Northeastern University

Module 12 Assignment:

ALY6980 Capstone Project

Authored by Group 4

Abhigna Ramamurthy,

Charu Priya Singh,

Jeseeka Shah,

Karolina Grodzinska

Submission Date: 7/2/2023

Introduction

This project focuses on examining the data provided by Free Float Media. Each group member worked individually to answer various intermediary questions and discover patterns in the data. Many of those questions were related to the diversity of boards of directors and female representation among directors. Our efforts as a group are to evaluate each variable responsible for affecting the performance of the board of directors leading up to a model accurately predicting the performance of the board. The dataset made available to us has been the starting point of our project. We began with Exploratory Data Analysis and tested various hypotheses. At this stage, we individually and as a group performed an in-depth analysis to understand the data and the industry. We have actively made efforts to combine our learnings from the literature review to further our efforts into building a successful project.

Board climate-responsible orientation (BCO), according to a peer-reviewed study published in Frontiers in Environmental Science, enhances company carbon performance. The study also indicated that corporate carbon performance is improved with more BCO experience and that carbon awareness lowers symbolic emission reduction activities while firm reputation increases symbolic emission reduction actions in carbon management. Another peer-reviewed study discovered that the composition of the board, remuneration, and carbon intensity influence both the decision to employ assurance and the supplier of assurance. According to a 2015 meta-analysis of 140 research articles on the link between female board representation and performance, there is a positive relationship between female board involvement and accounting returns, but no relevant relationship with market performance. Women on boards may improve a company's reputation, contribute to the creativity and originality of board debates, and help in the recruitment and retention of talented female workers who respect firms with women in executive and board leadership roles.

Studies like “Board Gender Diversity Codes, Quotas and Threats of Supranational Legislation: Impact on Director Characteristics and Corporate Outcomes”, “Do Women Leaders Promote Sustainability? Analyzing the Effect of Corporate Governance Composition on Environmental Performance”, and “Impact of Women Directors & Board Independence on Family Firm Performance: Evidence from India” literature reviews have helped me understand how the process can be achieved.

Analysis

We have used various techniques to analyze the dataset provided by the sponsor. The main research methodology used would be to go through peer-reviewed journals that can shed insight into the research already undertaken. This helps us not to reinvent the wheel, but to understand what we can add to the table given the constraints of the project. The software platforms that we have used for the project will include Jupyter Notebook, Tableau, Power BI, MS Excel, and Python programming language. The main libraries used include pandas, NumPy, and datetime for data cleaning and manipulation, seaborn, and plotly, for visualization, and statsmodels, and sklearn for data modeling. The steps I will follow are as follows:

1. **Design and data collection**

   As data analysts, the first step of the analysis is to understand the business objectives. The object is to hypothesize on the gender diversity of the company's board of directors contributing to the carbon intensity performance of the company. The influence of female directors also plays an important part of this hypothesis. Several data can help achieve this including the data provided by the sponsor which has director and company core metrics with already quantified data. The research from "The Effect of Board of Directors and Audit Committee Characteristics on Company Performance in Jordan" shows that board size, CEO duality, and gender diversity were all favorably connected to business performance, but director ownership, board meetings, and the proportion of board members with PhDs were all adversely related. These are the kind of hypotheses we will try to test with analytics methods.

2. **Data exploration and analysis**

   The next step is to understand how the data is spread across and quantify the relationship among the different variables available for analysis. Exploratory data analysis can help uncover hidden patterns in the data and generally help get to know more about the data. The next step would be to prepare the data for modeling by removing unwanted variables or feature engineering, balancing data, and scaling data as needed. Finally, model results can be showcased in the form of dashboards using Tableau. The overview of the method followed is given in the below pictorial representation of the whole process.

3. **Data Management and governance roadmap**

The total administration and control of data assets inside an organization is referred to as data governance. Guarantee that data is managed in a consistent, dependable, and compliant manner, it entails the construction of a structure, methods, and norms. Data governance mainly includes Data quality maintenance, Privacy and Security, Compliance, Data Stewardship, Data Management, conforming to the standards of accessing and sharing, data standards, and documentation. The 4 steps and methods followed for this project are:

a. Understanding the why behind the problems to illustrate the **challenges**. The key challenges I wanted to solve using data governance are Data quality issues affecting business decisions, Lack of data standards, Understanding more about the data sources, and Limited data analytics capabilities due to the low quality of data.

b. Understanding the **objectives** for data governance. This step included improving data quality by eliminating unnecessary details in the data provided, handling missing values by imputation or deletion, enhancing data analytics platforms and solutions, and delivering actionable insights.

c. Analyzing **capabilities** and course of **action**. Python libraries like bamboolib was helpful in quickly understanding what the issues are within the data, and libraries like Pandas, and Numpy helped resolve the issues. Data reporting platforms in the form of a detailed proposal, documentation for the code as well as Tableau reporting was used.

d. Creating a **visualization** of the **roadmap** to further enhance the capabilities of the dataset. The next steps that the sponsor company, Free Float Media can use to improve the data governance is shown in Figure 1 below. This would help further the insights and provide better recommendations for their clients.

Time Horizon	Stage 1	Stage 2	Stage 3	Stage 4
Year 1	Data quality assessment	Data governance framework	Data engineering team	Data reporting platform
Year 2	Data quality improvement	Data governance implementation	Data engineering solutions	Data analytics solutions

Figure 1: A possible roadmap for the data governance project for Free Float Media

Preliminary results

In this section, we will introduce our preliminary results. During the exploratory phase, we discovered that there are only about 22% of female directors.

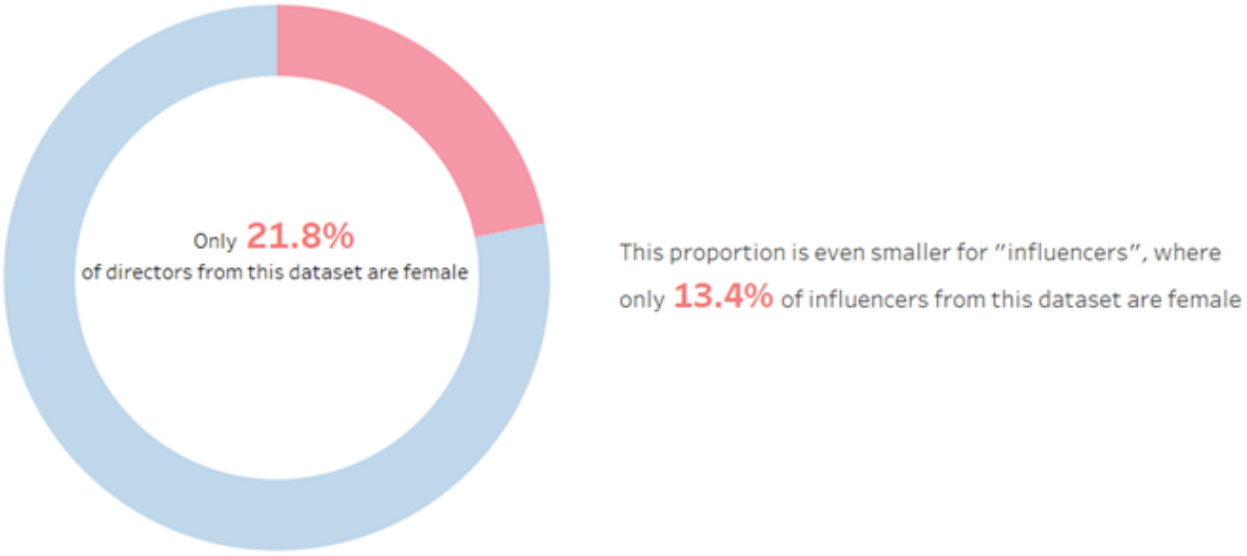

Figure 1: Proportion of female directors in the dataset

The percentage of women in the Industrials and Financial sectors is among the most diverse companies as seen in Figure 2.

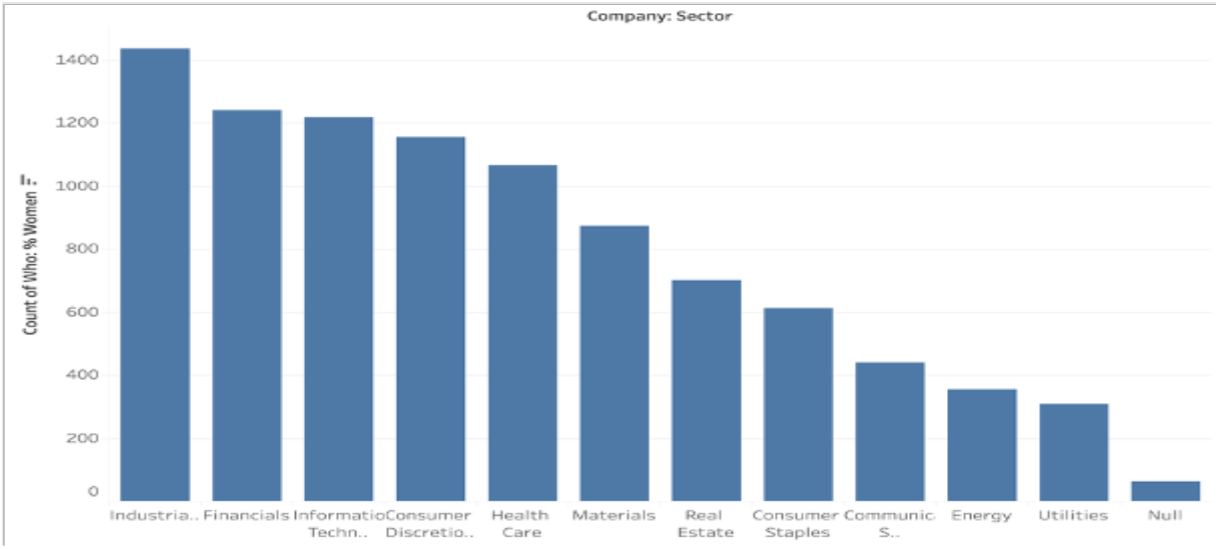

Figure 2: The percentage of women on boards of directors is the highest in Industrial and Finance firms

Additionally, the boards of directors of analyzed companies are mostly male-dominated (having over 50% of male directors):

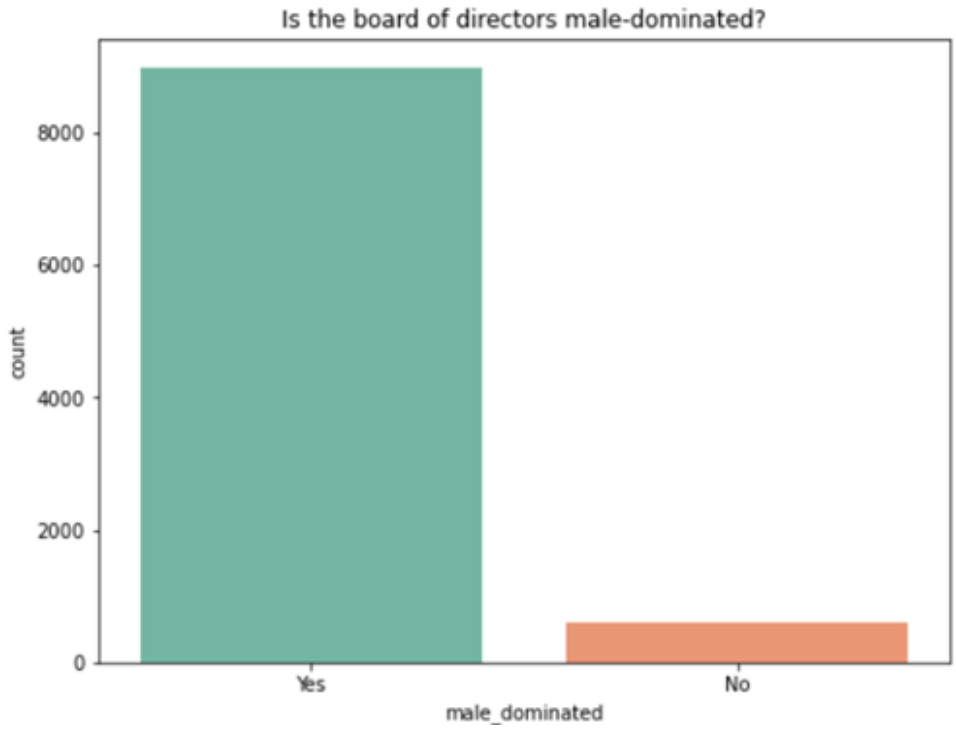

Figure 3: Most firms are dominated by male directors

Now, we can look at some differences between male and female directors.

Hillman et al. (2002), found out that male directors more often come from a business background and that female directors more often hold advanced degrees. The authors state that educational credentials give women an advantage in the male-dominated business world.

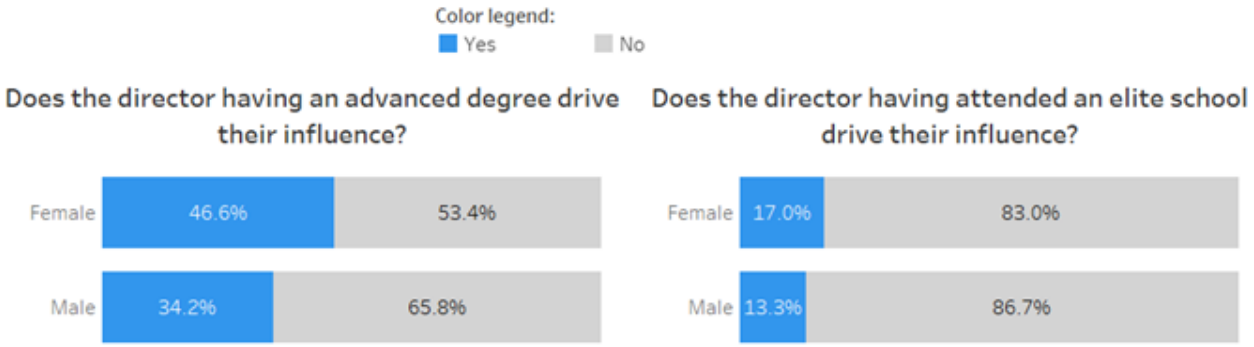

Figure 4: Education-based influence drivers

Based on the sponsor's data, we were also able to see that female directors rely on their educational background more than male directors do. There is a higher percentage of females both with advanced degrees and graduating from elite schools.

This finding is in line with other research (Dang, et al., 2014) claiming that women provide more evidence than males to be considered high achievers and female directors were more likely to hold degrees from elite schools.

Dang et al. (2014) have also found that female directors are often younger than male directors.

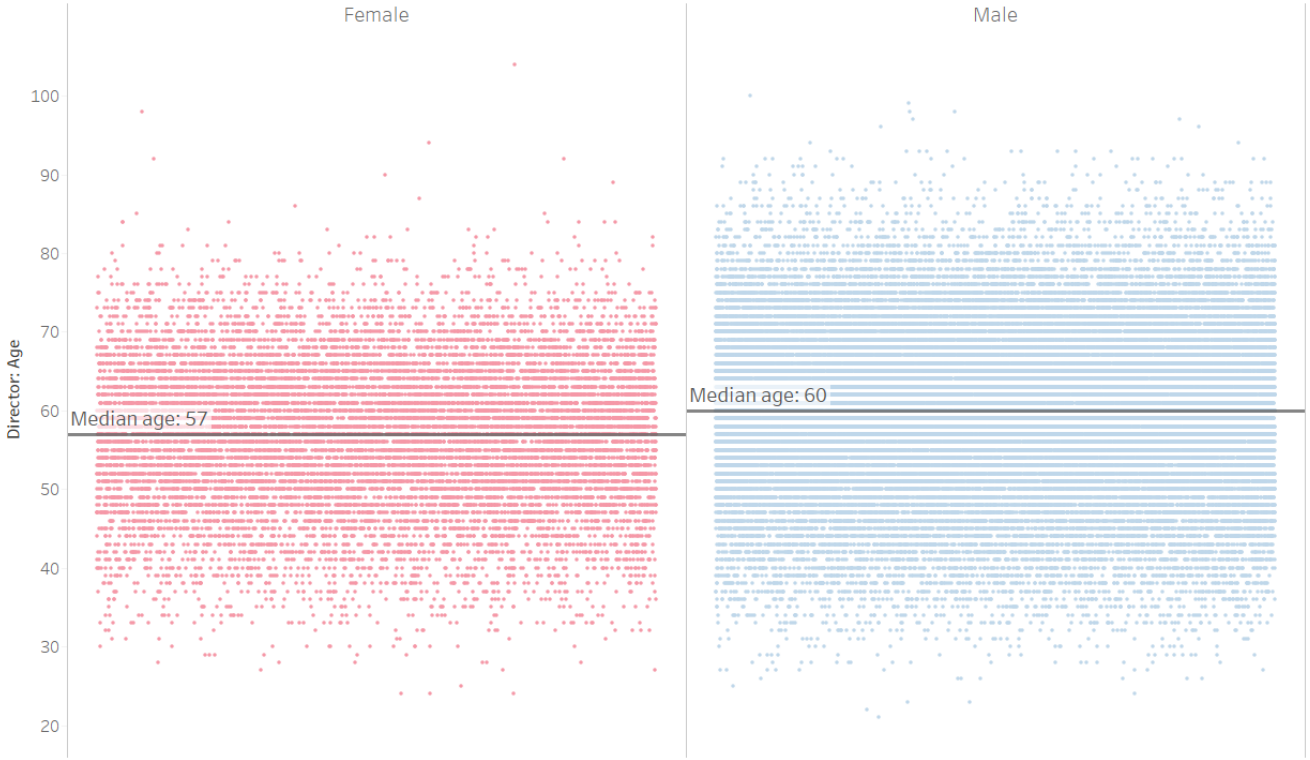

Figure 5: Female directors tend to be younger

Based on this jitter plot, male directors have a wider age range but also their median age (60) is higher than for the female directors (57).

Another hypothesis supported by Dang et al. (2014) was that women are more likely to hold non-management background.

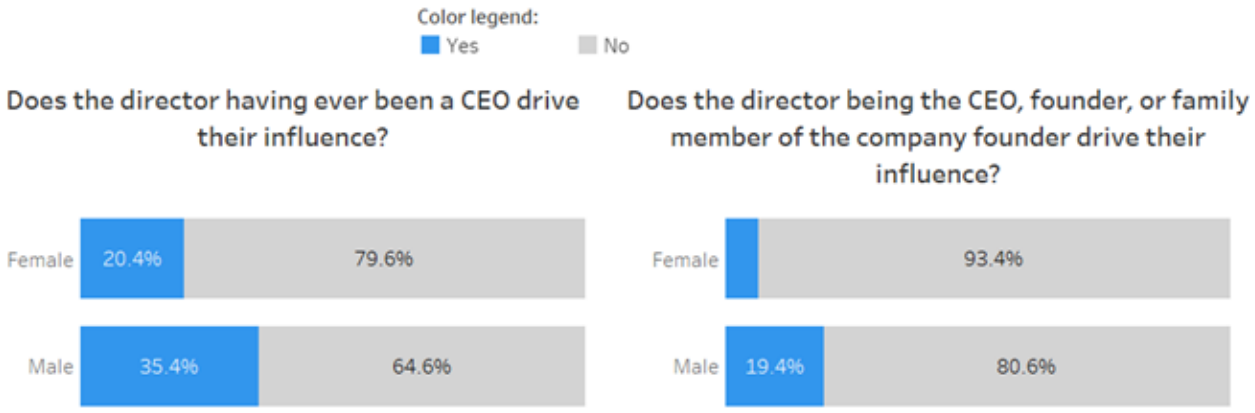

Figure 6: Influence drivers related to being a CEO or their family member

It seems that not only there is a higher proportion of CEOs among male directors but also, they are more often related to CEOs or founders of companies. This is another interesting difference between male and female directors serving on boards of various firms.

Dang et al. (2014) have found that female directors more often have family affiliation in the companies they serve on boards of, but sponsor data shows the opposite trend.

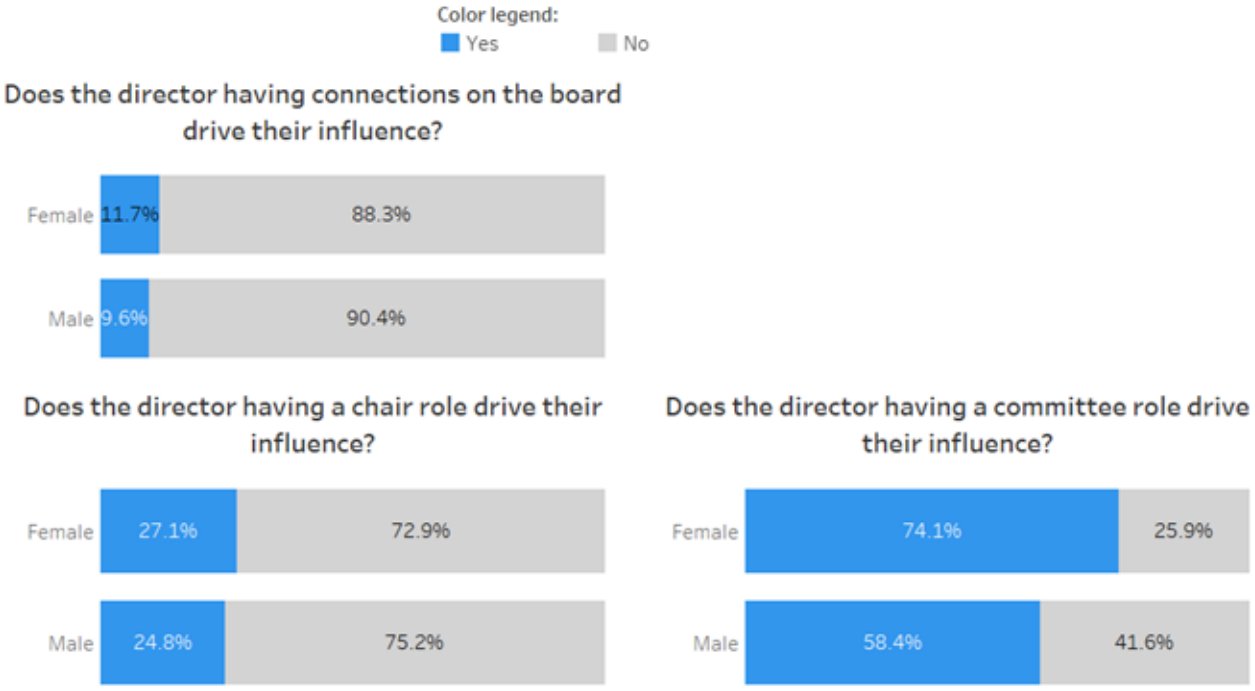

Figure 7: Influence drivers related to having a chair/committee role or connections on the board

In addition to having more connections, more women hold chair or committee roles, which makes them more influential in the companies they serve on the board of directors.

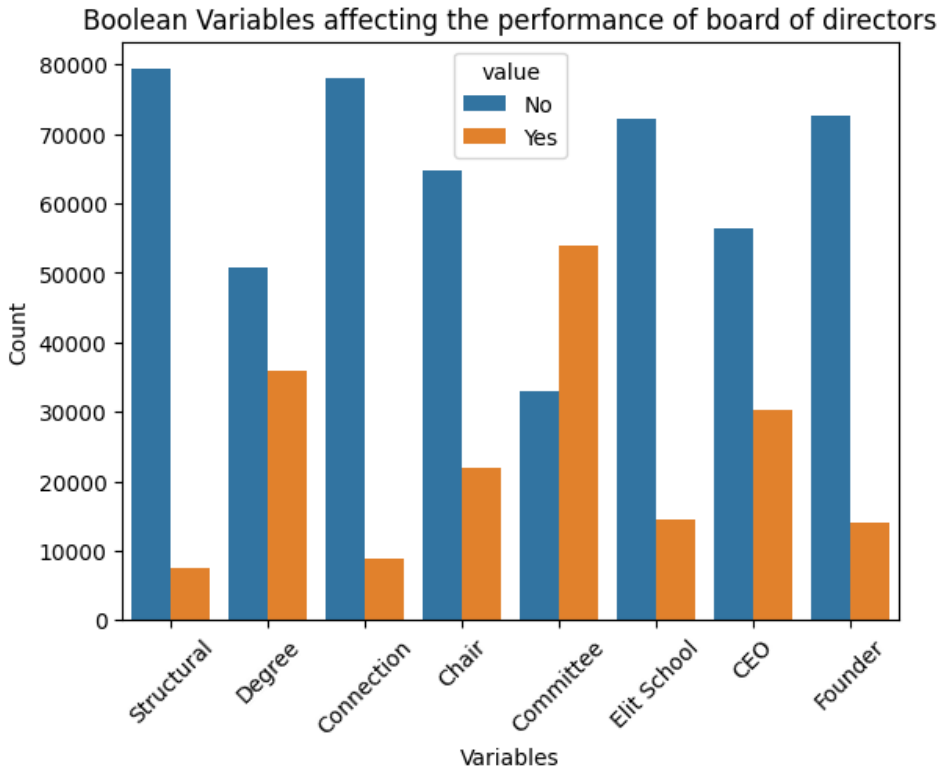

Figure 8: Summary of influence drivers of directors

Figure 8 elucidates the various Influence drivers mentioned in our dataset. The count of each boolean variable like INFLUENCE DRIVER: STRUCTURAL ADVANTAGE ~ Structural; INFLUENCE DRIVER: ADVANCED DEGREE ~ Degree; INFLUENCE DRIVER: BOARD CONNECTIONS ~ Connection; INFLUENCE DRIVER: CHAIR ROLE ~ Chair; INFLUENCE DRIVER: COMMITTEE ROLE ~ Committee; INFLUENCE DRIVER: ELITE SCHOOL ~ Elite School; INFLUENCE DRIVER: HAS BEEN CEO ~ CEO; INFLUENCE DRIVER: FOUNDER-CEO-FAMILY ~ Founder have been counted against each value to analyze the impact of each driver influence. It is found that having a structural advantage positively impacts the board of director's position. On the other hand, the board of directors having an advanced degree does not impact their position.

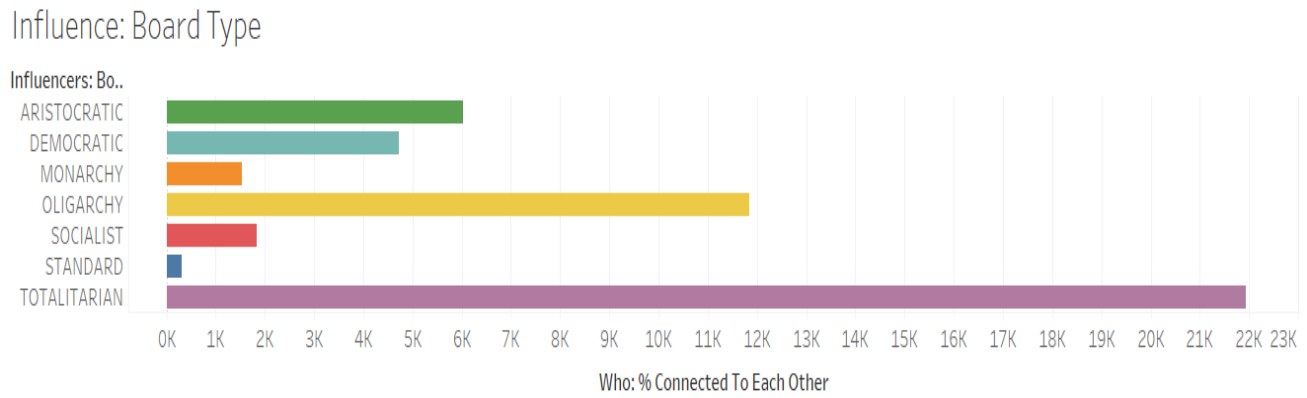

Figure 9: Influencers based on the board type

Now moving forward to board-type influence, we see that most influential people on the board firstly belong to totalitarian followed by oligarchy.

Analysis of the spread of directors included in the study based on geography can be seen in Figure 10 below. We can see that the majority of the CEOs are from US and China. The US also has the highest number of female CEOs.

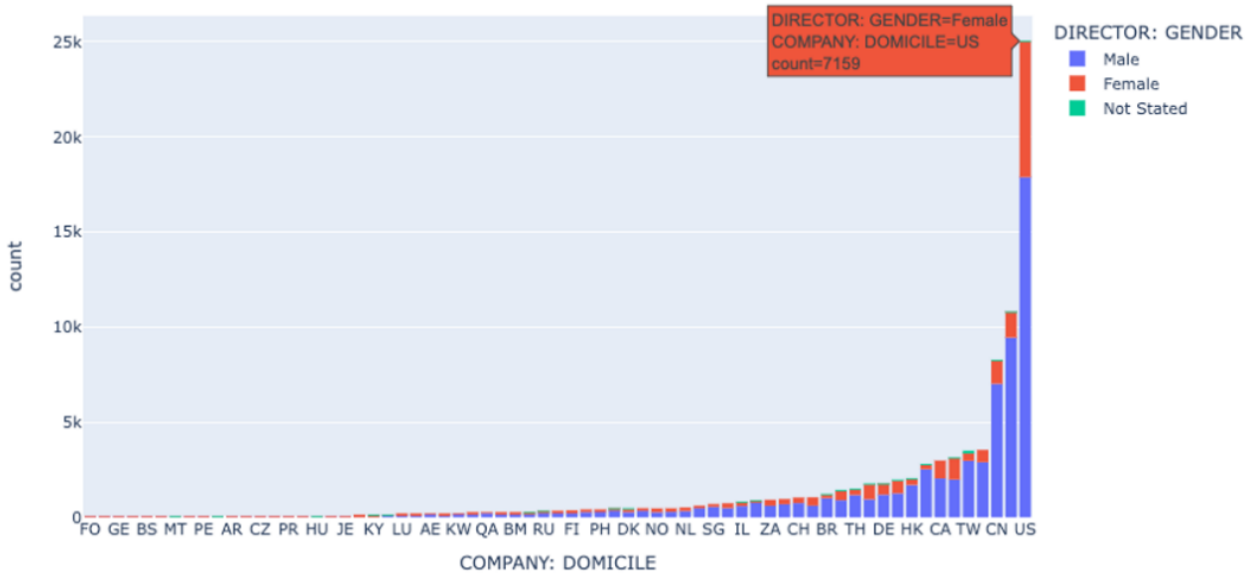

Figure 10: Histogram of Spread of Board of Directors based on domicile.

The sector of the company or the industry in which the data is spread across is also valuable to understand where the influence can be maximized and how carbon intensity can be linked to the

sector. Figure 11 below shows the spread of influential boards of directors based on the sector below. Financials have the overall highest number of female CEOs.

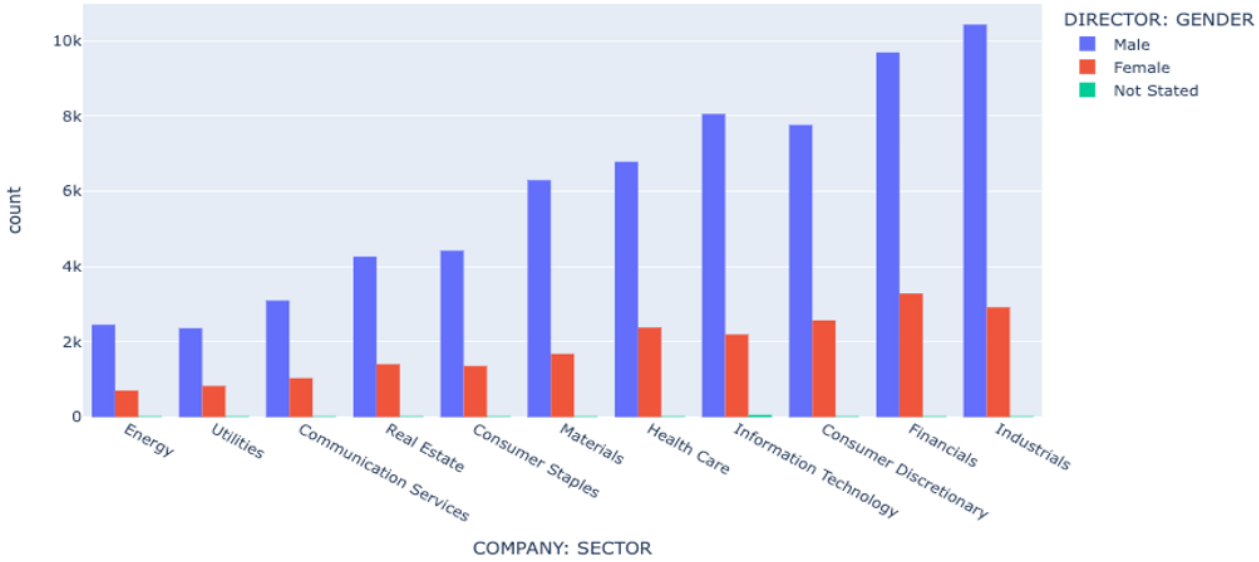

Figure 11: Grouped bar chart of count of board of directors by sector.

Figure 12: The total influence of active directors identified as female (green) versus the total influence of active directors identified as female minus the total percentage of active female directors (red).

Due to my background in business intelligence and data visualization, I concentrated on using Tableau to analyze the provided dataset. According to Smith, J. A., & Lee, K. H, if charts are constructed properly, it is considerably simpler for people to comprehend information displayed graphically than it is to read a large table of raw data. I tend to agree; it often takes me a few seconds to spot a trend on a chart, however, it can be challenging to infer a narrative from a collection of statistics. To make the data easier to understand, I concentrated on aggregating it and presenting it as summary statistics or graphs.

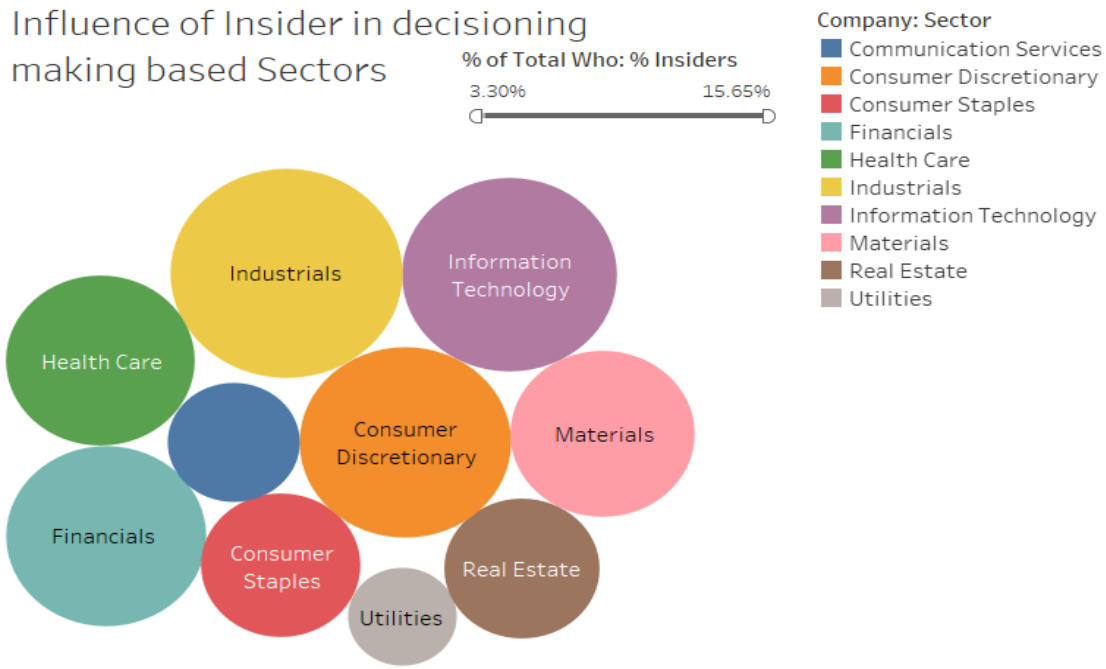

Figure 13: Influence of insiders in different sectors

We were interested in checking if the sponsor's statistics showed comparable patterns, so we compared my findings with academic literature. We will use data visualizations and summary statistics based on the dataset given by Free Float Media to address inquiries about the traits of directors.

To understand the board of directors' mindset it is important to figure out the insiders' influence on decision-making. We have considered the sector in which insiders play a key role in making decisions. It can be seen in Industrial trials, Information technology, Consumer discretionary etc. These sectors are heavily influenced by insiders.

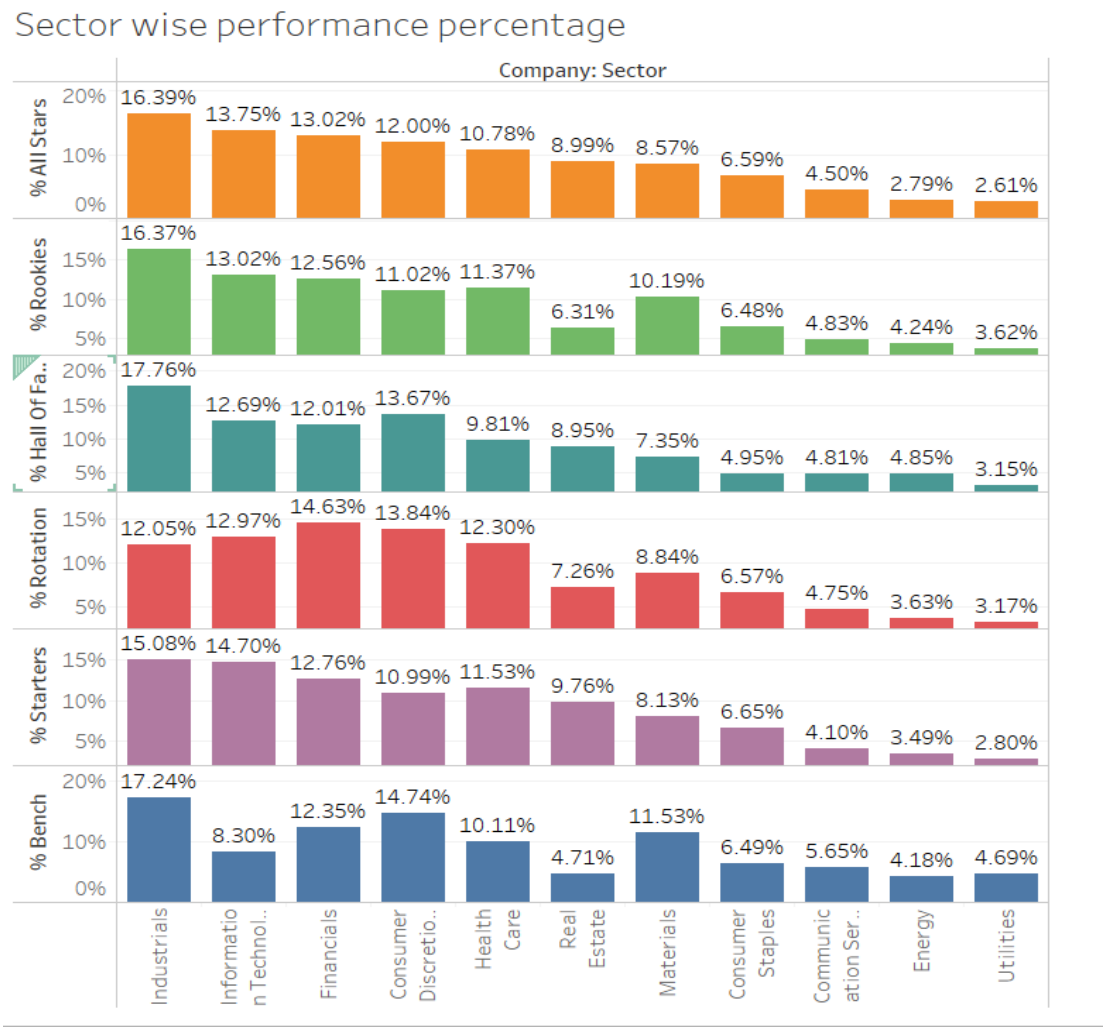

Figure 14: Sector-wise performance

Figure 14 shows the performance percentage of different sectors of the company. The above bar graph shows that hall of fame and bench-based people have better performance percentages in the industrial sector compared to other sectors.

It is interesting to see that the average age of the directors is 59 and are active on more than 1 board currently. With over 30 years of more than 10% influence in all the boards they were involved in, the mean influence at the highest point of their career in a company is 13%. In addition, the cross-tabulation table shown in Figure 15 below signifies that more women with prior CEO experience are considered for the board of directors whereas that is not the case in terms of male board of directors.

Cross-tabulation Table:

RESUME: EXPERIENCE DIRECTOR: GENDER	No	Yes
Female	12512	3196
Male	36508	19886
Not Stated	166	3

Figure 15: Cross-tabulation table of resume experience and director gender.

The skewness and kurtosis table below in Figure 16 provides a better understanding of the distribution of data within the director core metrics table. It is very evident that the average number of second-degree connections the director must board members at boards on which they sat, Resume: Average Board Connects, and the influence times market capitalization for the director and their first- and second-degree network connections, Resume: Network Power have high positive skewness and high kurtosis values. Positive skewness in a distribution indicates that the right side's tail is longer or fatter than the left side's tail. It implies that the data is more concentrated to the left and that the extreme values are located to the right. In contrast to the normal distribution, a distribution with a strong positive kurtosis (leptokurtic) features a sharp peak and thick tails. It suggests that the data is more concentrated around the mean and that the tails contain more outliers. This indicates that most of the directors have either slightly more than 1 board connects or higher numbers of 50-100 connections. This creates the skewness in the data. Same situation with the network power as well. This might also indicate a further investigation on the validity of the data by the sponsors.

Skewness and Kurtosis Table:

	Skewness	Kurtosis
DIRECTOR: ID	-0.957866	-0.710428
DIRECTOR: AGE	-0.097601	0.076950
DIRECTOR: ACTIVE BOARDS	3.345595	13.915101
DIRECTOR: BOARD HISTORY	3.078131	11.927664
RESUME: AVERAGE BOARD CONNECTS	10.715946	155.265761
RESUME: NETWORK POWER	10.477133	186.736669
INFLUENCE: % OF YEARS AS INFLUENCER	0.825067	-1.217994
INFLUENCE: HIGHEST INFLUENCE	2.370007	6.532678
INFLUENCE: HIGHEST INFLUENCE YEAR	-1.073644	0.276634
INFLUENCE: 5YR MEDIAN	2.722738	8.738247
PERFORMANCE: WIN RATE	-0.300472	-1.250160

Figure 16: Skewness and kurtosis table of director core metrics.

In understanding the controversy win rate of the board of directors for the companies in the study, the cross-tabulation shown from the director-company core metrics data in Figure 17 below showcases that almost 62% of female directors have better win rates than 52% of male directors. This showcases that women directors are more prone to talk about current controversial issues that ties up the corporate social responsibility of the company.

Cross-tabulation Table:

PERFORMANCE: CONTROVERSY WIN RATE	0.0	1.0
DIRECTOR: GENDER		
Female	7849	12676
Male	20596	45410
Not Stated	161	12

Figure 17: Cross-tabulation of controversy win rate by gender of the director.

Sector wise average age of directors

	Company: Domicile									
Company: Sector	AU	CA	CN	GB	HK	IN	JP	KR	TW	US
Consumer Discretio..	60.30	58.75	53.25	56.94	58.86	59.81	60.79	56.62	63.82	59.75
Financials	61.59	62.12	53.90	58.66	58.53	61.13	61.10	59.71	60.23	62.35
Health Care	63.00	58.67	53.68	59.22	58.36	61.03	61.23	56.97	66.56	60.64
Industrials	62.10	62.77	54.17	58.75	57.69	59.35	63.34	59.22	63.31	61.65
Information Technol..	58.43	58.33	52.93	58.28	56.06	60.13	60.66	57.57	63.46	59.40

Figure 18: Average Age of directors in different company sectors.

Looking at the above table, we see that based on domicile and sector the average to become a director is in the range of 53 to 63. The youngest director is of age 53 in the domicile of Canada.

To investigate the hypothesis of higher education and experience for female directors, we relied on the five features mentioned in Figure 19 below. Resume: Experience indicates if the director is or was a CEO at any company, Resume: Smarts shows if the director has attended an elite school or has an advanced degree, The first count under the Director: Gender is the total count of directors that fall under the category, the second count is the average number of second-degree connections the director has to board members at boards on which they sat, and the final count is the influence times market capitalization for the director and their first and second-degree network connections. We can see from the table that there is a higher number of highly educated, and higher experience female members than male members with no degree or experience. The influence as well as market capitalization is higher in the females with higher education and experience at executive levels.

		Director: Gender	
Resume: Experience	Resume: Smarts	Female	Male
No	No	6,933 0.2032 810,139,188,085	25,239 0.1803 502,590,988,502
	Yes	5,579 0.6707 2,253,519,728,706	11,269 0.7057 1,518,210,124,171
Yes	No	1,101 0.5504 2,321,549,980,619	10,517 0.4850 1,277,384,382,360
	Yes	2,095 0.9976 4,547,428,552,087	9,369 1.0139 2,787,493,505,964

Figure 19: Table showing different characteristics of male and female directors

.

Nielsen & Huse (2010) stated that there are differences in the leadership style between male and female directors, and those differences might affect the board's dynamics and effectiveness. Previous research concluded that it is difficult to establish a direct relationship between the board composition and the performance of a firm. It has been said that the strengths of male and female directors differ and while one gender might be good at performing certain tasks, the other gender will likely have stronger skills in a different domain.

To answer this question, we added a simple calculation grouping firms based on the percentage of female directors. Then, we compared the average performance metrics provided by the sponsor for each group of firms.

% of female directors	Avg. Performance: Influence Weighted Carbon Win Rate	Avg. Performance: Influence Weighted Controversy Win Rate	Avg. Performance: Influence Weighted Ebitda Win Rate	Avg. Performance: Influence Weighted Tsr Win Rate	Avg. Performance: Influence Weighted Win Rate
Under 10%	20.6%	64.3%	31.3%	17.6%	50.2%
10-20%	26.5%	64.8%	35.1%	25.5%	50.7%
21-30%	32.2%	68.3%	36.3%	33.5%	52.2%
31-40%	37.5%	66.7%	37.5%	35.1%	51.7%
41-50%	38.5%	65.8%	37.0%	31.5%	52.7%
51-60%	38.7%	67.2%	37.8%	33.8%	54.3%
61-70%	36.7%	71.3%	38.8%	32.9%	53.7%
Over 70%	37.8%	69.0%	30.8%	37.4%	51.8%

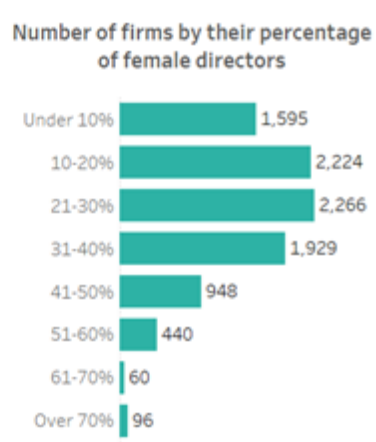

Figure 20: Table showing win rates based on the percentage of female directors in a firm.

There are very few firms with over 60% of female directors. Most boards of directors are male-dominated and have under 30% of female directors.

However, it is also interesting to note that based on the average performance metrics, firms with higher percentages of females on boards of directors seem to be performing better across all 5 metrics.

Welsch's T-test to understand significant difference between the performance of male and female-dominated boards

Based on Welsch's T-test, there is a statistically significant difference between the performance of male and female-dominated boards, based on the variable "PERFORMANCE: INFLUENCE WEIGHTED WIN RATE" from the company data.

```
print(stats.ttest_ind(female_dom, male_dom, equal_var = False))
```

```
Ttest_indResult(statistic=3.119883707408206, pvalue=0.0018833130277899916)
```

Therefore, it is worth exploring how the board's diversity and female representation affects a company's performance.

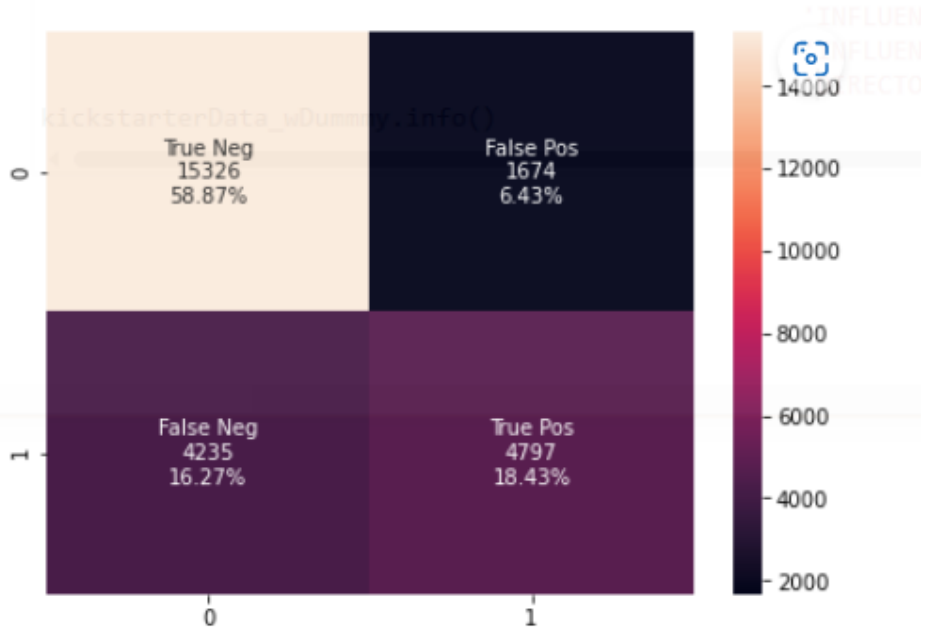

Figure 21: Confusion matrix considering test and predicted data.

Our hypothesis considered that having a higher education or degree is important to join a board of directors but it's clear that it's not important to have a higher education degree to become a director. True Positive (TP): The model predicted a positive class, and it was actually positive. In this case, the model correctly predicted 15326 positive cases. False Positive (FP): The model predicted a positive class, but it was actually negative. In this case, the model predicted 4235 positive cases, but they were actually negative. False Negative (FN): The model predicted a negative class, but it was actually positive. In this case, the model predicted 1674 negatives but they were actually

positive. True Negative (TN): The model predicted a negative class. In this case, the model correctly predicted 4797 negative cases. Using these values, we can calculate different evaluation metrics such as accuracy, precision, recall, and F1 score to evaluate the performance of the classification model.

Problem statement Understanding the relationship between gender and the composition of the board. The board is composed of Age, Active Boards, and Board History.

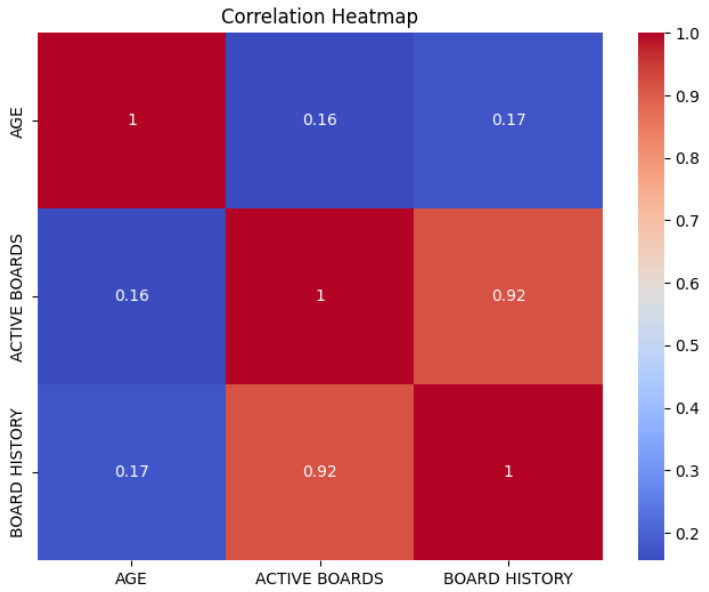

Figure 26: Correlation Matrix for Composition of Board

Algorithm	***Model Accuracy***	***Precision***	***Recall***	***F1 Score***
Decision Tree	78%	74.1%	78.8%	72.3%
SVM	78.6%	61.8%	78.6%	69.2%
Random Forest	78.8%	74.16%	78.8%	72%
Logistic Regression	78.3%	63.8%	78.3%	69.1%

Table 5: Benchmarking metrics for the problem statement 3

We wanted to evaluate the relationship between gender and the composition of the board. The board is composed of Age, Active Boards and Board History. The correlation matrix was used asa a key tool to evaluate the correlation between the variables Age, Active Boards, and Board History as shown in Figure 26. It was observed that Board History and Active Boards were highly

correlated. The ultimate goal was to understand how gender affects the board. We built several models to understand this relationship, including Support Vector Machine, Logisitic Regression, Decision Tree, and Random Forest. Cross-validation was performed to handle the imbalance in the dataset. All the models exhibit similar characteristics. The accuracy of all models is similar and falls in the range of 78% to 78.8%. Although the precision varies. Decision Tree and Logisitic Regression are the best models for our use case scenario.

Problem statement Potential for Becoming a Director Overview

The features 'INFLUENCE DRIVER: ADVANCED DEGREE, INFLUENCE DRIVER: ELITE SCHOOL' contribute the most to the target value, which is Has_Been_CEO.

S.no	Algorithm	Accuracy	Precision	Recall	F1-Score
1	Logistic Regression	77%	78%	90%	84%
2	Decision Tree	75%	78%	86%	82%
3	SVM	76%	75%	95%	84%

Table 1: Comparison of algorithms

The best algorithm for determining the potential for becoming a director is Logistic Regression. When the class distributions are extremely skewed, accuracy might lose its usefulness as a measure of model performance. Not just accuracy percentage is better but its precision, recall and F1-score support this. The dataset being so sparse in nature this is bring the best outcome after using feature engineering.

Problem Statement Performance: Controversy Win Rate

Analyzing the performance controversy win rate on the decision-making of the board members. As this is one of the most important and influential factors for my individual and group project. Here, I want to divide my test and train in 70-30 manner to get the stats. Before starting with the modeling based on the dataset. I have taken into consideration the factors that are termed as "YES" such as "INFLUENCE DRIVER: ADVANCED DEGREE_YES", "INFLUENCE DRIVER: ELITE SCHOOL_YES" etc. I have sorted the "PERFORMANCE: WIN RATE" and labeled all the factors above 0.5 as 1 rating as a success for impact on the win rate and the rest as 0. All the "unrated" marked in performance: controversy win rate was converted to "0". After converting

features to the required dummy variables all the relevant factors were ready to start with my further analysis.

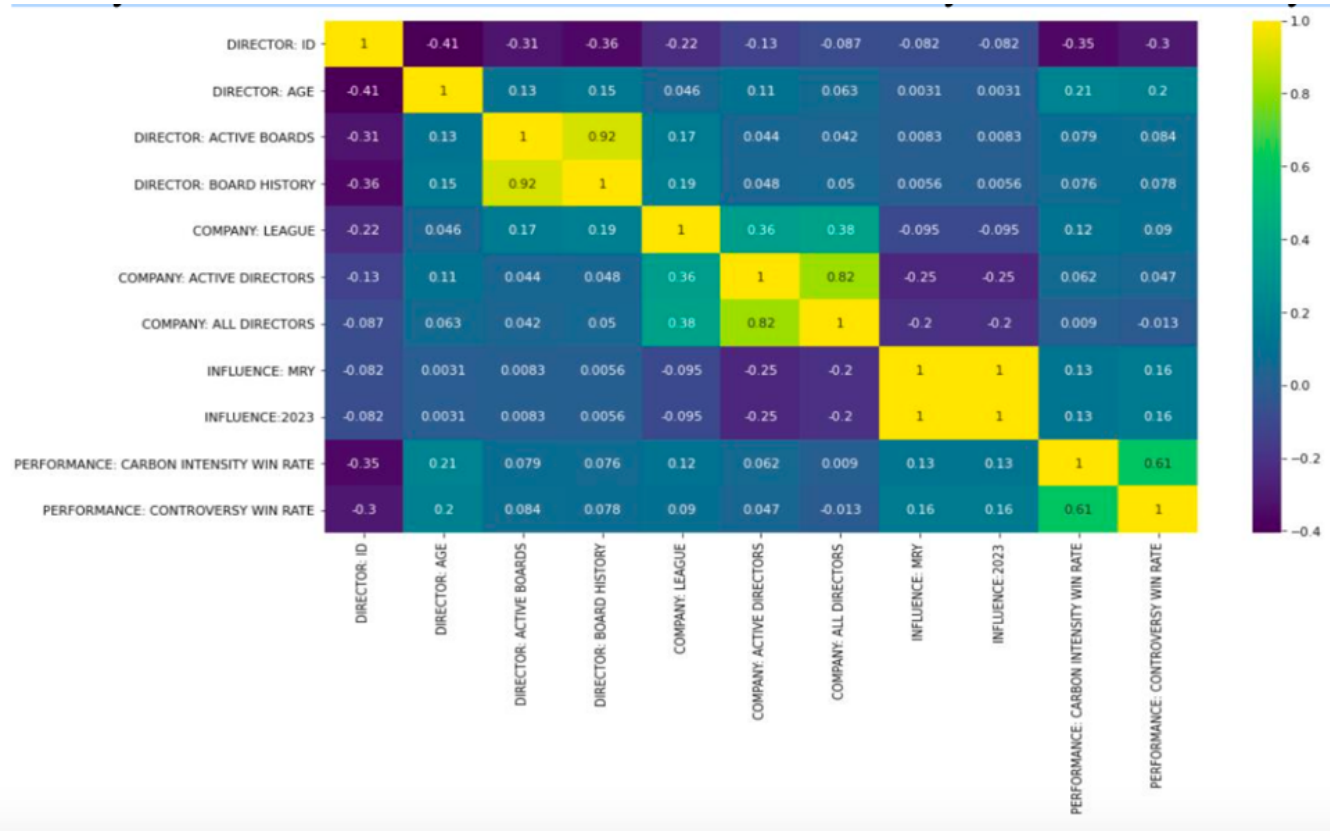

Figure 22: Correlation matrix

Looking at the correlation matrix, we found that the most important parameter for this model is "**carbon intensity win rate" and "age".** To understand deep inside it. I have considered understanding the "yes" and the "no" ratio of the impacted factors. Parameter with "No" label in *'DIRECTOR: GENDER','INFLUENCE DRIVER: ADVANCED DEGREE','INFLUENCE DRIVER: ELITE SCHOOL', 'INFLUENCE DRIVER: HAS BEEN CEO', 'INFLUENCE DRIVER: FOUNDER-CEO- FAMILY', 'INFLUENCE DRIVER: CHAIR ROLE','INFLUENCE DRIVER: COMMITTEE ROLE','INFLUENCE DRIVER: BOARD CONNECTIONS','INFLUENCE DRIVER: STRUCTURAL ADVANTAGE'*. Therefore, I have removed all the "No" parameter and considered only the "Yes" marked factor while moving forward.

The figure below shows as stated above by the correlation matrix. We have removed all the negative correlations on the target variable "**performance controversy win rate"**.

We will be implementing three models to get a conclusive decision on this factor. First will be Logistic regression then SVM and lastly KNN. To get a good understanding on our analysis We will be not just looking at accuracy but precision, recall and F1-score too.

Here, we have compared all the model and will be focusing on the F1score as this will consider the precision and recall factor to compare and choose the best model.

S.no	Algorithm	Accuracy	Precision	Recall	F1-Score
1	Logistic Regression	93%	88%	92%	90%
2	SVM	93%	87%	93%	89%
3	KNN	82%	75%	68%	71%

Table 2: Comparison of algorithms

Performance controversy win rate is one of the most important factors to decide the influence of decision making on the board table. Understanding the insider effect /age effect on the board table creates questions doubt. So, the potential for becoming part of board of directors a factor reflected here.

Problem statement Understanding the factors affecting the influence of the directors

The process followed to prepare the data from exploratory data analysis to model building were to clean model, one hot encoding and check for multicollinearity. This is the step where correlation between the predictor variables is quantified using variance inflation factor and helps get rid of unnecessary attributes to improve the model performance. After careful consideration, the independent variable included in the analysis are age, director active boards, average board connects, highest influence, 5-year median influence, win rate, elite school education, and experience as a CEO to understand the factors affecting the influence of a directors. A correlation analysis was also conducted on the dataset to determine nature of relationship of various variables with the predicting variable Influence as seen in Figure 10.

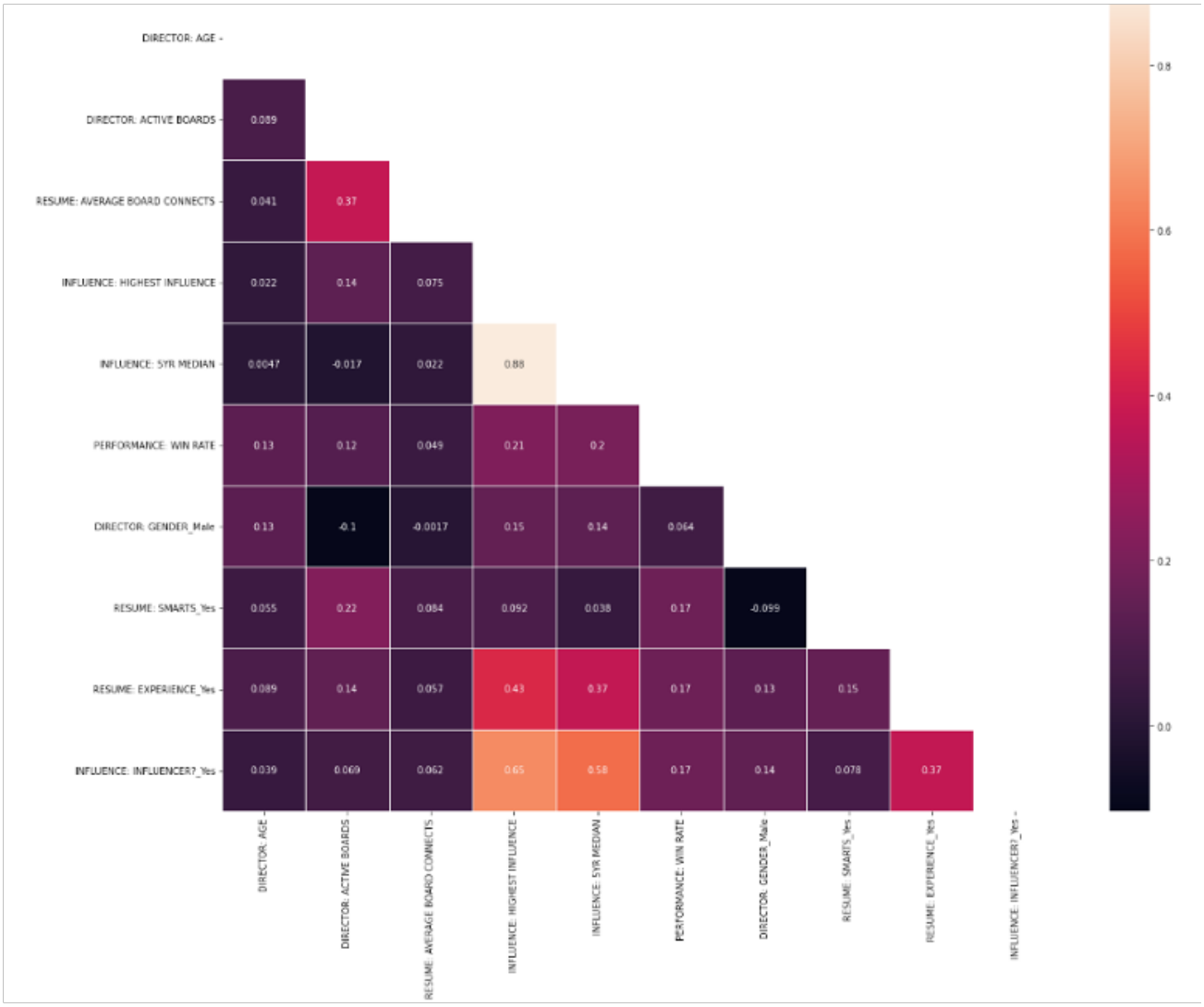

Figure 23: Correlation plot of contributing factors to determining influence of the director. We can see that prior experience and experience as a CEO can contribute positively.

Algorithm	***Model Accuracy***	***Precision***	***Recall***	***MSE***
Logistic Regression	87.5%	86.9%	72%	0.125
Decision Tree	88.4%	80.6%	84%	0.116
Random Forest	87.4%	80.7%	80%	0.126
KNN	87.15%	82%	78%	0.124

Table 3: Benchmarking metrics for the problem statement 1

We implemented 4 different classification models with the results as shown above. Though Decision Tree showed the best accuracy, Logistic Regression seems to work best when considering precision and recall as the data is highly imbalanced. Accuracy can lose its validity as a gauge of model performance when the class distributions are highly skewed. The Logistic Regression results are shown in detail below with Figure 11: Logistic regression results for problem statement 1: Understanding the factors affecting the influence of the directors. The logistic regression model output displays significant variables which can be determined by the Pr(>|t|), p-value, and depicts the positive and negative impact over dependent variable price through the coefficients. The Significance of the variable is determined if the P-Value is less than 0.05, a significance level of 0.05 indicates a 5% risk, then it terms that the model fits the data well. We can see that higher influence experience that is more than 3 years of directorship, and prior experience as a CEO showcase the highest contributing factors to being an influential director. We can also see that these results are in coherence with the correlation analysis done and correlation matrix shown in Appendix Figure 2.

Results: Logit

Model:	Logit	Pseudo R-squared:	inf
Dependent Variable:	INFLUENCE: INFLUENCER?_Yes	AIC:	inf
Date:	2023-05-22 11:26	BIC:	inf
No. Observations:	51023	Log-Likelihood:	-inf
Df Model:	8	LL-Null:	0.0000
Df Residuals:	51014	LLR p-value:	1.0000
Converged:	1.0000	Scale:	1.0000
No. Iterations:	8.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
DIRECTOR: AGE	0.1100	0.0154	7.1294	0.0000	0.0798	0.1403
DIRECTOR: ACTIVE BOARDS	-0.1394	0.0159	-8.7719	0.0000	-0.1705	-0.1082
RESUME: AVERAGE BOARD CONNECTS	0.0733	0.0152	4.8388	0.0000	0.0436	0.1030
INFLUENCE: HIGHEST INFLUENCE	2.0182	0.0458	44.0945	0.0000	1.9285	2.1079
INFLUENCE: 5YR MEDIAN	3.3684	0.0598	56.3293	0.0000	3.2512	3.4856
PERFORMANCE: WIN RATE	-0.1473	0.0157	-9.3547	0.0000	-0.1782	-0.1165
DIRECTOR: GENDER_Male	0.0123	0.0151	0.8157	0.4147	-0.0173	0.0420
RESUME: SMARTS_Yes	-0.0679	0.0154	-4.4197	0.0000	-0.0980	-0.0378
RESUME: EXPERIENCE_Yes	0.1247	0.0154	8.1054	0.0000	0.0945	0.1548

speed: 0:20:02.339774

Figure 24: Results of the logistic regression

Problem statement Understanding the factors affecting the total Carbon Intensity wins over losses versus sector and market capitalization sized peers attributed to the director

Algorithm	***Model Accuracy***	***Precision***	***Recall***	***MSE***
Logistic Regression	77.1%	67.7%	97.1%	0.229
Decision Tree	77.2%	67.6%	97.7%	0.228
Random Forest	72%	69.2%	71.6%	0.280
KNN	73.97%	69%	79%	0.282

Table 4: Benchmarking metrics for Problem Statement 2 - Understanding the factors affecting the total Carbon Intensity wins over losses versus sector and market capitalization sized peers attributed to the director.

From the above Table 4 benchmarking metrics, I have chosen Decision tree for this problem statement to provide influencing factors for carbon intensity win rate for the directors. The tree shown in the below figure 8 shows the tree model created in the process. The tree signifies that the two main contributing factors to determining carbon intensity win rates are controversy win rate of the directors and percentage of influence in the latest year of 2023 of each director in their respective companies in accordance with the correlation analysis shown in Figure 3, Appendix and confusion matrix as shown in Figure 4, Appendix.

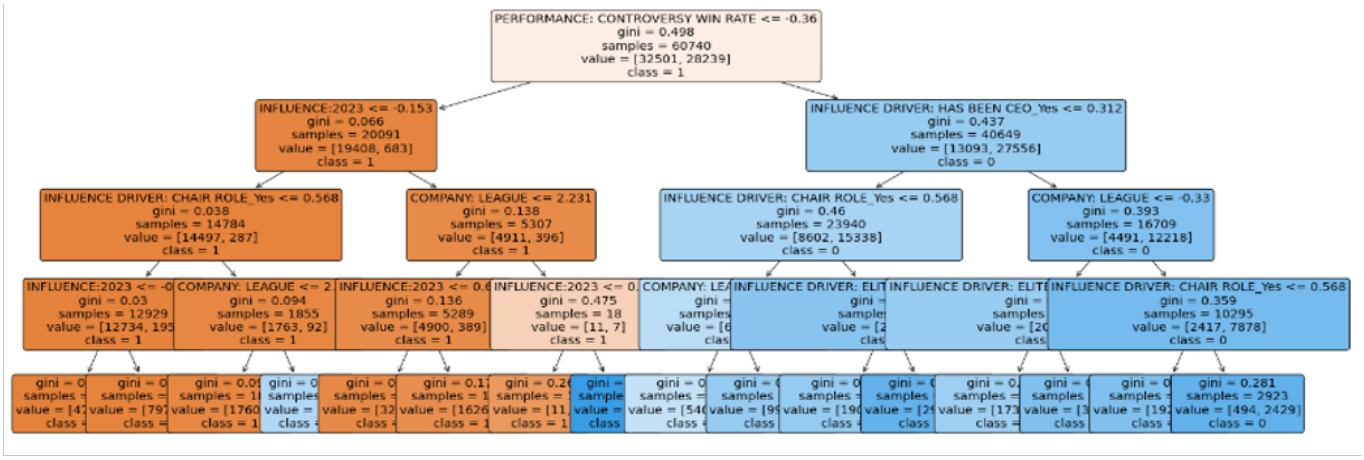

Figure 25: Decision Tree model results for the problem statement 2. Understanding the factors affecting the total Carbon Intensity wins over losses versus sector and market capitalization sized peers attributed to the director.

Conclusion

Through our primary business questions, we have tried to analyze gender diversity, prior experiences of directors, their influence, and finally leading up to understanding carbon intensity win rates. We found out that gender bias plays a vital role in understanding the potential of joining a board of directors. Females are required to have more advanced and elite degrees to be part of a board of directors. However, most directors who are male do not have it. Therefore, having an advanced degree was not significant in our predictive models, only because of the gender imbalance among directors.

The controversy win rate helps us understand that age is an important factor in the decision-making of the boards. Greater experience translates to increased influence and decision-making power thus a higher positive impact on the carbon intensity win rates.

Based on the analysis results, the model demonstrates moderate performance in classifying instances into the "No" and "Yes" categories. The precision, recall, and f1-score metrics indicate that the model is better at predicting instances belonging to the "No" category compared to the "Yes" category. This states that Has_been_director does not depend on the education as the factor.

Consolidating multiple research topics from various team members into a single cohesive proposal and resolving any differences was our main task. Moving forward we have considered how implementation of performance: controversial win rate being an important factor in decision making of the board of directors. This fact connects the dots to make us understand the potential of becoming board of director has other than just the influence. The higher the controversy win rate, the more the decision to be in favor of the experienced candidate than others.

Based on the models we have built in this project we realized that gender plays a vital role in the composition of the board. We were able to identify hidden patterns in the dataset provided to us by our sponsor Free Flot Media. Additionally, it is important to note that our basis of this claim is supported by the Decision Tree and Logistic Regression model that we have built in this project.

To summarize everything in a nutshell, we have taken influence, performance and gender as major in understanding the potential of becoming a board member and what influences the decision making in an organization. While conducting feature engineering in both the cases we understand the primary factor that we could consider while performing our analysis. In the first part we

considered HAS_BEEN_DIRECTOR as factor and second one took performance metric of controversy win rate.

The higher the controversy win rate, the more the chance of a decision falling in the insider circle of the board of directors. And to become director based on gender, we concluded that if the director is female then higher education is a most important factor else it does not impact your chance. Moreover, majority of the director are male. Therefore, as my model conformed it does not matter if you have an elite degree.

References:

Adams, R. B., & Ferreira, D. (2009). *Women in the boardroom and their impact on governance and performance.* Journal of Financial Economics, 94(2), 291-309.

Albawwat, A. H., Almansour, A., Al Zobi, M. K., & Alrawashedh, N. H. (2020). The Effect of Board of Directors and Audit Committee Characteristics on Company Performance in Jordan. International Journal of Financial Research, 11(6), 10. https://doi.org/10.5430/ijfr.v11n6p10

Burgess, Z., & Tharenou, P. (2002). Women board directors: Characteristics of the few. *Journal of business ethics*, *37*, 39-49.

Chams, N., & García-Blandón, J. (2019). Sustainable or not sustainable? The role of the board of directors. *Journal of cleaner production*, *226*, 1067-1081.

Dang, R., Bender, A. F., & Scotto, M. J. (2014). Women on French Corporate Board of Directors: How do they differ from their male counterparts? *Journal of Applied Business Research (JABR)*, *30*(2), 489-508.

Data Governance Capability Map in 2023 | Roadmap, Data architecture, Data. (2023). Pinterest. https://www.pinterest.com.au/pin/data-governance-capability-map-in-2023--488781365820690783/

Data Management Roadmap Example: A Strategic Approach. (2023). Jibility. https://www.jibility.com/data-governance-strategic-roadmap/

Failure of Classification Accuracy for Imbalanced Class Distributions - MachineLearningMastery.com. (2021). MachineLearningMastery.com. https://machinelearningmastery.com/failure-of-accuracy-for-imbalanced-class-distributions/

Glass, C., Cook, A., & Ingersoll, A. R. (2015). Do Women Leaders Promote Sustainability?Analyzing the Effect of Corporate Governance Composition on Environmental Performance. Business Strategy and the Environment, 25(7), 495–511. https://doi.org/10.1002/bse.1879

Hillman, A. J., Cannella Jr, A. A., & Harris, I. C. (2002). Women and racial minorities in the boardroom: How do directors differ? *Journal of Management*, *28*(6), 747-763.

How to Evaluate Your Company's Carbon Risk. (2022). Harvard Business Review. https://hbr.org/2022/01/how-to-evaluate-your-companys-carbon-risk

Jones, B. (2020). Learning to See Data: How to Interpret the Visual Language of Charts. *Data Literacy Press;* 1st edition.

Martínez-García, I., Terjesen, S., & Gómez-Ansón, S. (2021). Board Gender DiversityCodes, Quotas and Threats of Supranational Legislation: Impact on Director Characteristics and Corporate Outcomes. British Journal of Management. https://doi.org/10.1111/1467-8551.12517

Nielsen, S., & Huse, M. (2010). The Contribution of Women on Boards of Directors: Going beyond the Surface. *Corporate Governance: An International Review*, *18*(2), 136–148.

Python for Data Science. Independent T-test. Retrieved from: https://www.pythonfordatascience.org/independent-samples-t-test-python/

Sakai, T. (2016). Two sample t-tests for IR evaluation: Student or Welch?. *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 1045-1048.

Saha, R. (2023). The impact of board-level female directors on firm performance: evidence from India. Equality, Diversity and Inclusion: An International Journal. https://doi.org/10.1108/edi-07-2022-0172

The Board's Role in Sustainability. (2020). Harvard Business Review. https://hbr.org/2020/09/the-boards-role-in-sustainability

Women on Boards: The Hidden Environmental Connection? (2023). Powering better investment decisions - MSCI. https://www.msci.com/www/blog-posts/women-on-boards-the-hidden/02361079223

Xia, M., Cai, H. H., & Yuan, Q. (2022). Can board climate-responsible orientation improve corporate carbon performance? The moderating role of board carbon awareness and firm reputation. Frontiers in Environmental Science, 10. https://doi.org/10.3389/fenvs.2022.1043325

Appendix

Descriptive Statistics Table:

	DIRECTOR: ID	DIRECTOR: AGE	DIRECTOR: ACTIVE BOARDS
count	72891.000000	65266.000000	72891.000000
mean	501881.346531	59.291377	1.192959
std	217743.856068	9.738864	0.526497
min	10027.000000	21.000000	1.000000
25%	236902.000000	53.000000	1.000000
50%	597762.000000	59.000000	1.000000
75%	666716.500000	66.000000	1.000000
max	727365.000000	104.000000	8.000000

	DIRECTOR: BOARD HISTORY	RESUME: AVERAGE BOARD CONNECTS
count	72891.000000	72388.000000
mean	1.286894	0.484445
std	0.684065	3.229220
min	1.000000	0.000000
25%	1.000000	0.000000
50%	1.000000	0.000000
75%	1.000000	0.000000
max	9.000000	100.000000

	RESUME: NETWORK POWER	INFLUENCE: % OF YEARS AS INFLUENCER
count	7.289100e+04	72891.000000
mean	1.380317e+12	30.784377
std	4.786638e+12	44.063017
min	0.000000e+00	0.000000
25%	2.742527e+08	0.000000
50%	7.009669e+10	0.000000
75%	7.621955e+11	100.000000
max	1.908540e+14	100.000000

	INFLUENCE: HIGHEST INFLUENCE	INFLUENCE: HIGHEST INFLUENCE YEAR
count	72891.000000	72891.000000
mean	13.061603	2021.913652
std	14.457380	1.351678
min	0.000000	2018.000000
25%	4.100000	2021.000000
50%	8.200000	2023.000000
75%	16.400000	2023.000000
max	100.000000	2023.000000

	INFLUENCE: 5YR MEDIAN	PERFORMANCE: WIN RATE
count	56772.000000	72891.000000
mean	11.597429	0.440134
std	13.005756	0.313590
min	0.000000	0.000000
25%	4.000000	0.000000
50%	7.000000	0.540000
75%	14.000000	0.679000
max	100.000000	1.000000

Figure 1: Director descriptive statistics summary

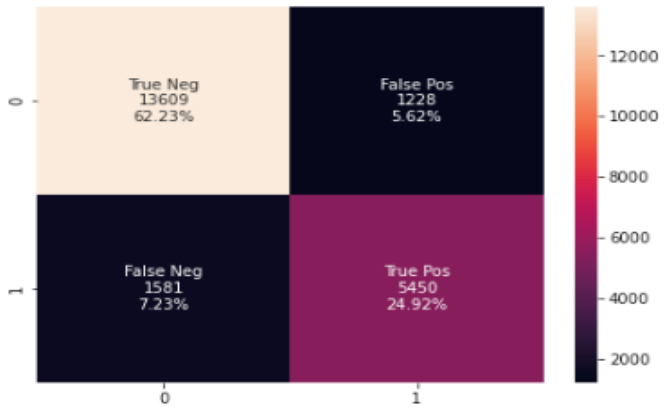

Figure 2: Confusion matrix of logistic regression model for the problem statement Understanding the factors affecting the influence of the directors.

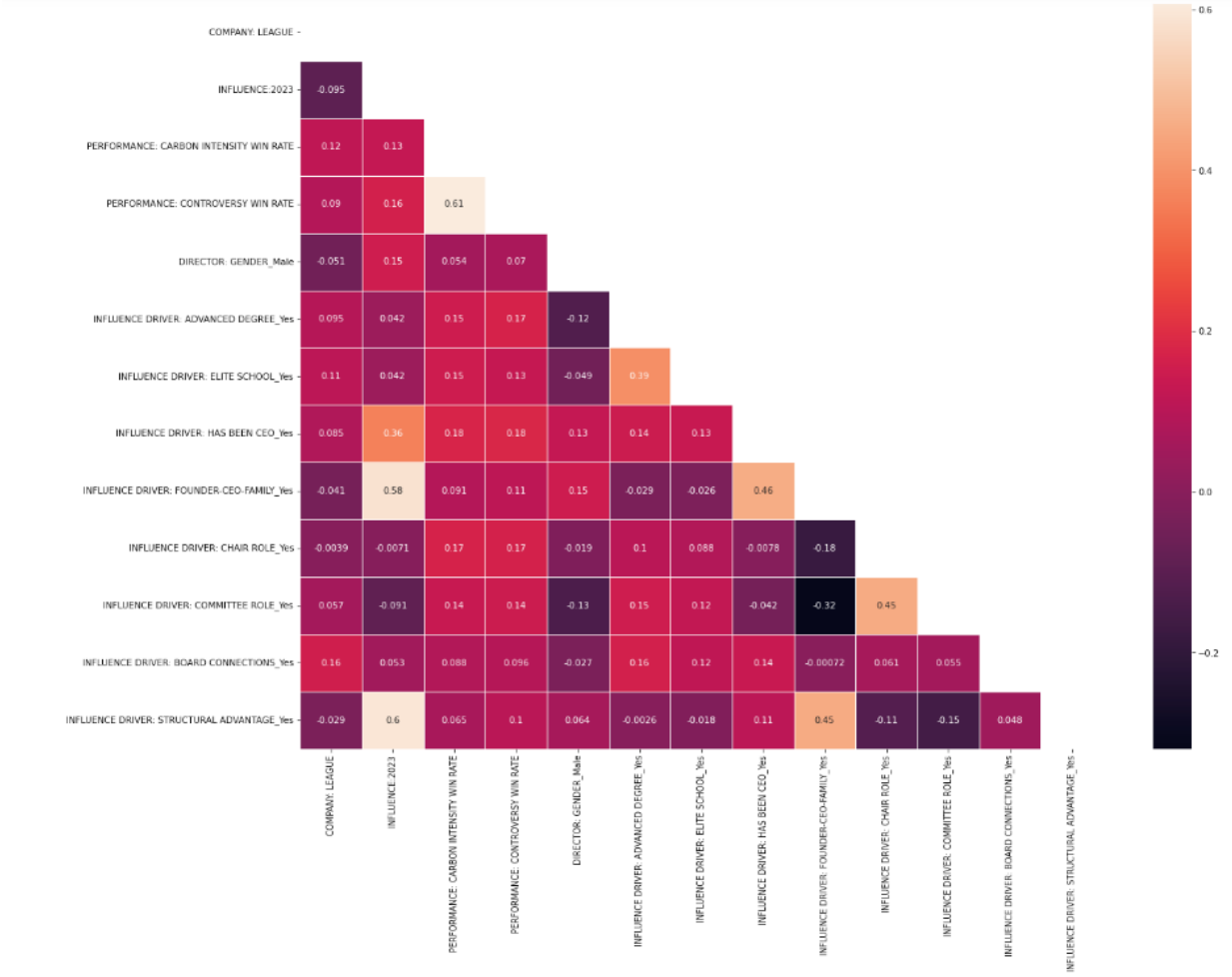

Figure 3: Correlation matrix of problem statement to understand the carbon intensity win rate

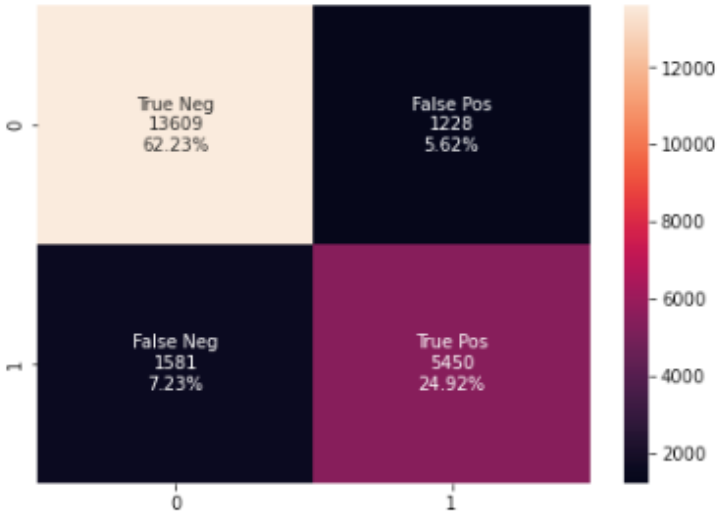

Figure 4: Confusion matrix of problem statement understand the carbon intensity win rate