Bias Ratio* (GBR) will be calculated simply by dividing the sensitivity resulting from the SXG-model by the AAPXG-model's sensitivity, as is shown in formula ABC

$$GenderBiasRatio_{firm} = \frac{Sensitivity_{SXG,\,firm,}}{Sensitivity_{AAPXG,\,firm,}} \qquad (2)$$

Where $GenderBiasRatio > 0$, $Sensitiviy = (0,1]$

$GenderBiasRatio$ does not significantly differ from one if no gender-based discrimination is present in the hiring process. There will be a $GenderBiasRatio > 1$ in case of discrimination based on gender. In case a model performs better when leaving out gender, the GBR will be lower than one, this could occur due to interaction effects with other firms whose GBR are higher but have comparable firm specific characteristics.

The GBR measures direct discrimination, thus discrimination emanating directly from registering someone's gender, albeit in text, images, or physical presence. In practice a lot of gender-based discrimination also prevails in other factors, e.g., experience; women are less likely to have relevant experience due to earlier experienced discrimination on the labor market. This can also be said for the education level of women, their job title, and the chance of being promoted (Watkins et al., 2006).

It should be noted that indirect discrimination is not in the scope of this research. Although indirect discrimination is a big problem for women, a hiring manager has the contractual obligation to hire the best applicant. During the hiring process, women will have less

qualifications because society has most likely discriminated against them during earlier events in their lives. Therefore, a non-biased rational manager will be less likely to hire a female applicant due to gender-based discrimination imposed by other parties in society.

Thus, GBR will help to quantify the costs of discrimination against women as imposed purely by their hiring managers. Therefore, the discrimination cost - if found significant – does not suffer from any biases coming from external sources, i.e., biasedness of others.

When the GBR is calculated for each firm, it can then be used to investigate the cost of discrimination. For this, the annual stock performance per firm and the GBR per firm will be used. Kendall's rank correlation tau test will be conducted to determine if there is any correlation between the market performance of a Dutch publicly traded firm and its GBR.

Building AAPXG and SXG

As mentioned above, this paper will use two variations of the traditional Extreme Gradient Boost model (XGBoost) to obtain results.

The XGBoost model is chosen firstly for its vast amount of hyperparameters that can be set, especially when comparing to regular boosting methods, this is important for this specific paper because no comparable research has been done or elaborately described, thus added flexibility will help to gain robust results.

Further, XGBoost manages missing values automatically and without significant losses in performance, which is important for this specific data as a lot of people do not completely fill out their LinkedIn profile, causing missing data points.

The third advantage of using XGBoost as the basis of this research can be considered the main reason for its success. XGBoost is mathematically optimized for cache optimization and CPU-parallelization, meaning it can manage a lot more data and calculations than its more traditional counterparts, such as Gradient Boosting and Random Forests. Given the vast amount of data in this research and because multiple models will have to be built, this third reason is also particularly important for choosing the best fitting model for this paper.

The main disadvantage of XGBoost models for this research is that XGBoost models are sensitive to outliers and overfitting (Choudhury, 2021). To avoid this problem, the data cleaning section, as elaborated in the respective section, checked for outliers and to avoid

overfitting a large dataset will be used and the hyperparameter tuning will be conservatively approached.

Tuning hyperparameters

XGBoost is – as the name implies – an extreme variation among machine learning algorithms, meaning it is very versatile and can be tuned in many ways. This is of immense importance as this tuning is focused on the use of big datasets. A more mathematical explanation of what XGBoost exactly entails can be found in its original paper by Chen & Guestrin (2016). In this section, the most notable features and characteristics of XGBoost will be considered.

The first step of XGBoost is similar to regular Gradient Boosting. XGBoost makes an initial Classification tree by randomly choosing the first classification, the so-called base score. This score is 0.5 by default, meaning it initially predicts the chance of a person working for a specific firm at 0.5. As there are nineteen different firms, and the biggest firm employs 'only' ~16% of the total scraped workforce, the default base score is much too high. Therefore, it will be changed to the average chance of working for a firm, being ~0.053.

Similarity scores, lambda, weights, and cover

This initial tree is evaluated based on its similarity scores. The similarity score of a tree consists of the sum of residuals squared, divided by the used probabilities for each residual (thus ~0.053 for the first tree) plus the lambda (λ). λ is added as a regularization term in the denominator with the goal of lowering the similarity score, especially for leaves with very few observations. By adding λ, the effect of outliers to the model are reduced and the results will be less prone to overfitting. The default setting for λ is one. This paper uses a grid search to tune all hyperparameters. For this, the input values are λ = [0.4, 0.8, 1.6]. Most existing literature uses lambdas within a range of 0.4 to 3.2. After running multiple versions of this paper's model, it turned out that lambdas above 1.6 results in very conservative models, making 1.6 a good cutoff. For each fitted hyperparameter, the used value is shown in table 6.

The models will also use 'minimal child weights'. This parameter determines the minimum number of observations that must be included in a tree's leaves to be added. This is to make sure trees will not make a leaf for each single observation, which would be very valid internally but useless in the real world as characteristics that hardly occur are most likely determined by chance. To prevent this overfitting behavior, a minimum child weight range of

[1, 5, 10] is used. This range is based on projects with similar observation and variable numbers. Another important thing to note is that XGBoost uses a cover. The cover is the percentage of observations that is categorized by a node based on a specific feature over multiple trees. This is calculated for each feature in each node, if the cover does not exceed the preset threshold, the node is not considered important enough and will not be used. This is another measure to fight overfitting in the model. By default, the cover is set to 1. Given the fact this paper already has multiple measures against overfitting in place, the default cover is used.

Gains, gamma, and pruning

After calculating the similarity score for each leaf and node, the gain of each node can be calculated. This is done with the sum of the similarity scores of each node's leaves minus the similarity score of the node itself. The gain is used to evaluate the relative quality of a node, with a higher gain meaning the node and its leaves are more useful than nodes with a lower gain. Not only is the gain important when choosing the right threshold for each node, it is also used to prune the tree.

Pruning a tree is necessary as it ensures the XGBoost model does not grow indefinitely, causing the tree to overfit and become unworkable due to its size and lack of external validity. Pruning is done by subtracting the gamma (γ) from the gain. If the outcome is positive, the node will be kept, if the outcome is negative, the node will be pruned. For this, it is important to set the right γ, or Lagrangian multiplier as some call it. γ makes trees shallower, this paper however has quite a lot of data and different variables and thus needs relatively deep trees. Therefore, the γ will be included in the grid search and a conservative range is searched: γ = [0.4, 0.8, 1.6]. The optimal value can be found in table 6.

Building new trees and learning rate.

After pruning, XGBoost, in a similar fashion to regular Gradient Boosting, calculates the output value for each leaf, and uses the log(odds) plus the learning rate (ε) multiplied by the output of the tree to determine the values of the next tree, after which the calculations and pruning starts from the beginning.

ε is set to 0.3 by default and decides how much of the previous tree is considered when building a new tree. Setting ε to low will result in a very slow learning and tedious model,

setting ε to high will give very quick results but a high likelihood of a local optimum, as the machine learning process only considers specific patterns and does not generalize enough. To find the optimal ε for this paper, ε is included in the grid search: ε = [0.005, 0.01, 0.1, 0.2]. These values are chosen because they are in line with existing literature and because after the first trials it turns out the dataset is too small for lower etas, making models with lower etas very conservative. Choosing etas higher than 0.2 resulted in inconsistent results, suggesting local optima. Again, the chosen value for ε can be found in table 6.

Max depth, subsample, and data split

For XGBoosting, it is important to set the maximum depth for the trees. The maximum depth denotes the distance from the root of the tree to the farthest leaf. When a tree is too shallow, relevant information and splits will be left out of the tree. On the other hand, a very deep tree can result in overfitting as leaves will be built based on little sample data. Thus, the maximum depth controls the complexity of the tree to be built. The default maximum depth equals six. Because this paper uses a relatively high number of variables and data points, it may be possible to find better results with deeper trees. The search for the optimal depth will be in the following range: max_depth = [4, 7, 10].

Another method used to prevent overfitting is adding a 'subsample rate'. The subsample rate denotes what percentage of the training data will be randomly chosen for each folding iteration. Using all training data for each fold creates a low external validity. Therefore, the subsamples will be set at; subsample = [0.5, 0.75].

The same procedure has been conducted for the variables. By adding the same parameter range for the columns of the dataset, each iteration only uses 50 or 75 percent of the explanatory variables, this is used to make sure only useful variables are in the final model and to give an idea of the importance of each inputed feature.

The use of a column and row subsample rate comes with the added value of making the overall model less influenced by correlating variables. Although the data in this paper is carefully selected and filtered, it is possible that, when calculating relative importances, correlated variables are affecting the overall results. In a standard XGBoost model, two variables which are perfectly correlated lead to the removal of one of the variables; this is a built-in function (Chen et al., 2018). When two variables are partially correlated, however, XGBoost tends to under evaluate the variable that is selected last, which is determined by

chance and the order of data. This is due to the fact that the last chosen variable is contributing less to the gain of the overall model. This phenomenon is especially important for the gender variable in this research because it correlates with several other variables such as education, which may cause the models to underestimate the importance of gender or any of the other variables. As gender is one of the major focuses in this paper, it is important to circumvent this problem.

The addition of the two subsample rates in combination with the 5-fold cross validation will help the model to give each variable the correct importance. The model will select, for example, variable A in 50% of the trees and variable B in the other 50% of the trees, which will be done 5 times for each parameter combination. XGBoost is designed in such a way the model will not refocus on variables it has already learned (Chen et al., 2018). Thus, if it has determined the effects of using variable A, it will hardly use variable A in new trees and it will focus on variable B. The models in this paper will be able to see the correlating variables functioning independently form each other because of the subsampling, which in combination with the fact the XGBoost models is designed to focus on learning variables it does not know yet makes sure the calculation of feature importances is very robust against correlations (Chen et al., 2018).

To make sure the XGBoost model is not only internally valid, the trees will be built using a training set and will then be evaluated on the test set. The training set contains 80% of all LinkedIn data and the test set contains the remaining 20% of the data.

Cross validations and number of rounds

The number of rounds is a great parameter to influence the performance as well as the time it takes to train the model. Initially, the number of iterations will be set at one thousand, with the early stopping parameter of 50, meaning that the machine learning process will be done at most 1000 rounds but will be evaluated every 50 iterations. If the performance of the model declines during fifty iterations, the model will automatically stop and save the best performing generation. Table 6 shows the optimal number of iterations for both the SXG and AAPXG model.

It is important that the final model is not only internally valid but can also be used externally. Therefore, on top of subsampling and using test data, K-Fold cross validation is used. In this case, the K will be set to 5. This means each tree will be evaluated each iteration against five different subsamples of data, giving a more accurate and meaningful accuracy for the respective iteration.

Used parameters

After tuning all parameters above, which resulted in 1296 5-folded models, the parameters shown in table 6 are found to be optimal. This table shows four different models, the normal AAPXG and SXG, but also two extra models named 4AAPXG and 4SXG. These two extra models are included to dive deeper into the most significant firms, on which the result section will elaborate further. The methodology of these extra models is identical to the original two, with the only difference being the usage of only four firms as the dependent variable instead of the original 19 firms. These four firms are ING, ASML, ABN AMRO and Philips.

Table 6 shows a high consistency among the optimal hyperparameters. The main differences can be found in the learning rate, which is respectively ten and twenty times lower for the initial models. This can be explained by the higher amount of data and classes that the model must accommodate for, making it harder for XGBoost to find underlying patterns quickly. This difference is further shown by the lower number of trees in the 4SXG and 4AAPXG models, as a higher learning rate makes sure you reach the ultimate model quicker.

Model	Depth	Min. child weight	Subsample	Subsample columns	ε	γ	# Trees	λ
SXG	4	5	0.5	0.75	0.01	1.6	1000	0.4
AAPXG	4	5	0.5	0.75	0.005	0.8	1000	0.4
4SXG	4	1	0.75	0.5	0.1	3.2	490	0.8
4AAPXG	7	5	0.75	0.75	0.1	3.2	507	1.6

Table 6: Optimal hyperparameters for each model given the parameter grid.

With the tuned models it is now possible to take a deeper dive and discuss the results coming from the models.

Kendall's rank correlation tau

To see if there is any relation between gender-based discrimination and performance, Kendall's rank correlation tau will be used. This method was chosen as it can deal with non-normal distributions and small datasets. This is important as only nineteen firms will be compared and because the GBRs cannot be expected to be normally distributed because it is plausible many firms have ratios around one with only part of the firms having significantly different ratio's, creating a skewed distribution.

The performance of each company will be based on the stock returns over the last three years. Three years is chosen because the average employee in this dataset has been working at their current employer for almost exactly three years. This makes sure the performances are relevant to the current workforce of the company. This test is done with both the output data of the models and using ranks to normalize the market returns. All firms receive two ranks from 1 to 19 based on their GBR and performance. The firm with the lowest GBR will receive rank one, the second lowest GBR will rank 2nd, etc. For the performance ranking the firm with the highest returns over the last three years will rank first and the firm with the lowest returns will be ranked last.

In case multiple GBRs or returns are exactly equal, the used rank will also be equal. The rank for the following firm in this case will only increase by one. In other words, if three firms finish second based on their GBR, these firms all receive the rank '2'. The firm with the next lowest GBR will receive rank 3 instead of 5. These two tests together will be used to check for any sensitivities coming from the distribution in the market data, this will further be checked using other common correlation measures as a robustness check.

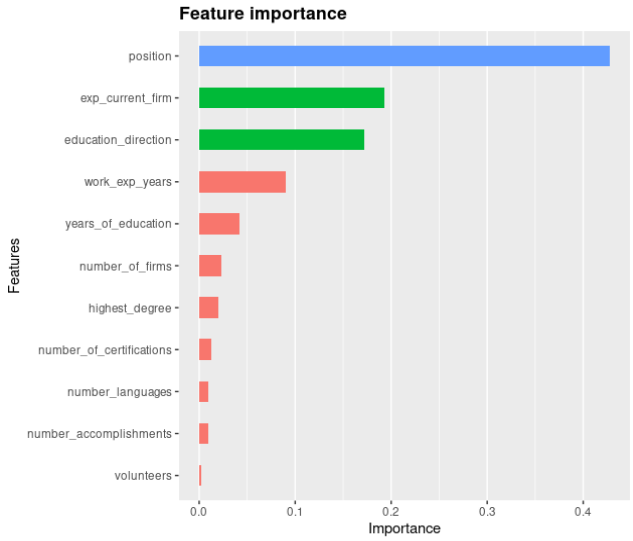
Results

SXG and AAPXG

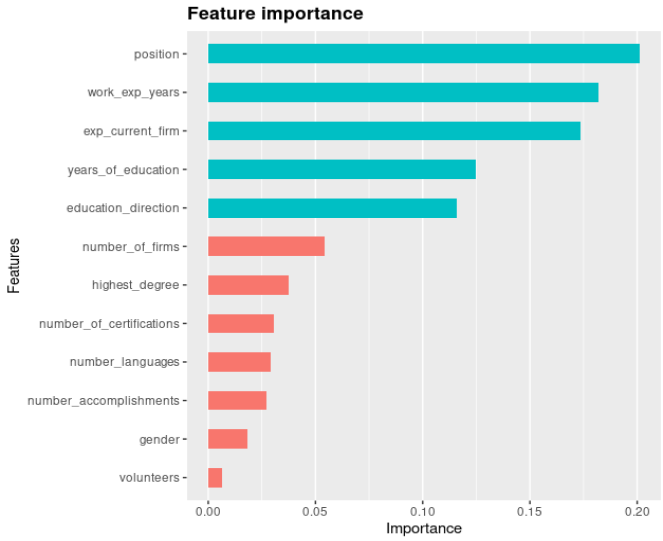
The initial results of the SXG and AAPXG models can be found in table 7. This table shows how both models manage to classify about 39% of the LinkedIn profiles into the correct firm. As discussed before, the difference between these models is the exclusion of gender in the second model. The accuracy of both models does not show any significant difference between the models. This means deleting gender does not significantly influence the accuracy of the models.

Measure	SXG	AAPXG
Accuracy	0.3937	0.388
95% CI of Accuracy	(0.3641, 0.4240)	(0.3585, 0.4182)
Kappa	0.3087	0.2925

Table 7: Accuracy and Kappa for SXG and AAPXG. Accuracy is denoted as a percentile



Graph 1: Feature importance for AAPXG. The importance is shown as gain. Exact numbers can be found in Appendix B2. Colors indicate relative importance.



Graph 2: Feature importance for SXG. The importance is shown as gain. Exact numbers can be found in Appendix B1. Colors indicate relative importance.

.

It is possible this means gender does not hold a significant amount of information for the models, or gender might be negatively correlated with other variables that cancel each other's effects.

To investigate this, graphs 1 and 2 show the feature importances, the exact numbers can be found in appendix B1 and B2. The feature importances are measured using the gain for each feature i.e., how much the accuracy improves due to a specific feature. Two other measures that can be found in the appendix are the cover, giving the percentage of observations affected by the feature, and the frequency, which shows the percentage of splits in which the feature is used.

To determine whether someone is working for a specific firm, the SXG is mostly interested in someone's position and experience, whether it is current or total. The top five features account for 79% of all gains in accuracy and these five are used in 72% of all splits in the trees. Gender does not seem particularly important when looking at the SXG as it is considered in only 2% of all splits. Gender is ranked second to last when using the SXG, but things get more interesting when looking at the feature importances of the AAPXG model.

The AAPXG as shown in graph 1 and appendix B2 tells a different story when it comes to feature importances. The absence of someone's gender results in the halving of the importance of educational and working experience. It also results in someone's position to become by far the most important driver when classifying into firms. The position variable more than doubles its importance and now accounts for 43% of the total gain, the top 5 features now have a total gain of 92%. The frequency measure of the top 5 does not undergo a substantial change as it increases only 6% to 78%. This smaller change can be explained by the positioning of the splits. If someone's position is mostly considered in the top of the trees, this will result in a lower frequency than features which are mainly considered in the later stages of the trees.

An explanation for the sizable increase in the importance of positions can be found in a negative correlation between gender and positions. However, when looking at the data, as shown in table 8, there seems to be no significant difference in the positions of male and female employees in this dataset. This is confirmed by both the pairwise and ranked spearman correlations which are around -0.035.

Another explanation can lie in the same correlation but for specific firms. When looking at the correlation between gender and positions for each specific firm, which is shown in appendix C, it is found that several firms have a negative correlation between their employees' gender and positions. At a significance level of 5% Unilever, ABN AMRO and Boskalis have a negative correlation, at 10% Adyen and KPN must be added to this list. These companies account for 26.2% of all included employees so a negative association between gender and position can be a crucial factor for the differences between SXG and AAPXG.

The size of the position importance is large and remains large when altering parameters, training sets and test sets. The size can be a bit more nuanced when considering the cover and frequency of the features, as can be found in Appendix B1 and B2. Position ranks as third highest when it comes to its cover in the AAPXG, and it scores fourth place for its frequency. This is still a sizable jump compared to the original SXG, but it does suggest the gain on itself might be exaggerating the position's importance.

Level	male	%	female	%
graduate	186	8.62%	144	11.83%
medior	1185	54.94%	692	56.86%
senior	221	10.25%	111	9.12%
management	682	31.62%	372	30.57%
top	69	3.20%	42	3.45%
Total	**2157**	**100%**	**1217**	**100%**

Table 8: Relation between gender and position. Both numerical and as a percentage of all male or female employees.

To quantify gender-based discrimination, it is important to look at the sensitivities and the GBR of each company; this is shown in table 9. This table shows that for some companies it is hard to classify their employees, resulting in sensitivities equal to zero. ASML and ABN AMRO turned out to be the most distinctive companies as they have the highest sensitivities by far. This might be caused by actual firm specific effects, but the size of their workforce also means the models have more information to work with during the classification process.

For six firms the classification rate completely diminished resulting in zero sensitivity in AAPXG, i.e., none of their employees are classified correctly in AAPXG. The differences in the sensitivities for some of these six companies look small at first glance but for bigger companies such as Shell

it means that 21 employees were correctly classified in the SXG and none in the AAPXG. This decline to zero for these firms is a consistent occurrence over multiple runs and varying hyperparameters.

For four other firms the sensitivities decline as well but to a value greater than zero. For ASML and Philips the sensitivities are lower but stable, resulting in GBRs close to 1. For ING and Unilever, the GBRs are 1.8956 and 2.0011 respectively, meaning half as many of their employees are classified into the right firm in the AAPXG. For ING this means 68 employees are misclassified due to the exclusion of gender, for Unilever the performance decline is nominally smaller with only 13 extra misclassifications. This is a small decline and may seem insignificant. However, this decline of roughly eight percent, or about 13 profiles, is very consistent when changing parameters, seeds, training sets, testing sets and evaluation methods. These two firms thus show a high GBR, which suggests these firms can be susceptible to gender-based discrimination.

On the other side of the GBR-spectrum ABN AMRO has the only GBR below zero, meaning the gender variable was holding the model back to make the best decisions for ABN AMRO. The low GBR results in a nominal increase of 114 correctly classified ABN AMRO employees.

Culturally it is unlikely that ABN AMRO is subconsciously biased against both men and women, a more likely explanation for this increase lies in the higher GBR of ten other firms. Logically, if SXG misclassified ABN AMRO employees into these other ten firms due to their gender, these employees are more likely to be correctly classified in the AAPXG where gender is not included. This is because the ten firms now have lost a characteristic that caused employees to be classified into their respective firms.

One could expect the same effect for ASML, but it is quite likely that ASML's workforce is a lot less affected by characteristics of other firms as they are much more specialized towards tech than any other firm in the Netherlands.

Given the results of the SXG and AAPXG models, it becomes apparent that excluding gender makes it a lot harder to classify employees for most firms. To take a closer look at his results, the four firms that had a sensitivity significantly above zero (ABN AMRO, ASML, Philips and ING) are examined in two extra models; the 4SXG and 4AAPXG.

Firm	Sensitivity SXG	Sensitivity AAPXG	GBR
ABN AMRO	0.6780	0.8475	0.8000
Adyen	0.0303	0	inf
Ahold Delh.	0.08	0	inf
AkzoNobel	0	0	-
ASML	0.6504	0.6260	1.0390
ASR	0	0	-
DSM	0	0	-
Heineken	0	0	-
ING	0.2813	0.1484	1.8956
KLM	0	0	-
KPN	0.0313	0	inf
NN	0	0	-
Philips	0.3529	0.3294	1.0713
PostNL	0.0250	0	inf
Shell	0.0877	0	inf
Boskalis	0	0	-
Ordina	0.0909	0	inf
Unilever	0.1765	0.0882	2.0011
Unknown	1	1	1.0000

Table 9: Sensitivities for each firm in SXG and AAPXG with their resulting GBR. The unknown is included but holds little information as an unknown employer results in an unknown work experience in the current firm, making it extremely easy to classify.

Given the high number of zeroes in table 9, other measures are considered to evaluate the models. However, no other measure resulting from an XGBoost model could be found that gives the possibility to investigate the research question without having many missing values.

4SXG and 4 AAPXG

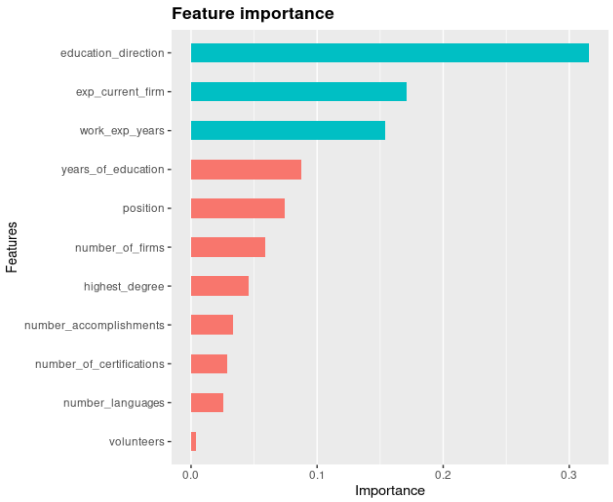
The 4SXG and 4AAPXG are identical to their predecessors, with the difference being only four firms are included. These firms are ING, ASML, ABN AMRO and Philips. For both models the accuracy and kappa are next to identical, as shown in table 10. Graphs 3 and 4 show the feature importances, which can also be found in appendix D1 and D2. The feature importances in this instance do not vary as much as they did in previous models. The main difference can be found in the education direction which becomes less important when excluding gender. All other features -except for volunteering- do not change much but are leveled out in the 4AAPXG as compared to the 4SXG. The minor changes suggest that for these four companies, gender does not seem to be influencing other factors as it did in the initial model.

Measure	**4SXG**	**4AAPXG**
Accuracy	0.5000	0.4866
95% CI of Accuracy	(0.4563, 0.5437)	(0.4582, 0.5356)
Kappa	0.3124	0.3156

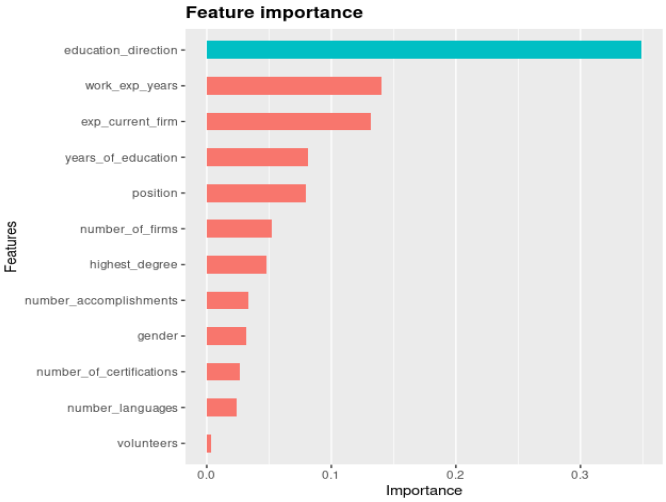
Table 10: initial performance results for 4XG and 4AAPXG where accuracy is denoted as percentile

When looking at the GBR and the sensitivities for each firm (Table 11) it is even clearer that gender does not influence the results of the newer models much. For ASML the GBR is almost exactly one and for both Philips and ABN AMRO the GBRs lie close to 1. ING is the only firm with a more differing GBR but given it is less than 1 and because the difference affects less than 2% of ING's workforce classifications it is not expected to indicate any gender-based discrimination for ING or any of the other three firms.

Based on the feature importances, overall model performances and the GBRs it becomes clear that the 4SXG and 4AAPXG do not quantify gender-based discrimination and should therefore not be used for the performance analysis in the following paragraph.



Graph 3: Feature importance for 4AAPXG. The importance is shown as gain. Exact numbers can be found in Appendix D2. Colors indicate relative importance.



Graph 4: Feature importance for 4SXG. The importance is shown as gain. Exact numbers can be found in Appendix D1. Colors indicate relative importance.

Firm	Sensitivity 4SXG	Sensitivity 4AAPXG	GBR
ABN AMRO	0.7169	0.6627	1.0817
ASML	0.6212	0.6288	0.9879
ING	0.2139	0.23913	0.8945
Philips	0.3523	0.32955	1.0690

Table 9: Sensitivities for each firm in SXG and AAPXG with the resulting GBR.

Kendall's rank correlation tau

Kendall's rank correlation tau has been used to gain insights on the relation between stock performances and GBRs. Kendall's rank correlation tau gives some remarkably interesting insights. The tau coefficient has an estimate of -.4182, with a p-value equal to 0.08656. To ensure the robustness of this coefficient, Spearman's ranked test and Pearson's product correlation are also calculated. These two alternative methods lead to terribly comparable results.

The negative correlation suggests there is a negative relation between the GBR-rank and the stock performance ranking of firms. A low GBR in that case also correlates with low returns,

suggesting gender-based discrimination leads to higher monetary gains for Dutch publicly traded companies. Of Course, the p-value is not significant at 5% which is often the norm, but it is significant at 10% which suggests further research is needed to confirm these results.

Assuming Kendall's tau is indeed negative it is interesting to look at any rationale behind this. One explanation could be the small number of firms in this paper in which case further research comparing more companies will gain more robustness. Another explanation could be that firms which do not focus on gender when establishing their workforce are more socially oriented and less driven by pure profits and vice versa. If this is the case, then a behavioral study on Dutch firms may help to gain better understandings on the relation between gender-based discrimination and stock performances.

Conclusion

This paper used four different variations of the XGBoost model to gain insights on gender-based discrimination. Over four thousand different LinkedIn profiles were scraped from each of which fifteen different variables could be extracted and used in the models. By comparing classification models that either could or could not use an employee's gender some interesting results were obtained. Evidence has been found suggesting that for the models it becomes harder to classify employees into the correct firm when excluding gender. Although this difference can be completely countered by firms without any gender-based discrimination. The GBR shows a decline in sensitivity for ten out of the eleven firms with significant results. This suggests the GBR can be used to quantify gender-based discrimination and gives insights into discrimination over different firms. When taking a deeper dive, the results suggest the position of an employee becomes extremely important when excluding genders from the dataset. Although no evidence for a relation between gender and position can be found for the dataset, the firms with the biggest jumps in sensitivity also turned out to be the firms that did have a correlation between their workforce's gender and the positions within the firm.

Ultimately the goal of this paper was to find if there is any monetary consequence of gender-based discrimination. After running multiple tests, it must be concluded that no significant results can be found at 5% significance but the results available suggest a positive relation between gender-based discrimination and stock performance at 10% significance. This is unexpected and further research should therefore focus on behavioral explanations for this as the results might be due to socially focused firms having less discrimination but also less focus on monetary profits. Another interesting addition that would counter the current limitation in the number of firms researched would be adding more firms, if possible, internationally, to get more significant and externally valid results.

With these results it is possible to conclude that there is gender-based discrimination but its effect on the performance of a firm is extremely limited or non-existent.

References

The Drum, (2022), ''Dutch women change their name to Peter on LinkedIn to protest lack of diversity'', Retrieved from: https://www.thedrum.com/news/2022/02/02/dutch-women-change-their-name-peter-linkedin-protest-lack-diversity

Senden, L. (2018), "Een vrouwenquotum gaat om het rechtzetten van discriminatie". Retrieved from: https://www.uu.nl/nieuws/een-vrouwenquotum-gaat-om-het-rechtzetten-van-discriminatie

Åslund, O. & Nordström Skans, O. (2012), ''Do anonymous job application procedures level the playing field?'', Industrial and Labor Relations Review, Vol. 65, No. 1, 85-107

Hall, D. J., & Davis, R. A. (2007). Engaging multiple perspectives: A value-based decision-making model. *Decision Support Systems*, *43*(4), 1588-1604.

Bernile, G., Bhagwat, V., & Yonker, S. (2018). Board diversity, firm risk, and corporate policies. *Journal of financial economics*, *127*(3), 588-612.

Rivera, L. A. (2012). Hiring as cultural matching: The case of elite professional service firms. *American sociological review*, *77*(6), 999-1022.

GRPR, 2018, 'General Data Protection Regulation' retrieved from: gdpr-info.eu

Chapple, L., & Humphrey, J. E. (2014). ''Does board gender diversity have a financial impact? Evidence using stock portfolio performance'', Journal of business ethics, 122(4), 709-723.

Gul, F. A., Srinidhi, B., & Ng, A. C. (2011). ''Does board gender diversity improve the informativeness of stock prices?'', Journal of Accounting and Economics, 51(3), 314-338.

Watkins, M. B., Kaplan, S., Brief, A. P., Shull, A., Dietz, J., Mansfield, M. T., & Cohen, R. (2006). ''Does it pay to be a sexist? The relationship between modern sexism and career outcomes'', Journal of Vocational behavior, 69(3), 524-537.

Page (2007), ''The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies'' - New Edition

Chen T., He T., Benesty M., Tang Y., "Understand your dataset with XGBoost", retrieved from: https://cran.r-project.org/web/packages/xgboost/vignettes/discoverYourData.html#numeric-v.s.-categorical-variables

DeGiro, (2021), "Company Profile", retrieved from: https://trader.degiro.nl/trader/#/products/[IDENTIFICATION_NUMBER_OF_FIRM]/company

Francesconi, M., & Parey, M. (2018). Early gender gaps among university graduates. *European economic review*, *109*, 63-82.

Choudhurry (2021). "Top XGBoost interview questions for Data Scientists'', Retrieved from: https://analyticsindiamag.com/top-xgboost-interview-questions-for-data-scientists/

Chen and Guestrin (2016) 'XGBoost: A Scalable Tree Boosting System'

CBS (2021). '' Bevolking; leeftijd, migratieachtergrond, geslacht, regio, 1 jan. 1996-2020'',

Retrieved from:

https://opendata.cbs.nl/statline/#/CBS/nl/dataset/37713/table?dl=FE77&ts=1534860821677

Appendix A

Firm name	Index	Sector	Employees
ADYEN	AEX	FS&B	2,180
Aegon	AEX	FS&B	3,500
Kon. Ahold Delhaize	AEX	Food	100,911
Akzo Nobel	AEX	Chemicals	2,500
ASML Hold.	AEX	Tech	26,614
Kon. DSM	AEX	Food	3,858
Heineken	AEX	Food	3,700
IMCD	AEX	Chemicals	3,740
ING groep	AEX	FS&B	14,754
Kon. KPN	AEX	Household	9,699
Kon. Philips	AEX	Household	11,142
RANDSTAD NV	AEX	Outsourcing	4,320
NN Group	AEX	FS&B	5,000
Shell PLC	AEX	Commodities	11,123
Signify NV	AEX	Household	2,250
UNILEVER PLC	AEX	Food	2,500
Air France-KLM	AMX	Transport	22,918
ABN AMRO Bank NV	AMX	FS&B	21,670
ASR Nederland NV	AMX	FS&B	4,500
Boskalis Westminster	AMX	Construction	9,500
Corbion NV	AMX	Chemicals	2,493
InterTrust NV	AMX	FS&B	2,500
JDE Peets NV	AMX	Food	2,400
PostNL NV	AMX	Logistics	20,421
SBM Offshore	AMX	Commodities	5,019
TKH Group	AMX	Tech	5,583
BAM	ASCX	Construction	8,000
Heijmans	ASCX	Construction	4,839
Ordina	ASCX	Tech	2,583
Sligro Food Group	ASCX	Food	11,000

Table A1: Publicly traded firms that employ more than 2000 employees within the Netherlands. All firms are traded on the three major indices in the Netherlands and include their sector and employee number within the Netherlands only.

FS&B	Financial Services & banking
Food	Food manufacturers and/or distributors
Chemicals	Manufacturers and/or distributors of chemicals
Tech	Firms that sell and develop innovative machines
Household	Manufacturers and/or distributors of household appliances
Logistics	Distributors of third-party goods
Transport	Firms that transport people and goods
Construction	Construction companies
Commodities	Firms active in commodities such as oil, gas, and other nonrenewable sources

Appendix B

Feature SXG	Gain	Cover	Frequency
Position	0.20	0.08	0.06
Work experience	0.18	0.22	0.22
Experience current firm	0.17	0.17	0.18
Years of education	0.12	0.16	0.17
Education direction	0.12	0.11	0.09
Number of firms	0.05	0.07	0.07
Highest degree	0.04	0.05	0.05
Number of certifications	0.03	0.05	0.04
Number languages	0.03	0.04	0.04
Number accomplishments	0.03	0.04	0.04
Gender	0.02	0.02	0.02
Volunteers	0.01	0.01	0.01

Table B1: Feature importances for SXG, presented in Gain, Cover and Frequency and sorted on Gain. All used features are presented.

Feature AAPXG	Gain	Cover	Frequency
Position	0.43	0.16	0.11
Experience current firm	0.19	0.16	0.18
Education direction	0.17	0.16	0.13
Work experience	0.09	0.20	0.21
Years of education	0.04	0.12	0.15
Number of firms	0.02	0.07	0.07
Highest degree	0.02	0.04	0.05
Number of certifications	0.01	0.04	0.04
Number languages	0.01	0.02	0.03
Number accomplishments	0.01	0.02	0.03
Volunteers	0.00	0.01	0.01

Table B2: Feature importances for AAPXG, presented in Gain, Cover and Frequency and sorted on Gain. All used features are presented.

Appendix C

Firm	Kendall's Tau	P-value
ABN AMRO	-0.08	0.080
Adyen	-0.13	0.078
Ahold Delh.	-0.09	0.174
AkzoNobel	0	0.505
ASML	-0.04	0.207
ASR	-0.04	0.342
DSM	0	0.480
Heineken	-0.03	0.396
ING	-0.04	0.213
KLM	0.01	0.558
KPN	-0.12	0.097
NN	0.12	0.926
Philips	0.1	0.973
PostNL	0.02	0.570
Shell	0.04	0.269
Boskalis	0.25	0.041
Ordina	-0.04	0.347
Unilever	-0.16	0.030

Table C1: The correlation coefficient between gender and the position of an employee within each firm and the respective p-values.

Appendix D

Feature 4SXG	Gain	Cover	Frequency
Education direction	0.35	0.20	0.18
Work exp years	0.14	0.17	0.19
Experience current firm	0.13	0.14	0.16
Years of education	0.08	0.08	0.12
Position	0.08	0.10	0.08
Number of firms	0.05	0.08	0.07
Highest degree	0.05	0.06	0.05
Number accomplishments	0.03	0.05	0.05
Gender	0.03	0.03	0.03
Number of certifications	0.03	0.05	0.04
Number languages	0.02	0.04	0.04
Volunteers	0.00	0.01	0.01

Table D1: Feature importances for 4SXG, presented in Gain, Cover and Frequency and sorted on Gain. All used features are presented.

Feature 4AAPXG	Gain	Cover	Frequency
Education direction	0.30	0.15	0.13
Work experience	0.16	0.18	0.20
Experience current firm	0.16	0.17	0.19
Years of education	0.10	0.11	0.14
Position	0.07	0.09	0.07
Number of firms	0.06	0.09	0.08
Highest degree	0.05	0.06	0.05
Number accomplishments	0.03	0.04	0.04
Number of certifications	0.03	0.05	0.05
Number languages	0.03	0.04	0.04
Volunteers	0.00	0.01	0.01

Table D2: Feature importances for 4AAPXG, presented in Gain, Cover and Frequency and sorted on Gain. All used features are presented.