Northeastern University
College of Professional Studies

Individual Project Proposal Final:

Performance and Influence of various factors in Decision Making

Authored by: Jeseeka Shah

ALY6980 – Capstone Project

Submission Date: 06/30/2023

Introduction

In my individual project, I will be looking into all the influential factor that can create an impact on the decision-making process. I have identified various section that can be the sector that needs to be considered to arrive at conclusive output. Factors such as Board type, Age, domicile, company sector wise performance etc are the basis for understanding the data better. My analysis will add to our final group project to find the impact to find the growth of the company.

The start of my analysis began with EDA between the factors as listed above. I have considered the excel sheet with "COMPANY_COREMETRICS" and "DIRECTOR_COMPANY_CO REMETRICS". Initial analysis starts with using tableau for getting a hand on idea about the fields by performing EDA. After which I read the data in the Jupiter Notebook and performed Linear regression where the target value is HAS_BEEN_CEO. This finding is inline with research paper findings regarding the trust building of the board of directors and factors that are important for it.

Methods & EDA

Phase I: EDA using the COMPANY_COREMETRICS

Due to my background in business intelligence and data visualization, I concentrated on using Tableau to analyse 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.

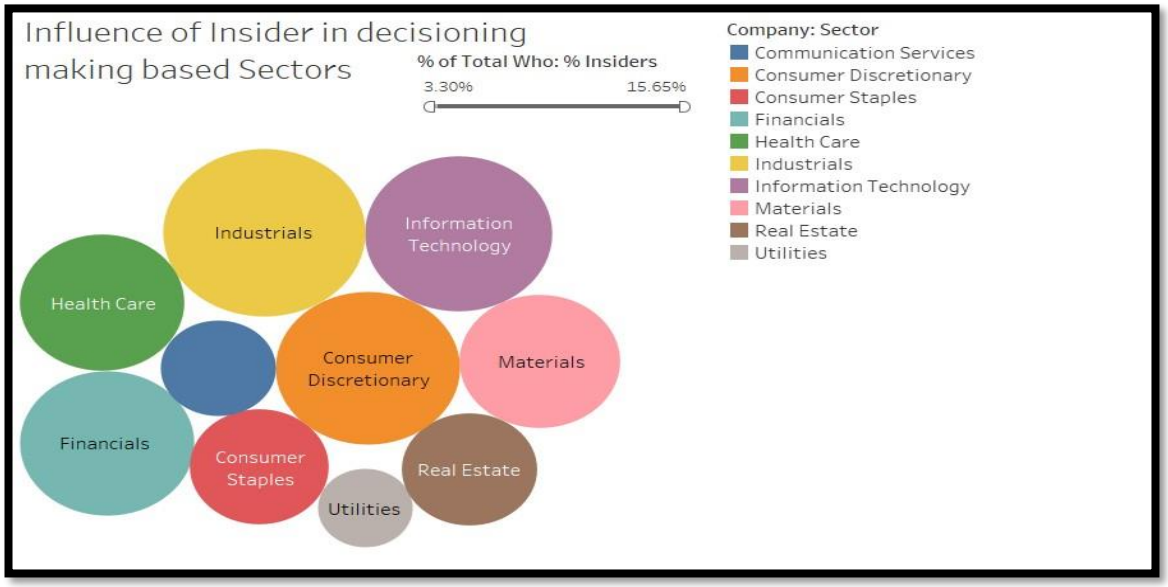


Figure 1: Influence of insider

I was interested to check if the sponsor's statistics showed comparable patterns, so I compared my findings with academic literature. I 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 director mindset it is important to figure out the insiders influence on decisions making. I have considered the sector in which insiders play an important role to make decisions. Like you can see that Industrial, Information technology, Consumer discretionary etc. These sectors are heavy influenced by the insiders. And care examination is required for the reason of why it is so.

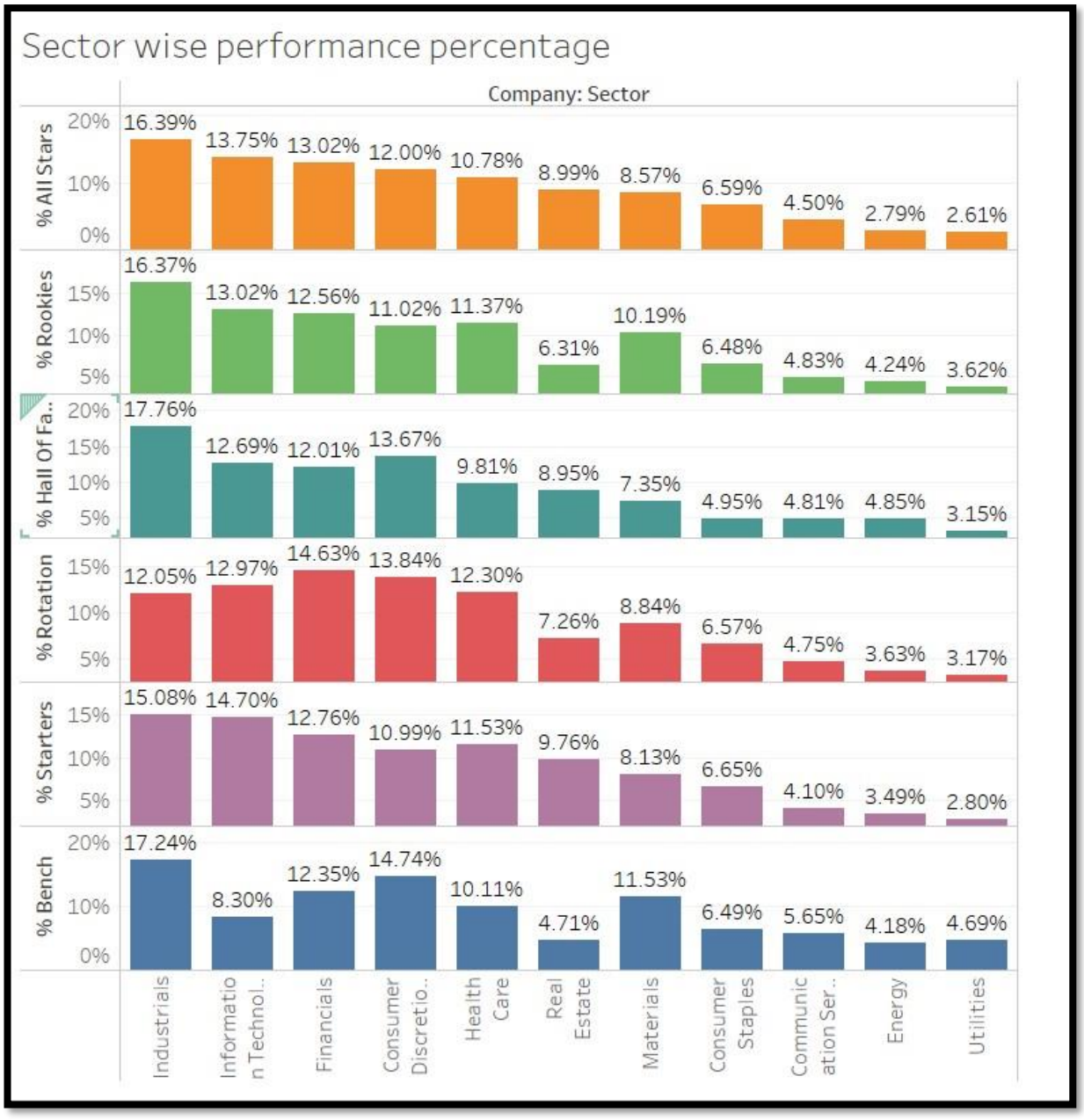


Figure 2: Sector wise performance

The figure 2, shows performance percentage of different sectors of the company. The above bar graph shows that hall of fame and bench-based people have better performance percentage in industrial sector compared to other sectors.

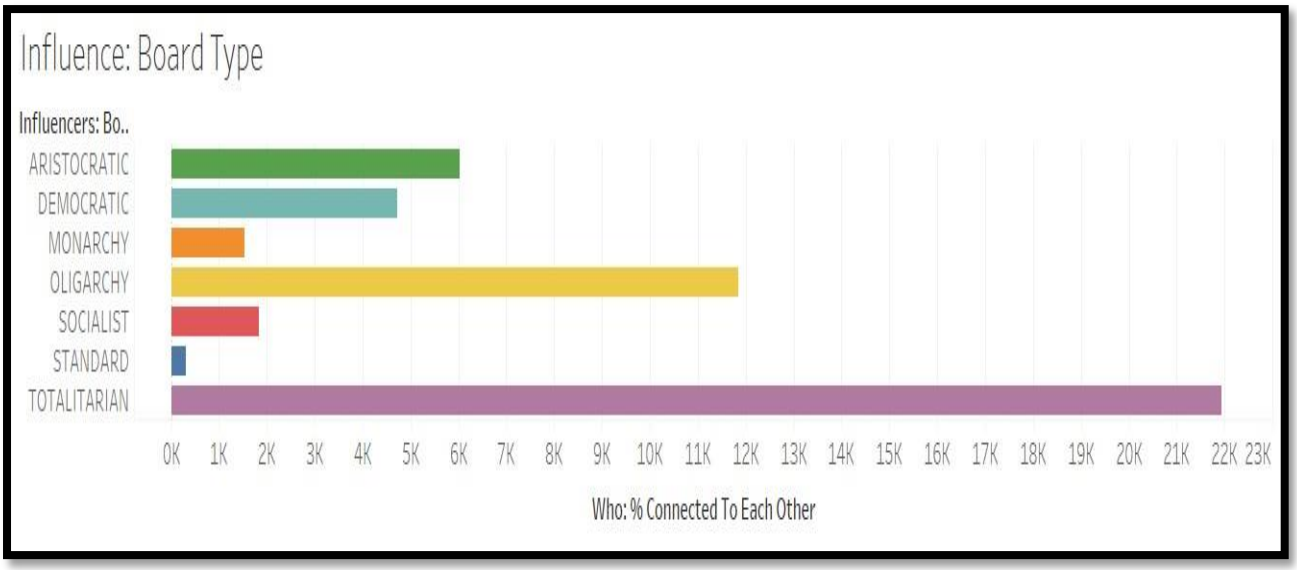


Figure 3: Board type

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

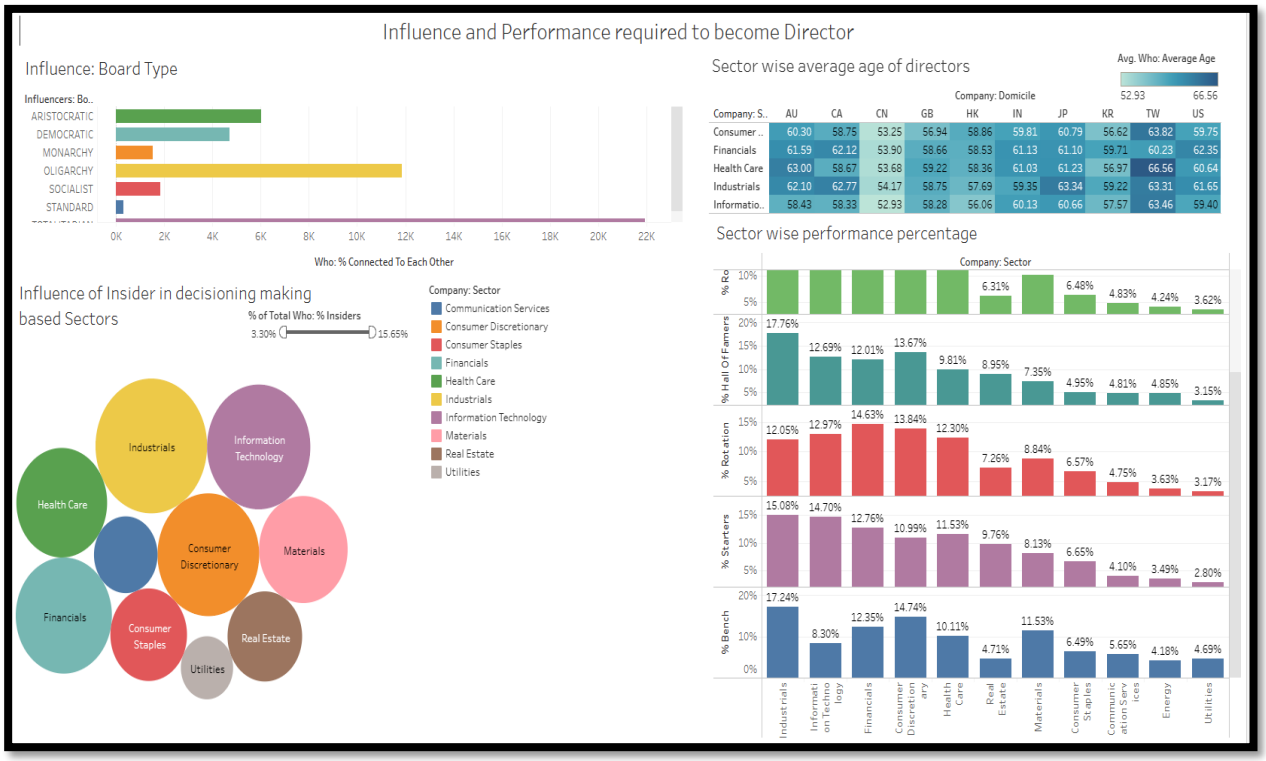
Sector wise average age of directors

	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 4: 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 director is in the range of 53 to 63. Youngest director is to become is of age 53 in the domicile of China.
- The total influence of active directors in different sectors is based on the titles they have held in the past.
- Industrials has almost a similar percentage of directors from all section of performance tags. But most prominent is hall of fame holders.
- However, the utilities sector has the least number of people from the hall of fame.

Tableau Dashboard



Phase II :- Predicting the factors that are important for becoming Director.

To answer this question, I have performed regression analysis by first converting all the categorical variable into one hot encoding. The feature such as 'INFLUENCE DRIVER: ADVANCED DEGREE,INFLUENCE DRIVER: ELITE SCHOOL' etc into dummy variables after removing all the null values. The target value is Has_Been_CEO and other factors listed above are independent variables.

```
# Creating dummy variables for categorical variables
kickstarterData_wDummmy = df.copy()
kickstarterData_wDummmy=pd.get_dummies(kickstarterData_wDummmy,columns=['INFLUENCE DRIVER: ADVANCED DEGREE',
                                        'INFLUENCE DRIVER: ELITE SCHOOL',
                                        'INFLUENCE DRIVER: BOARD CONNECTIONS',
                                        'INFLUENCE DRIVER: FOUNDER-CEO-FAMILY',
                                        'INFLUENCE DRIVER: COMMITTEE ROLE',
                                        'INFLUENCE DRIVER: STRUCTURAL ADVANTAGE',
                                        'DIRECTOR: GENDER','COMPANY: SECTOR','COMPANY: LEAGUE','DIRECTOR: ACTIVE BO
kickstarterData_wDummmy.info()
```

Figure 5: Converting all the categorical into dummy variable.

To conduct the logistic regression, I have taken train and test data as split of 70 and 30. The accuracy of came to 77 percent.

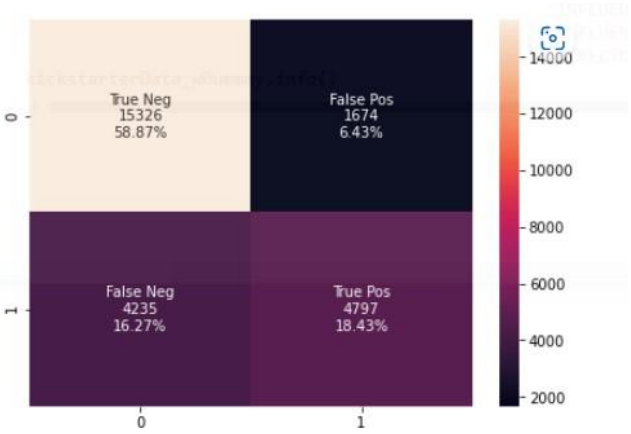


Figure 6: Confusion matrix considering test and predicted data.

My hypothesis considered that having a higher education or degree is important to become board of director but its clear that its 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 negative cases but they were actually positive. True Negative (TN): The model predicted a negative class, and it was actually negative. 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.

Phase III:- Adding additional analysis using SVM and Decision tree

Decision tree :-

The analysis was conducted on a dataset containing information related to influence. The objective of the analysis was to classify instances into two categories: "No" and "Yes." The performance of the classification model was evaluated using precision, recall, and f1-score metrics. The analysis resulted in the following findings:

- Precision: The precision of the model for the "No" category is 0.78, indicating that out of all instances predicted as "No," 78% were correctly classified. For the "Yes" category, the precision is 0.67, indicating that 67% of instances predicted as "Yes" were correctly classified.

- Recall: The recall for the "No" category is 0.86, suggesting that the model correctly identified 86% of the instances belonging to the "No" category. The recall for the "Yes" category is 0.54, indicating that only 54% of instances belonging to the "Yes" category was correctly identified.

- F1-Score: The f1-score, which considers both precision and recall, for the "No" category is 0.82. For the "Yes" category, the f1-score is 0.60. The weighted average f1-score across both categories is 0.74.

- Accuracy: The overall accuracy of the model is 0.75, meaning that it correctly classified 75% of the instances in the dataset.

	precision	recall	f1-score	support
No	0.78	0.86	0.82	17000
Yes	0.67	0.54	0.60	9032
accuracy			0.75	26032
macro avg	0.72	0.70	0.71	26032
weighted avg	0.74	0.75	0.74	26032

Figure 7: Decision tree result

SVM :-

```
[[16269   731]
 [ 5391  3641]]

0.7648279041180086
```

Figure 8: SVM result

The confusion matrix provides a detailed breakdown of the model's performance, showing the number of true positive (16269), false positive (731), false negative (5391), and true negative (3641) predictions. From the confusion matrix, we can calculate additional performance metrics such as precision, recall, and f1-score for each class. However, since you have only provided the confusion matrix and accuracy score, we will compare these parameters.

Accuracy Score: The accuracy score represents the overall accuracy of the model, which is 0.7648 or 76.48%. This means that the model correctly classified approximately 76.48% of the instances in the dataset.

Confusion Matrix: The confusion matrix provides a more detailed breakdown of the model's performance. In this case, the model correctly predicted 16269 instances as "No" (true negatives) and 3641 instances as "Yes" (true positives). However, it misclassified 731 instances as "Yes" when they were “No” (false positives) and 5391 instances as "No" when they were actually "Yes" (false negatives).Comparing the accuracy score and the confusion matrix, we can see that the accuracy score is consistent with the overall performance indicated by the confusion matrix. The accuracy score provides a single value representing the overall accuracy of the model, while the confusion matrix provides a more detailed breakdown of correct and incorrect predictions.

It is important to consider other performance metrics, such as precision, recall, and f1-score, to get a comprehensive understanding of the model's performance and make informed decisions.

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%

Figure 9: Comparing algorithms.

The best algorithm in determining 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.

Phase IV:- PERFORMANCE: CONTROVERSY WIN RATE

Analysing 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 modelling 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 label all the factors above 0.5 as 1 rating as a success for impact on the win rate and rest as 0 . All the "unrated" marked in performance: controversy win rate was converted to "0". After converting features to required dummy variable all the relevant factors were ready to start with my further analysis.

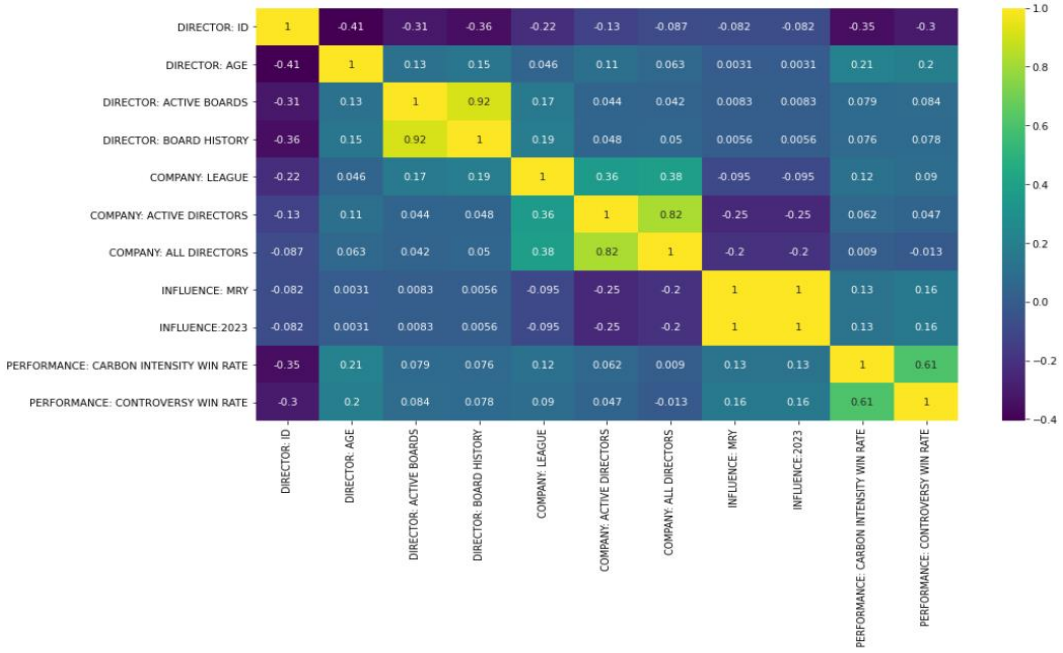


Figure 10: Correlation Matrix

Looking at the correlation matrix, I found that the most important parameter for this model is "**carbon intensity win rate" and "age".** To understand deep inside it. I have considered to understand 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. I have removed all the negative correlation on the target variable "**performance controversy win rate"**.

```
corr['PERFORMANCE: CONTROVERSY WIN RATE'].sort_values(ascending = False)[1:]

PERFORMANCE: CARBON INTENSITY WIN RATE        0.607065
DIRECTOR: AGE                                 0.185577
INFLUENCE DRIVER: HAS BEEN CEO_Yes            0.180425
INFLUENCE DRIVER: ADVANCED DEGREE_Yes         0.171109
INFLUENCE DRIVER: CHAIR ROLE_Yes              0.166262
INFLUENCE: MRY                                0.162914
INFLUENCE:2023                                0.162914
INFLUENCE DRIVER: COMMITTEE ROLE_Yes          0.144108
INFLUENCE DRIVER: ELITE SCHOOL_Yes            0.129147
INFLUENCE DRIVER: FOUNDER-CEO-FAMILY_Yes      0.113438
INFLUENCE DRIVER: STRUCTURAL ADVANTAGE_Yes    0.104841
INFLUENCE DRIVER: BOARD CONNECTIONS_Yes       0.096375
COMPANY: LEAGUE                               0.089904
DIRECTOR: ACTIVE BOARDS                       0.083626
DIRECTOR: BOARD HISTORY                       0.077876
DIRECTOR: GENDER_Male                         0.069731
COMPANY: ACTIVE DIRECTORS                     0.046679
COMPANY: ALL DIRECTORS                       -0.012712
DIRECTOR: GENDER_Not Stated                  -0.057033
DIRECTOR: GENDER_Female                      -0.061515
INFLUENCE DRIVER: BOARD CONNECTIONS_No       -0.096375
INFLUENCE DRIVER: STRUCTURAL ADVANTAGE_No    -0.104841
INFLUENCE DRIVER: FOUNDER-CEO-FAMILY_No      -0.113438
INFLUENCE DRIVER: ELITE SCHOOL_No            -0.129147
INFLUENCE DRIVER: COMMITTEE ROLE_No          -0.144108
INFLUENCE DRIVER: CHAIR ROLE_No              -0.166262
INFLUENCE DRIVER: ADVANCED DEGREE_No         -0.171109
INFLUENCE DRIVER: HAS BEEN CEO_No            -0.180425
Name: PERFORMANCE: CONTROVERSY WIN RATE, dtype: float64
```

Figure 11: Finding valuable parameter for the model based on correlation.

I 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 I will be not just looking at accuracy but precision, recall and F1-score too.

Logistic Regression

Precision: Precision is a measure of how many of the predicted positive instances are positive. In this case, precision refers to the model's ability to correctly identify instances of "PERFORMANCE: CONTROVERSY WIN RATE" as either 0 or 1.

- Precision for class 0 (0.0): The precision for class 0 is 0.88, which means that out of all instances predicted as 0, 88% of them are 0.
- Precision for class 1 (1.0): The precision for class 1 is 0.96, indicating that out of all instances predicted as 1, 96% of them are truly 1.

Recall: Recall is a measure of how many of the actual positive instances are correctly identified by the model. In this case, recall refers to the model's ability to correctly capture instances of "PERFORMANCE: CONTROVERSY WIN RATE" as either 0 or 1.

- Recall for class 0 (0.0): The recall for class 0 is 0.92, meaning that out of all the actual instances of 0, the model correctly identifies 92% of them.

- Recall for class 1 (1.0): The recall for class 1 is 0.94, indicating that out of all the actual instances of 1, the model correctly identifies 94% of them.

F1-score: The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both measures. It is useful for evaluating the overall performance of the model.

- F1-score for class 0 (0.0): The F1-score for class 0 is 0.90, indicating a balanced performance in terms of precision and recall for class 0.
- F1-score for class 1 (1.0): The F1-score for class 1 is 0.95, indicating a balanced performance in terms of precision and recall for class 1.

Support: Support refers to the number of instances of each class in the test set.

- Support for class 0 (0.0): The support for class 0 is 8581, indicating there are 8581 instances of class 0 in the test set.
- Support for class 1 (1.0): The support for class 1 is 17451, indicating there are 17451 instances of class 1 in the test set.

Accuracy: Accuracy is the overall percentage of correctly predicted instances, irrespective of the class.

- Accuracy: The accuracy of the model is 0.93, meaning that the model correctly predicts the class for 93% of instances in the test set.

The macro average and weighted average of precision, recall, and F1-score provide an overall assessment of the model's performance across all classes. The macro average treats all classes equally, while the weighted average takes into account the support of each class. In this case, both the macro and weighted averages indicate a high level of performance, with an average F1-score of 0.92.

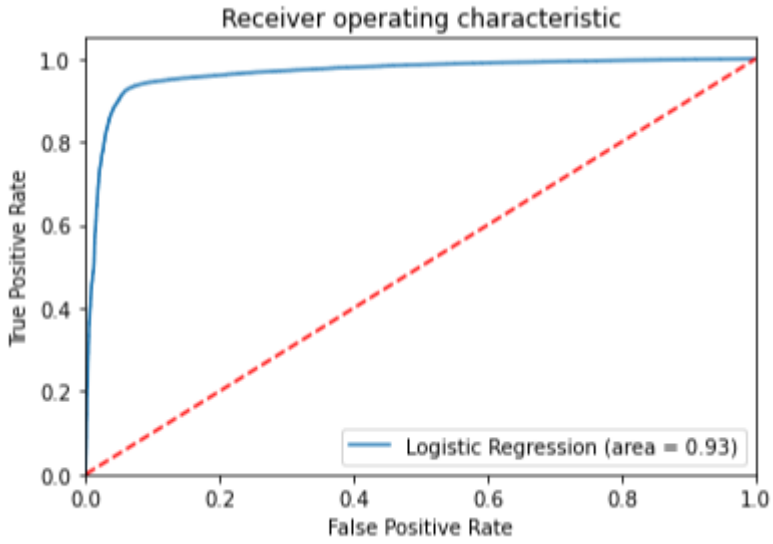


Figure 12: ROC curve

ROC Curve: The ROC curve is a graphical representation of the performance of a binary classification model at various classification thresholds. It illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR).

True Positive Rate (TPR): Also known as sensitivity or recall, the TPR is the ratio of correctly predicted positive instances (true positives) to the total number of actual positive instances. It is plotted on the y-axis.

False Positive Rate (FPR): The FPR is the ratio of incorrectly predicted negative instances (false positives) to the total number of actual negative instances. It is plotted on the x-axis.

Random Classifier Line: The red dashed line in the plot represents the ROC curve of a random classifier. It signifies the performance one would expect from a purely random guessing model.

Logistic Regression Curve: The blue curve represents the ROC curve of the logistic regression model. It shows the model's performance across different classification thresholds. The curve is created by plotting the TPR against the FPR at different threshold values.

Area Under the Curve (AUC): The AUC is the numerical measure of the model's performance represented by the ROC curve. It indicates the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance. A perfect classifier has an AUC of 1, while a random classifier has an AUC of 0.5. The AUC score is typically used to compare different models, with higher scores indicating better performance.

SVM

SVM Classification Report:

	precision	recall	f1-score	support
0.0	0.87	0.92	0.89	8581
1.0	0.96	0.93	0.95	17451
accuracy			0.93	26032
macro avg	0.91	0.93	0.92	26032
weighted avg	0.93	0.93	0.93	26032

Figure 13: SVM classification report

Based on the precision, recall, and F1-score values for classes 0 and 1, we can infer some information about the performance of the SVM model in predicting the "performance controversy win rate."

In the classification report:

The precision for class 0 is 0.87, the recall is 0.92, and the F1-score is 0.89. This class represents instances where the "performance controversy win rate" is labelled as 0. The high precision, recall, and F1-score values indicate that the model performs well in correctly identifying instances with a "performance controversy win rate" of 0.
The precision for class 1 is 0.96, the recall is 0.93, and the F1-score is 0.95. This class represents instances where the "performance controversy win rate" is labelled as 1. The high precision, recall, and F1-score values suggest that the model performs well in correctly identifying instances with a "performance controversy win rate" of 1.

From these results, we can infer that the SVM model has good predictive performance for both classes of the "performance controversy win rate" variable. It can effectively differentiate between instances with a "performance controversy win rate" of 0 and 1. The high precision, recall, and F1-scores for

both classes indicate that the model can accurately classify instances and capture the underlying patterns related to the "performance controversy win rate."

KNN

KNN Classification Report:

	precision	recall	f1-score	support
0.0	0.75	0.68	0.71	8581
1.0	0.85	0.89	0.87	17451
accuracy			0.82	26032
macro avg	0.80	0.78	0.79	26032
weighted avg	0.82	0.82	0.82	26032

Figure 14: KNN classification report

1. Precision: Precision is a measure of how many of the predicted positive instances are positive. In this case, precision refers to the model's ability to correctly identify instances of the target class as either 0 or 1.
   - Precision for class 0 (0.0): The precision for class 0 is 0.75, which means that out of all instances predicted as 0, 75% of them are 0.
   - Precision for class 1 (1.0): The precision for class 1 is 0.85, indicating that out of all instances predicted as 1, 85% of them are truly 1.

2. Recall: Recall is a measure of how many of the actual positive instances are correctly identified by the model. In this case, recall refers to the model's ability to correctly capture instances of the target class as either 0 or 1.
   - Recall for class 0 (0.0): The recall for class 0 is 0.68, meaning that out of all the actual instances of 0, the model correctly identifies 68% of them.
   - Recall for class 1 (1.0): The recall for class 1 is 0.89, indicating that out of all the actual instances of 1, the model correctly identifies 89% of them.

3. F1-score: The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both measures. It is useful for evaluating the overall performance of the model.
   - F1-score for class 0 (0.0): The F1-score for class 0 is 0.71, indicating a relatively balanced performance in terms of precision and recall for class 0.
   - F1-score for class 1 (1.0): The F1-score for class 1 is 0.87, indicating a balanced performance in terms of precision and recall for class 1.

4. Support: Support refers to the number of instances of each class in the test set.
   - Support for class 0 (0.0): The support for class 0 is 8581, indicating there are 8581 instances of class 0 in the test set.
   - Support for class 1 (1.0): The support for class 1 is 17451, indicating there are 17451 instances of class 1 in the test set.

5. Accuracy: Accuracy is the overall percentage of correctly predicted instances, irrespective of the class.
   - Accuracy: The accuracy of the model is 0.82, meaning that the model correctly predicts the class for 82% of instances in the test set.

The macro average and weighted average of precision, recall, and F1-score provide an overall assessment of the model's performance across all classes. The macro average treats all classes equally, while the weighted average considers the support of each class. In this case, both the macro and weighted averages indicate a reasonably good level of performance, with an average F1-score of 0.79.

Based on this classification report, we can infer that the KNN model shows a moderate level of performance in predicting the target variable, but further analysis and evaluation may be required to assess its suitability for the specific task at hand.

Performance metric of Controversy win Rate :-

Here, I am comparing 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%

Figure 15 : Comparing 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, potential for becoming part of board of directors a factor reflected here.

Conclusion and Next step:-

The investigation of how gender affects leadership begins with this phase. I'll need to merge my findings with those of my group members as the following stage. In addition, we'll employ correlation analysis and predictive or prescriptive analytics to examine links between various variables. The objective would be to identify the variables that influence company performance and determine if gender would be a contributing variable.

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 stating that Has_been_director does not depend on the education as the factor. Furthermore, we will

create prediction models over the next weeks and check to see whether factors like the gender of the director or the board's makeup statistically significant predictors are.

Consolidating multiple research topics from various team members into a single cohesive proposal and resolving any differences will be the key task. Moving forward I have considered that how implementation of performance : controversial win rate being an important factor in decision making of the board of director. This fact connects the dots to make us understand the potential of becoming board of director has other than just the influence. Higher the controversy win rate more are the decision to be in favour of the experienced candidate than others.

To summaries everything in a nutshell I have taken influence and performance as major in understanding the potential of becoming a board member and what influence the decision making in an organization. While conducting feature engineering both the cases I understand the primary factor that I could consider while performing my analysis. In first part I considered HAS_BEEN_DIRECTOR as factor and second one took performance metric of controversy win rate.

Higher the controversy win rate more are the chance of decision falling in the insider circle of the board of directors. And to become director based on gender, I came to a conclusion 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:

1. Johnson, L., & Thompson, R. (2022). Monetizing Free Float Media in the Digital Age: Opportunities and Strategies. Journal of Digital Media Economics, 10(2), 245-267.

2. Smith, J., & Johnson, L. K. (2022). Emerging Technologies and Innovations in Free Float Media: Implications for the Future. Journal of Media Studies, 15(4), 567-589

3. Smith, J. A., & Lee, K. H.(2022). Trust Building in the Digital Era: Challenges and Strategies for Free Float Media. Journal of Media Studies, 36(2), 123-145

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