Capstone Sponsor Deliverable

Empirical Study on the Directors in US-Issued Companies

Group 7:

Fatima Nurmakhamadova

Shuting Chen

Xiaolu Shen

Yumeng Xie

Instructor: Roy Wada
ALY 6980 - 70447: Capstone
Academic Term: Fall Quarter, 2022
College of Professional Study, Northeastern University
December 18, 2022

Contents

Introduction 4
Literature Review 6
Research Methodology 8
Research Assumptions 9
Research Hypotheses 9
Quantitive Influence 10
Identities 10
Network & Resources 11
Directors’ Options 12
Others 12
Exploratory Data Analysis on Original Data 14
Data Processing 18
Data Transformation and Integration 18
Feature Engineering 21
Target Variable 22
Multiple Directorships across Different Companies 24
Historical and Present Identities 25
IPS, Third-Degree Connection, and Controversies 27
Data Cleaning 30
Scope Narrowing 30
Missing Values 31
Imbalanced Categorical Features 33
Descriptive Analysis 35
Historical Departure Frequency 35
Departure with Age & Genders 36
Departure with Committee Members 37
Departure with Identities 38
Departure with Number of Directorships 39
Departure with Influence Percentage 40
Departure with Shareholder Percentages 41
Departure with Weighted Network Connectivity 42
Departure with IVY League Backgrounds 43
Predictive Models Part 1 45
Predictive Models and Performance Evaluation 46

Regression Analysis 50
Data Preparation 50
Model Evaluation 52
Model Interpretation 53
Predictive Models Part 2 57
Data Preparation 58
Model Performance 59
Conclusions 63
Hypotheses Test Results 63
Predictive Modeling 64
References 67

Introduction

This report aims to demonstrate our project, "*empirical study on board directors in the US-issued companies*" and the project goal is to analyze the structural aftermath of the director boards during their firms' disclosed controversial events and create predictive models to label their likelihoods and predicted binary departure results ("departed" & "not-departed"). The following report will cover various aspects of our research such as literature reviews, research methodologies, data processing, descriptive analysis, modeling, and so on.

This project is sponsored by Free Float Media. In the meantime, we are offered with multiple data sources, mainly including three aspects:

- Individual directors: names, educational backgrounds, estimated ages, genders, tenures of the boards, identities in their companies, positions of committees, various shareholder percentages, etc.
- Firms: issuer tickets, the number of disclosed controversial activities (2008 – 2022), market capitals (2012 – 2021), corporate parents, ownership category, sector, etc.
- Connectivity among the directors: existing connectivity between every two directors, overlapped time, number of third-degree connections, etc.

The original tasks are to analyze how the boards of directors will change during the controversial events and predict the likelihood of departures of directors if controversies happen and are disclosed to the public again. Both tasks rely on the question: who left (either was expelled or left the board from his/her own will) when their companies were involved in controversial events? On the one hand, we know exactly when each director entered and left the board. Comparing it with the time of controversial events, we can easily solve the first task. However, on the other hand, there is no evidence or proof that if one director left one company's board in the past, he/she would leave again from the next company with controversies. So, focusing on historical behavior is not a proper way in our research. But the directors who left (or stayed) when controversial events happened may have similar characteristics. For example, does a director with more resources that its company needs (expertise, network, stock shares …) have the same possibilities to leave or be departed when the company involves controversy compared to another director with fewer resources? Probably not. Thus, our research goal is to analyze how individual directors' characteristics influence departure results from their companies.

Literature Review

The whole research proposal is established on the assumption (or phenomenon) that when a firm involves one or more controversial events (like fraud, sexual assault, and so on) in a period of time, its director board may get changed. According to Srinivasan (2005), after the fraudulent activity is disclosed to the public, there is usually significant upheaval in the structures of the boards; and during this period of director departures, there are two mechanisms to explain the departures: 1) "jumping ship" – director protecting its individual reputation, and 2) "cleaning house" – firm mitigating the unfavorable judgment of outside audiences and protect the firm's access to needed resources. In the short sense, the action of leaving the board can be interpreted by the director's subjective action or the firm's objective action. However, due to the nature of the boards and the business secrets, we can never learn about which mechanism drove the leaving; but, it is for sure that the action of leaving the board is related to the controversial events.

The next thing is to find out who is more likely to be departed during the controversy and who is not. Instead of focusing on exactly who left the boards, our research wants to focus on what characteristics of directors will influence the departures during controversial events. In addition to the previous findings, to figure out which one of these two mechanisms drives the directors to leave the boards, Marcel and Cowen (2013) applied the logistic regression model to analyze which one of these two mechanisms drives the directors to leave the boards. The results of their variable analysis, such as director capital, whether the director is the nominating committee member, and so on, offer the reference to evaluate which outside directors having certain characteristics may leave the board due to these two mechanisms. On the other hand, D'Onza and Rigolini (2016) applied the regression model to analyze the relationship between "the capital of the director" and "departure after the fraud is announced to the public" and showed that "business expertise, industry expertise, and the number of board appointments have significant and negative relationships with the likelihood of a director leaving the board at the end of his or her term after an announcement of fraud".

Except for the characteristics mentioned above which may influence the departure of directors on boards after the fraud activities, the social networks among the directors are another thing we want to explore. Kuang and Lee (2017) proved that well-connected independent directors on board have a significant and negative impact on the detection of fraudulent activities and because of less possibility of fraudulent activities being detected, the firm with well-connected independent directors faces fewer charges and less severe consequences. Also, Kuang and Lee showed that "independent directors' connections to fraud firms significantly increase a firm's propensity to fraud commission and the likelihood of fraud detection is also higher", which inspires us to pay attention to the possibility of the directors' social network could influence the departures of the boards when fraud happens. In addition, Gao and other researchers (2016) found that directors with multiple directorships at other firms have relatively higher possibilities of leaving the firm involved in the controversial activity than those who have one single directorship in the fraudulent company.

Research Methodology

To achieve the project goal, our research will mainly include four phases:

Phase 1: Research Assumptions and Hypotheses. In the first phase, we will set several assumptions to overcome the limitations of the original data sets and raise hypotheses regarding the features we have about the individual directors and new features we will create (hypotheses will be tested in Phase 3 and Phase 4)

Phase 2: Data Processing and Feature Engineering. Considering the data sources provided by our sponsor are in different structures and focusing on different levels (company & director levels), the data processing should be completed before the analysis. In this phase, we will complete the data integration, data cleaning, and research scope filtering. On the other hand, we need to create new features (the characteristics of directors) from the processed data to fit the following research process and test the raised hypotheses, including identifying the departed directors, multiple directorships across companies, connectivity, and so on. In the meantime, we will apply the secondary data structure regarding the different approach to deal with historical features (influence percentage score, and so on) in the modeling part.

Phase 3: Descriptive Analysis. After Phase 2, we will apply statistical analysis and visualizations to compare and demonstrate distributions of departed and non-departed directors' characteristics, analyzing the historical changes due to the characteristics and gaining references for the following phase.

Phase 4: Modeling Analysis. The predictive models will be created in this phase, solving the original need of our sponsor: labeling the likelihood of individual director's departure. Also, we will apply a logistic regression model to quantitively evaluate characteristics' influence on the departures.

Research Assumptions

Due to the limits of original data, which only contains the yearly count of controversial events for each company, we cannot identify which controversial events harmed the companies and caused directors got departed, neither nor who should be responsible for the event (influence percentage data may indicate who are responsible, but it does not indicate who will take the consequence). As a matter of fact, we may never learn about the details and results of controversial events within all these companies because the duration of these events could last for years. Also, director departures involve business secrets and strategy considerations. For example, when the naturally responsible person may be more critical to the company (having more resources that the company must obtain, etc.) than another less important director, the less critical director got departed as the company needs to show commitment to admitting wrongdoing to the public after the controversial being disclosed. After talking with our sponsor about our concerns, we raised the following assumptions:

Assumption 1: all the controversial events recorded in the yearly total counts are negative events for the companies.

Assumption 2: if one director left the company in the year when his/her former company was recorded with one or more controversial events, his/her leave is regarded as a departure order from the company.

Assumption 3: if one director left the company in the year when his/her former company was recorded with NO controversial events, his/her leave is NOT regarded as a departure order from the company.

Research Hypotheses

As mentioned before, our research method is to analyze how directors' characteristics influenced their departures and stay when the controversial events happened and further apply the characteristic having significant influence (positive or negative) on departures in our classification model. According to scholars' research (Marcel & Cowen, 2013 and D'Onza & Rigolini, 2016), the departures may be influenced by the directors' capital, gender, multi-memberships in different companies, and so on. Thus, based on the information provided in the original data sets, we have the following hypothesis regarding departure influence from different types of features (characteristics of individual directors):

Quantitive Influence

H1.1: the influence percentage score of the director has a negative influence on departures when controversial events happen. The influence percentage score is the attribute created by our sponsor. It quantitively reflects how much influence an individual director had on the company. For example, if director A and director B both shared the director board of the same company and let's say director A had a 70% of influence percentage score while director B had only a 30% of influence percentage score at the same time, this means that director A had more influence on the company than director B did. We hypothesize that director A is less likely to be departed when the controversial event happened compared to director B.

H1.2: the percentage of dominant share holds, insiders' officers directors holds, principle share holds, and controlling share holds all have negative influence on departures when controversial events happen.

Identities

H2.1: having positions in the pay committee, audit committee, and nominating committee negatively influences the likelihoods of departures when controversial events happened compared to the situation of not having positions in these committees. Similar to the previous influence percentage scores, directors having positions in these committees are more important or more influential to the company, thus we hypothesize that the director having positions in committees has a lower likelihood to be departed than those who have not.

H2.2: directors who used to be or currently are chairmen of the board, lead directors, or CEOs have a lower likelihood to be departed than the directors who are not these characters in companies.

H2.3: directors who are company founder or executive directors are less likely to be departed than the directors who have not these identities.

Network & Resources

H3.1: the weighted connectivity of directors has a negative influence on the departure when the company involves controversial events. For example, director C and director D are both on the board of the same company; director C's weighted connectivity is 18 while director D's is 8. According to H3.1, director C is less likely to be departed from the board compared to director D when the controversial event happened.

H3.2: the number of third-degree connections has a negative influence on the departure when the company involves controversial events. First-degree connection means the contacts you know and can be contacted directly; second-degree connection means people connected to the first-degree contacts; third-degree connection means people you could efficiently build connections with through the first-degree and second-degree connections. The greater number of third-degree connections a director owns, the more resources he/she can apply for the firm.

H3.3: the number of directorships in different companies has a negative impact on the departure when controversial events happen. According to the research (Marcel & Cowen, 2013), the director with more capital is less likely to be departed and we consider that the multiple directorships across different companies are one of these capitals: the more directorships in different companies, the more capital director has. Thus, we hypothesize that the more directorships a director has across different companies, the less likely he/she will be departed when the controversial event happens.

H3.4: the tenure length has a negative impact on the departure when controversial events happen. We hypothesize that the longer a director has been on the board, the better connections he/she has built within the company and outside the company. Thus, a director with a longer tenure length is less likely to be departed than another one but with a shorter tenure length.

Directors' Options

H4.1: the options of multiple classes of voting stock have negative impacts on the departure when controversial events happen.

H4.2: the options of independent management both have negative impacts on the departure when controversial events happen.

Others

H5: male directors have a lower likelihood to be departed when controversial events happen than female directors. According to D'Onza and Rigolini (2016), female directors are more likely to leave the boards before controversial events are disclosed to the public. As we cannot confirm the exact time when controversial events were disclosed, according to our assumptions of identifying departed directors, we hypothesize female directors have higher likelihoods to be departed than male directors in the year when companies are involved in controversial events.

H6: age has a positive influence on the departure results when controversial events happen. We hypothesize that the company may prefer relatively younger directors rather than the elders.

H7: the educational background of IVY league schools has a negative influence on the departure when controversial events happen. We hypothesize that the director who graduated from an IVY league school (like Harvard University) has a lower likelihood to be departed compared to the director who has no such educational experience.

Exploratory Data Analysis on Original Data

This section will demonstrate initial Exploratory Data Analysis of the original data that was done using Power BI. This includes data transformation done for narrowing down the focus of the analysis that was also done in Power BI using the Power Query editor.

In the initial analysis, we decided to narrow the focus to Russell 3000 companies as the data on the historical influence of the directors are available only for that group. Thus, the EDA part represents analysis only for companies in the Russell 3000 Index. For this, we downloaded the spreadsheet that contains the list of the Russell 3000 companies which then was used to sort companies in the main file (*Board_of_directors.csv*) leaving only related ones. This was done by merging files using the Inner Join option based on the companies' tickers. Moreover, we merged the rest files using the Inner Join option by companies' id (ISSUERID) and directors' id (INDIVIDUAL_ID) depending on the file, to keep only records of the companies who committed the fraud, and directors who had records of influence share. We ended up with 37,211 records of the directors and companies related to the Russell 3000. Necessary data cleaning and data type correction were done, which includes replacing 0 values in age column with median value of 63.

Next, we visualized important features in the main file about the board of directors to understand them in detail and get some insights. The Figure 3-1 represents the dashboard report which gives an overall idea about the companies and directors of Russell 3000 companies. There are 1,691 companies left in total after the data transformation, where most of them are from financial, industrial, and IT sectors. The dominating ownership type is principal shareholder, followed by widely held, and then controlled. Across all available companies, 86.85% of them have non-executive directors, and only 13.15% of executive directors. The peak of getting in and out of board directors was from 2000 to 2008, which was then followed by an increase in the number of directors getting into the board and a decrease in those getting out. As for the outside directors, the number increases over the whole period of the companies' existence with a slight drop in 2008 which then increased almost twice by the current period.

Figure 3-1

Dashboard of main Features in Board of Directors in Russell 3000 Companies

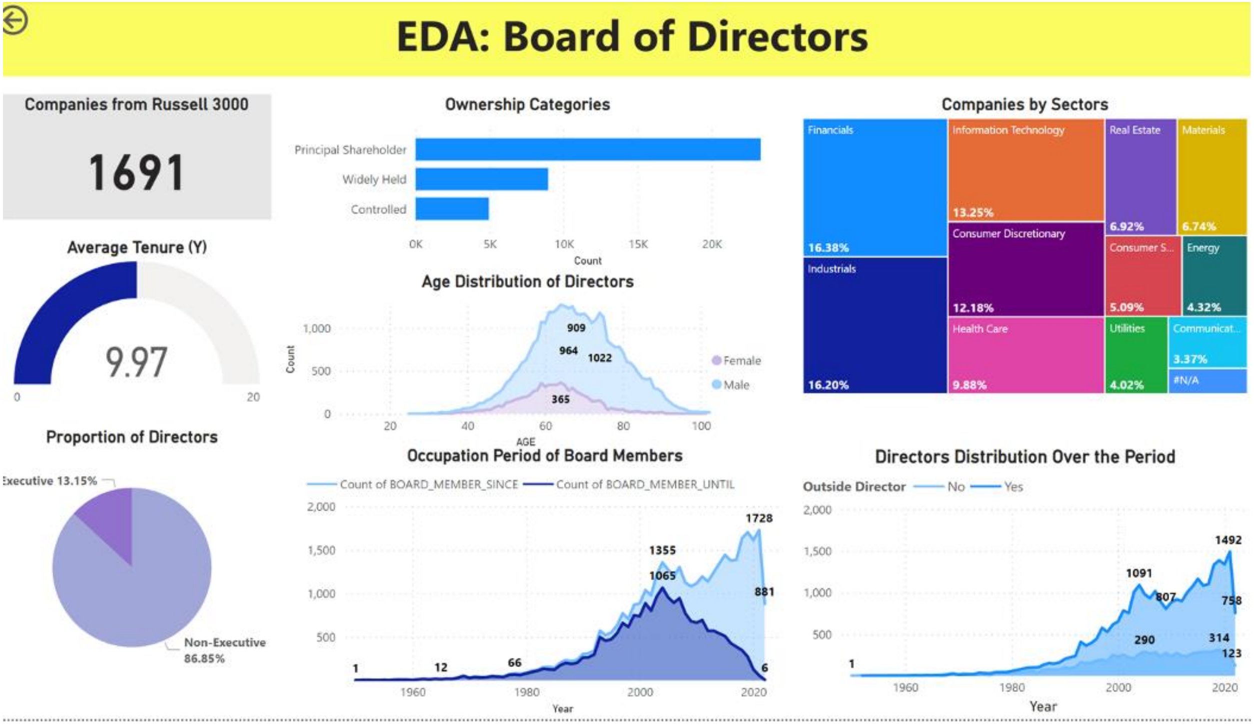
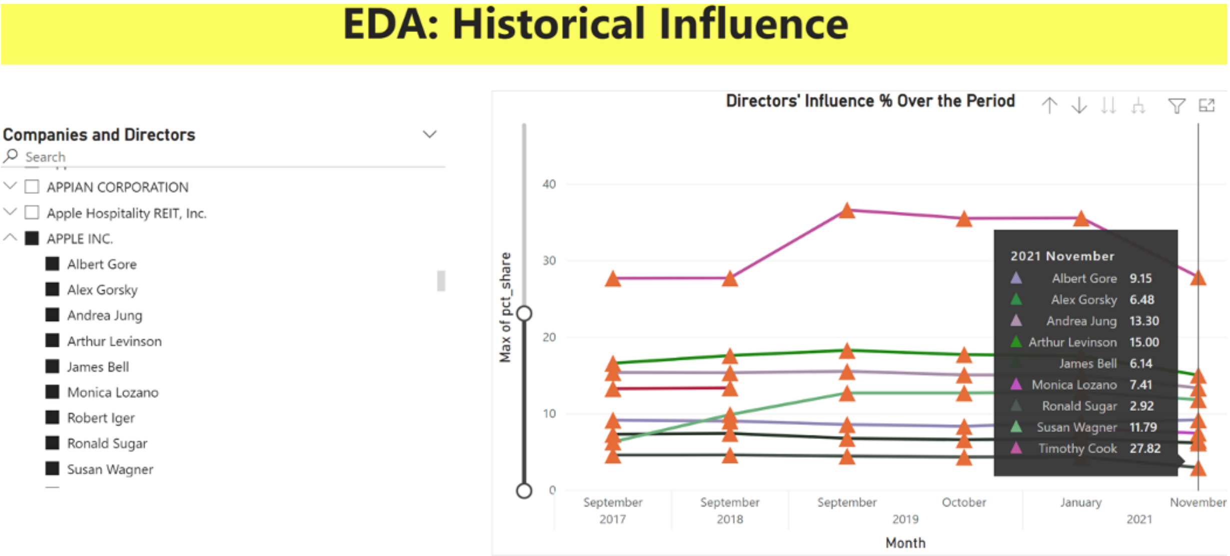


Figure 3-2 demonstrates the drill-down visual report for directors' historical influence data for the period from 2017 to 2022. This dashboard helps to trace each director's influence rate over their period of being on the board. The slicer visual on the left allows filtering data by company and director which makes the trace more convenient. In the line chart on the right, we can see the influence rate distribution of Apple's board of directors. The tooltip displays the names of each director and their influence rate for November 2021. In this dashboard we can see that Timothy Cook has the largest influence share among other directors, having the highest rate of 36.57% in September 2019 which then dropped to 27.6%, to the same level as it was at the beginning of the period. This is higher than the average influence share for directors in the Russell 3000 companies.

Figure 3-2

Drill-Down Visual Report of Historical Influence of Directors in Russell 3000 Companies



2021 November	
Albert Gore	9.15
Alex Gorsky	6.48
Andrea Jung	13.30
Arthur Levinson	15.00
James Bell	6.14
Monica Lozano	7.41
Ronald Sugar	2.92
Susan Wagner	11.79
Timothy Cook	27.82

The summary statistics table provided in Table 3-1 represents the board of directors and historical influence rate datasets. We can see the average age of the directors is 65 – 66 years. While the director's tenure ranges between 0 to 73 years, having an average of 10 years. The average percentage of the dominant shareholders in Russell 3000 companies is 16.4%, while for principal shareholders it is 11.2%, and for control shareholders, it is 7.3%. As for the 3rd-degree connections, most of the directors have 1 or no connection at all, while some of them have more than 10 connections in the years 2021 and 2022.

Table 3-1

Summary Statistics Table of Directors in Russell 3000 Companies

	Age	Tenure	Market cap usd	Dominant shareholder pct	Insiders officers directors held %	Principal shareholder pct	Control shareholder %	2022 3rd degree connect	2021 3rd degree connect
count	37211	37211	3.721100e+04	37211	37211	37211	37211	37211	37211
mean	65.7	9.9	2.755683e+10	16.4	7.6	11.2	7.3	0.46	0.59
std	16.3	8.5	9.580644e+10	17.88	16.76	10.21	19.86	1.42	1.70
min	0	0	8.859080e+07	0	0	0	0	0	0
25%	59	4	1.841465e+09	10.1					
50%	67	8	5.378030e+09	12.6	1.5	11.6	0	0	0
75%	75	14	2.130153e+10	16.6	4.3	15.4	0	0	0
max	109	73	2.310000e+12	100	100	50	100	12	14

Table 3-2 demonstrates that the boards' count ranges from 4 to 34 directors, having on average 10 – 11 directors. As for the influence share percentage, it obviously ranges from 0 to 100%, where directors on average have 10% of the influence share on the board.

Table 3-2

Summary Statistics Table of Directors' Influence Share in Russell 3000 Companies

	Board count	% Share
count	203858	203858
mean	10.4	10.1
std	2.5	9.9
min	4	0
25%	9	4.1
50%	10	7.7
75%	12	12.3
max	34	100

Data Processing

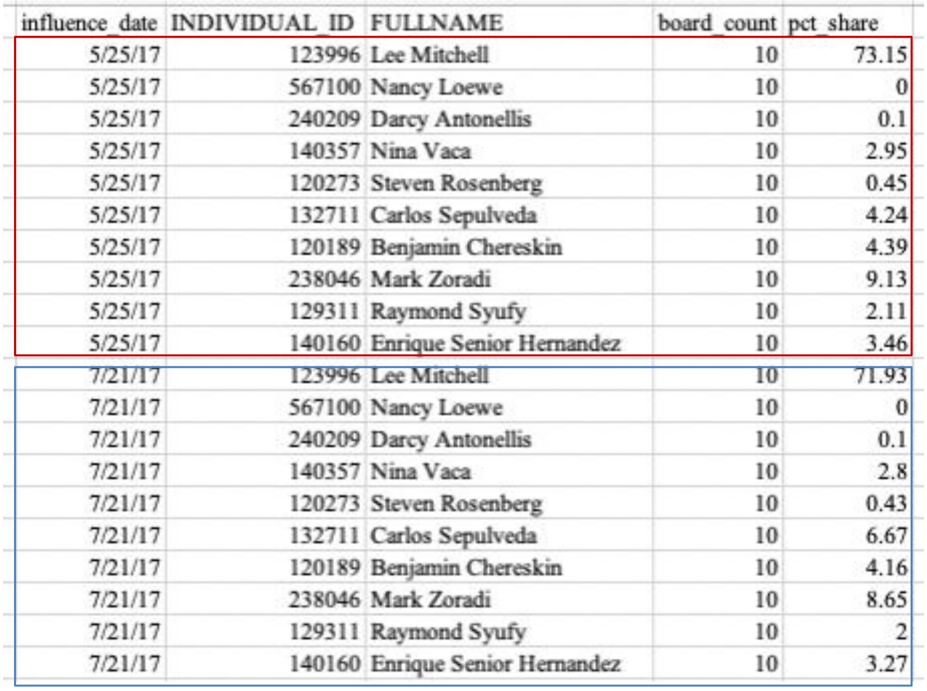
This section will demonstrate processes we have done for the data manipulation, including data transformation, integration, feature engineering, and scope filtering.

Data Transformation and Integration

Although the structures of the data sets provided are almost the same: each director or company has only one record in the original data (controversy count data, director information, etc.). However, the data of influence percentage score and connectivity among the directors is recorded in different structures – each individual director has multiple observations with different objects or at different time points (Figure 4-1). Thus, before merging these two data sets, we need to transform their structures to the other data sources.

Figure 4-1

Screenshot: Part of Original Influence Percentage Score Data

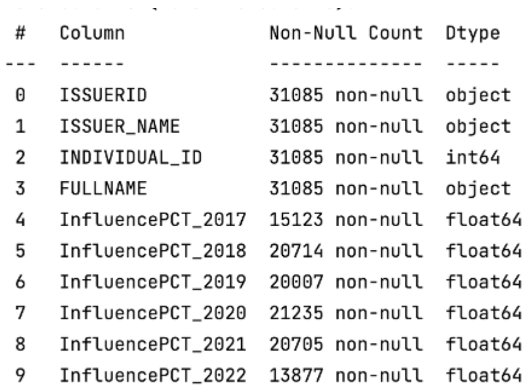


influence_date	INDIVIDUAL_ID	FULLNAME	board_count	pct_share
5/25/17	123996	Lee Mitchell	10	73.15
5/25/17	567100	Nancy Loewe	10	0
5/25/17	240209	Darcy Antonellis	10	0.1
5/25/17	140357	Nina Vaca	10	2.95
5/25/17	120273	Steven Rosenberg	10	0.45
5/25/17	132711	Carlos Sepulveda	10	4.24
5/25/17	120189	Benjamin Chereskin	10	4.39
5/25/17	238046	Mark Zoradi	10	9.13
5/25/17	129311	Raymond Syufy	10	2.11
5/25/17	140160	Enrique Senior Hernandez	10	3.46
7/21/17	123996	Lee Mitchell	10	71.93
7/21/17	567100	Nancy Loewe	10	0
7/21/17	240209	Darcy Antonellis	10	0.1
7/21/17	140357	Nina Vaca	10	2.8
7/21/17	120273	Steven Rosenberg	10	0.43
7/21/17	132711	Carlos Sepulveda	10	6.67
7/21/17	120189	Benjamin Chereskin	10	4.16
7/21/17	238046	Mark Zoradi	10	8.65
7/21/17	129311	Raymond Syufy	10	2
7/21/17	140160	Enrique Senior Hernandez	10	3.27

First, for the influence percentage score, we modify the data structure of the influence percentage score data by converting the values in different rows of individual directors into the values in different columns – generating the data in which each observation contains one unique individual director’s yearly average influence percentage score (because we do not know the exact date of the controversial event, and according to our assumptions, we only need the directors’ average scores in different years) (Figure 4-2).

Figure 4-2

Screenshot: Modified Structure of Influence Percentage Scores

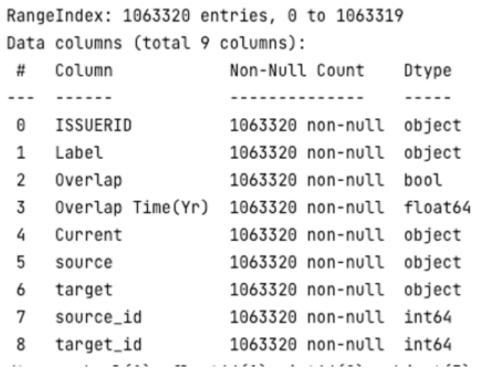


```
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   ISSUERID           31085 non-null  object
 1   ISSUER_NAME        31085 non-null  object
 2   INDIVIDUAL_ID      31085 non-null  int64
 3   FULLNAME           31085 non-null  object
 4   InfluencePCT_2017  15123 non-null  float64
 5   InfluencePCT_2018  20714 non-null  float64
 6   InfluencePCT_2019  20007 non-null  float64
 7   InfluencePCT_2020  21235 non-null  float64
 8   InfluencePCT_2021  20705 non-null  float64
 9   InfluencePCT_2022  13877 non-null  float64
```

On the other hand, for the network edge data (Figure 4-3) in which each observation represents the connection between two individual directors, we calculate the average values of overlap time (how long two directors have known each other) for each director and get the weighted connectivity for individual directors.

Figure 4-3

Screenshot: Structure of Connectivity Data



```
RangeIndex: 1063320 entries, 0 to 1063319
Data columns (total 9 columns):
 #   Column           Non-Null Count    Dtype
---  ------           --------------    -----
 0   ISSUERID         1063320 non-null  object
 1   Label            1063320 non-null  object
 2   Overlap          1063320 non-null  bool
 3   Overlap Time(Yr) 1063320 non-null  float64
 4   Current          1063320 non-null  object
 5   source           1063320 non-null  object
 6   target           1063320 non-null  object
 7   source_id        1063320 non-null  int64
 8   target_id        1063320 non-null  int64
```

After modifying the data structures, now, we need to combine all the necessary data sources together to get a final data set containing all the information on directors, companies, influence percentage, connectivity, and so on (only in this way, can we perform the feature engineering and scope narrowing processes in the following sections). The integration processes include the following steps:

Merge 1 – individual director data & company controversial count data. We chose the director data as our base data source as it has the greatest number of features (53) and left-join it with the company controversial count data on the primary key "Issuer_ID" (the unique identifier of companies). Thus, we have the data set with 156,507 observations of different individual directors (the observations of one director on multiple boards of different companies are regarded as different observations). Among these observations, there are 31,085 directors in 1,037 different companies which were recorded with controversial event counts.

Merge 2 – Merge 1 & influence percentage data. We left-join the Merge 1 data processed previously with the transformed influence percentage data. However, the primary key for this step is the combination of "Issuer_ID" and "Individual_ID". This is because we take the observations of one director on multiple boards of different companies are regarded as different observations and influence percentage data are transformed to the level of individual directors. Thus, we have the data set with 156,507 observations of different individual directors, and among these observations, 31,085 directors' companies have controversial event records, and 31,085 directors are with historical influence percentage values.

Merge 3 – Merge 2 and Network data. We left-join the Merge 2 data with the transformed network edge data (weighted connectivity) on the primary key "Individual_ID". Among all the observations, there are 16,582 directors with weighted connectivity values.

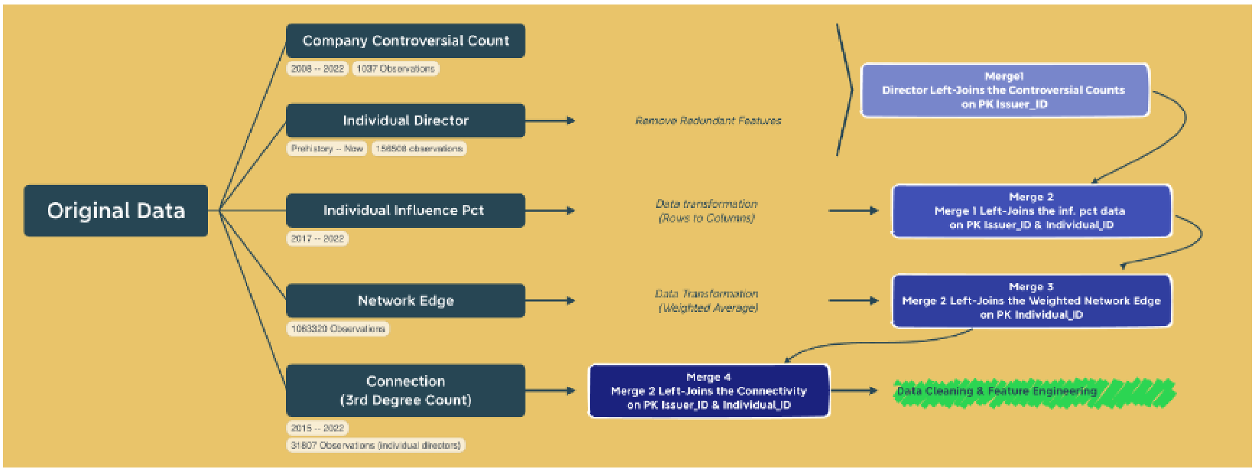
Merge 4 – Merge 3 and Connection data. Here, we left-join the Merge 3 data with the connection data (third-degree connection) on the primary key "Issuer_ID" and

“Individual_ID”. Among all the observations, there are 31,807 observations with third-degree connection values.

After all the merging processes above (Figure 4-4), the data set now has 156,507 observations of individual directors. Among these observations, 31,085 director records have controversial event values of their companies, 31,085 director records have historical influence percentage values, 16,582 directors have weighted connectivity values, and 31,807 directors have third-degree connection values (here, the same directors in different companies are regarded as different ones). The reason we did not apply inner join to merge data sources is that some directors may have other memberships on other companies’ boards and if we apply inner join to merge, that will only give us the overlapped observations of all data sources – causing us to lose the potential insights of our hypothesis. (But we will do this after the feature engineering to ensure we get every possible obtainable data of directors)

Figure 4-4

Mind Map of Data Integration


Original Data
Company Controversial Count
2008 -- 2022
1037 Observations
Individual Director
Prehistory -- Now
156508 observations
Individual Influence Pct
2017 -- 2022
Network Edge
1063320 Observations
Connection
(3rd Degree Count)
2015 -- 2022
31807 Observations (individual directors)
Remove Redundant Features
Data transformation
(Rows to Columns)
Data Transformation
(Weighted Average)
Merge1
Director Left-Joins the Controversial Counts
on PK Issuer_ID
Merge 2
Merge 1 Left-Joins the inf. pct data
on PK Issuer_ID & Individual_ID
Merge 3
Merge 2 Left-Joins the Weighted Network Edge
on PK Individual_ID
Merge 4
Merge 2 Left-Joins the Connectivity
on PK Issuer_ID & Individual_ID
Data Cleaning & Feature Engineering

Feature Engineering

The feature engineering of our research includes four parts: 1) generating the target variable (identifying directors were departed or stayed on the boards when controversial events happened); 2) identifying directors' multiple directorships across different companies; 3) identifying the historical and present identities (CEO, lead director, and chairman) of individual directors in different companies; 4) calculate the average influence percentage scores, average third-degree connections of individual directors, and total controversies of each company.

Target Variable

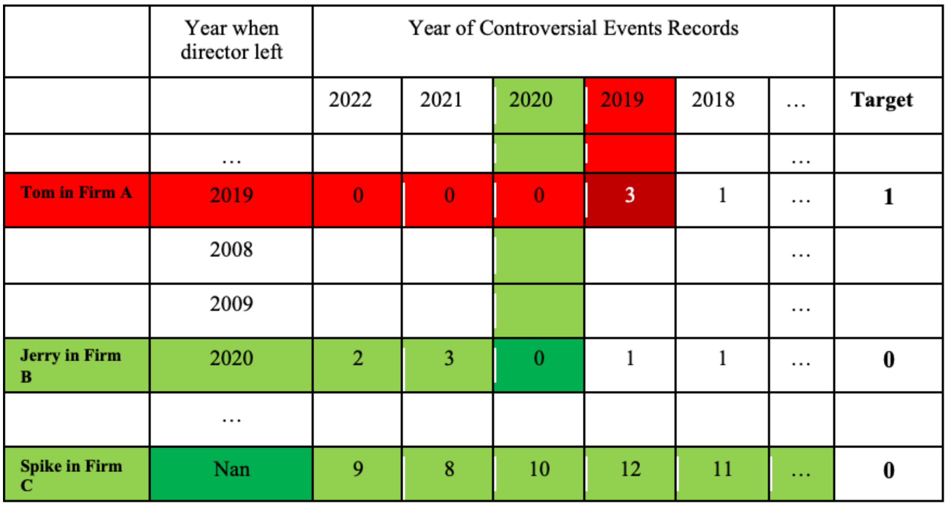
Under the assumptions we made, the logic to identify directors were departed or stayed on the boards when controversial events happened is:

when the company of director was recorded with one or more controversial events in a year (let's say 2022), we compare the year value of the date feature "Board_Member_Until" (the date when the director left the company) with the 2022 – if the year with controversial events matched the year when the director left, then this director would be regarded as "being departed from the board when the controversial event happens" and assigned with label "1"; otherwise, it would be assigned with label "0".

For example (Figure 4-5), Tom in Firm A left the company in 2019, and in that year, Tom's company had controversial events records –this observation of Tom will be identified as "departed, 1"; while for Jerry in Firm B left its company in 2020 but in that year, Jerry's company has no disclosed controversy being recorded – this observation of Jerry will be identified as "non-departed, 0" (it left from its own will); On the other hand, even though Spike's company had controversies recorded every year, but Spike never left the company board – it will also be identified as "non-departed, 0".

Figure 4-5

Chart Example of Identifying Target Variable

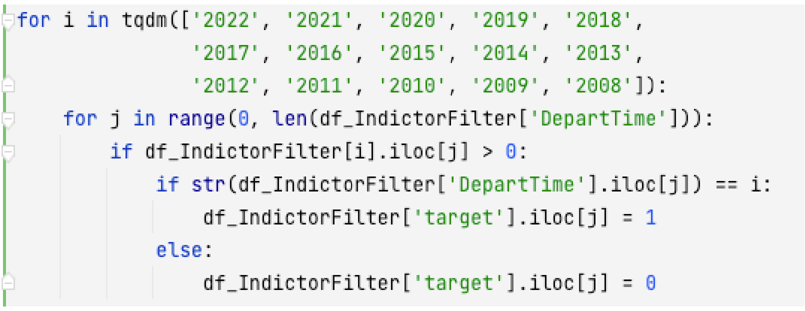


	Year when director left	Year of Controversial Events Records						
		2022	2021	2020	2019	2018	…	**Target**
	…						…	
Tom in Firm A	2019	0	0	0	3	1	…	**1**
	2008						…	
	2009						…	
Jerry in Firm B	2020	2	3	0	1	1	…	**0**
	…							
Spike in Firm C	Nan	9	8	10	12	11	…	**0**

This process will go through from 2008 to 2022 – all the years with controversial event records and compare it with the year when each director left. Thus, we obtain our target variable: whether the director departed or not when the controversial event happened. And of course, this process cannot be done with manual operation, so a loop function was developed, which can automatically generate the target labels (0 and 1) by the processes mentioned above in our target feature (Figure 4-6).

Figure 4-6

Screenshot: Loop for Generating Target Value



```
for i in tqdm(['2022', '2021', '2020', '2019', '2018',
               '2017', '2016', '2015', '2014', '2013',
               '2012', '2011', '2010', '2009', '2008']):
    for j in range(0, len(df_IndictorFilter['DepartTime'])):
        if df_IndictorFilter[i].iloc[j] > 0:
            if str(df_IndictorFilter['DepartTime'].iloc[j]) == i:
                df_IndictorFilter['target'].iloc[j] = 1
            else:
                df_IndictorFilter['target'].iloc[j] = 0
```

Note. Please see details in the attached Jupyter Notebook file, <Data Preprocessing, Primary Analysis, and Preparation>

In the end, among the 31,085 directors whose companies have controversial event records, there are 2,033 directors departed when their companies had controversial events. And one interesting finding is that every director was only departed once by their former companies, meaning they did not go back to the companies departing them or they went back and have not been departed until 2022. This may sound cliché, but it will serve an important role in our next data processing.

Multiple Directorships across Different Companies

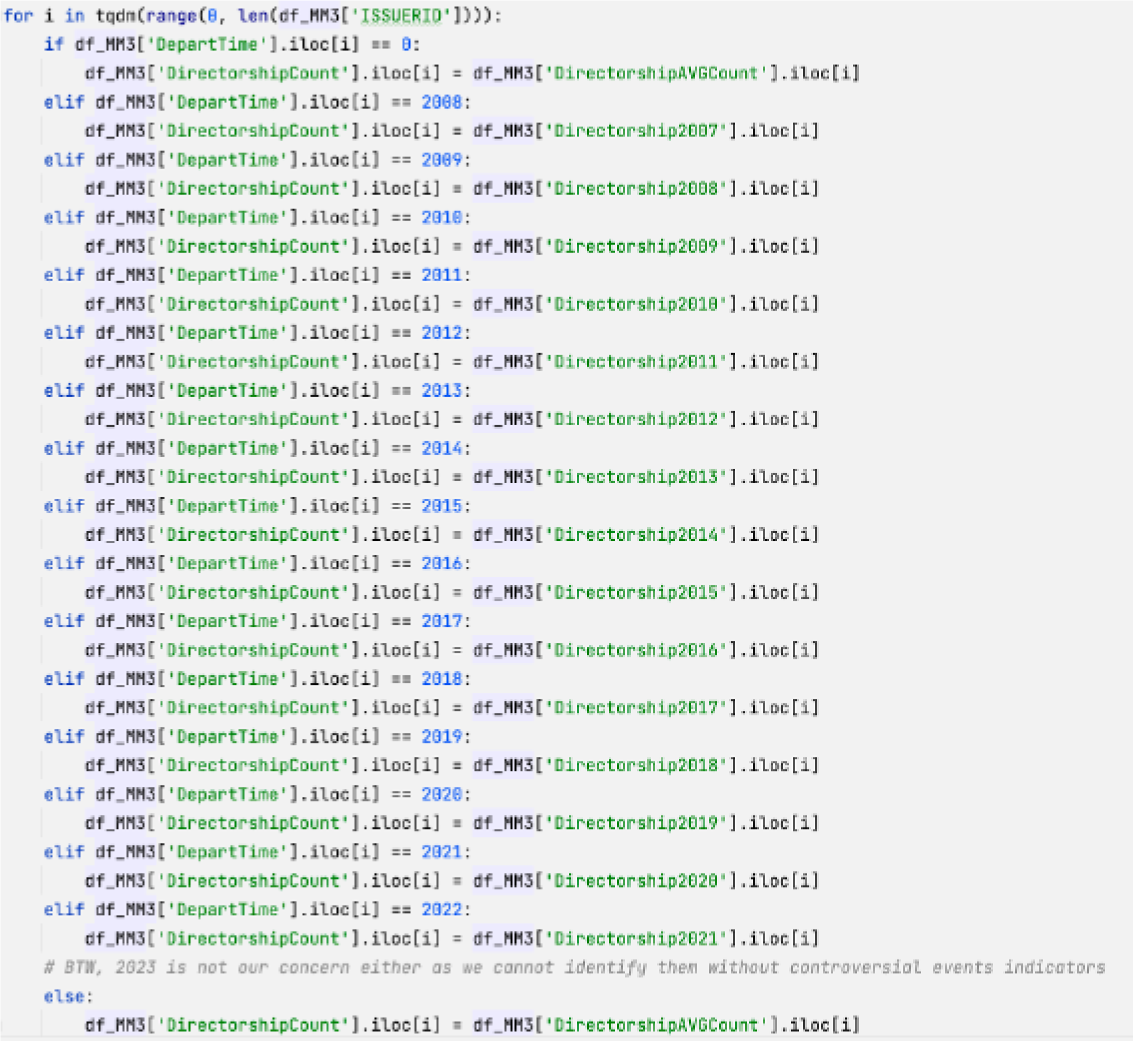
The second feature engineering we did is to identify how many directorships each individual director has across different companies. The logic is like the previous one but a little more complex:

Because the same director cannot have multiple observations of one company's board, the frequency of its Individual_ID is the number of its directorships in different companies. First, we separately filter the observations of directors who were (or are) still on the boards from 2007 to 2021 and count the frequencies of directors as the number of directorships in these years. Then, if one director was identified as "departed" in one year from 2008 to 2022, then we will assign the number of directorships before this year as the count of directorship when he/she got departed (one year ahead is because once the directors got departed, its directorship number would minus one (or more if he/she got departed for multiple times in that year)—but what we want is the number he/she was about to be departed, not after that); if the directors were identified as "non-departed" from 2008 to 2022, then we will assign their average number of directorships as their count of directorship.

Same as before, it is unpractical to manually complete the process above. Thus, another loop is created for this process (Figure 4-7)

Figure 4-7

Screenshot: Loop for Identify the Count of Directorships



```
for i in tqdm(range(0, len(df_MM3['ISSUERID']))):
    if df_MM3['DepartTime'].iloc[i] == 0:
        df_MM3['DirectorshipCount'].iloc[i] = df_MM3['DirectorshipAVGCount'].iloc[i]
    elif df_MM3['DepartTime'].iloc[i] == 2008:
        df_MM3['DirectorshipCount'].iloc[i] = df_MM3['Directorship2007'].iloc[i]
    elif df_MM3['DepartTime'].iloc[i] == 2009:
        df_MM3['DirectorshipCount'].iloc[i] = df_MM3['Directorship2008'].iloc[i]
    elif df_MM3['DepartTime'].iloc[i] == 2010:
        df_MM3['DirectorshipCount'].iloc[i] = df_MM3['Directorship2009'].iloc[i]
    elif df_MM3['DepartTime'].iloc[i] == 2011:
        df_MM3['DirectorshipCount'].iloc[i] = df_MM3['Directorship2010'].iloc[i]
    elif df_MM3['DepartTime'].iloc[i] == 2012:
        df_MM3['DirectorshipCount'].iloc[i] = df_MM3['Directorship2011'].iloc[i]
    elif df_MM3['DepartTime'].iloc[i] == 2013:
        df_MM3['DirectorshipCount'].iloc[i] = df_MM3['Directorship2012'].iloc[i]
    elif df_MM3['DepartTime'].iloc[i] == 2014:
        df_MM3['DirectorshipCount'].iloc[i] = df_MM3['Directorship2013'].iloc[i]
    elif df_MM3['DepartTime'].iloc[i] == 2015:
        df_MM3['DirectorshipCount'].iloc[i] = df_MM3['Directorship2014'].iloc[i]
    elif df_MM3['DepartTime'].iloc[i] == 2016:
        df_MM3['DirectorshipCount'].iloc[i] = df_MM3['Directorship2015'].iloc[i]
    elif df_MM3['DepartTime'].iloc[i] == 2017:
        df_MM3['DirectorshipCount'].iloc[i] = df_MM3['Directorship2016'].iloc[i]
    elif df_MM3['DepartTime'].iloc[i] == 2018:
        df_MM3['DirectorshipCount'].iloc[i] = df_MM3['Directorship2017'].iloc[i]
    elif df_MM3['DepartTime'].iloc[i] == 2019:
        df_MM3['DirectorshipCount'].iloc[i] = df_MM3['Directorship2018'].iloc[i]
    elif df_MM3['DepartTime'].iloc[i] == 2020:
        df_MM3['DirectorshipCount'].iloc[i] = df_MM3['Directorship2019'].iloc[i]
    elif df_MM3['DepartTime'].iloc[i] == 2021:
        df_MM3['DirectorshipCount'].iloc[i] = df_MM3['Directorship2020'].iloc[i]
    elif df_MM3['DepartTime'].iloc[i] == 2022:
        df_MM3['DirectorshipCount'].iloc[i] = df_MM3['Directorship2021'].iloc[i]
    # BTW, 2023 is not our concern either as we cannot identify them without controversial events indicators
    else:
        df_MM3['DirectorshipCount'].iloc[i] = df_MM3['DirectorshipAVGCount'].iloc[i]
```

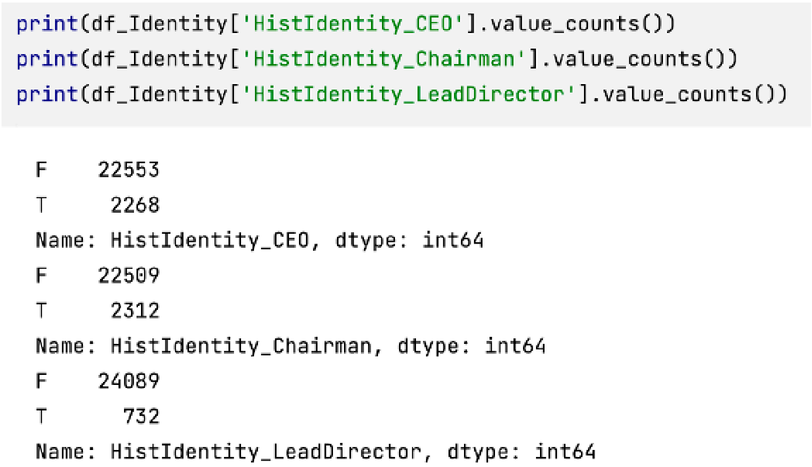
Note. Please see details in the attached Jupyter Notebook file, <Data Preprocessing, Primary Analysis, and Preparation>

Historical and Present Identities

Being the CEO or Chairman of the board usually indicates greater influence in the company. However, someone who used to be in these positions could also have that influence. Thus, instead of applying the present indicators of important identities in the company, we also want to identify if the directors were or are in these positions. By examining the not-null values of tenures of these positions, we can easily get the indicators of these identities that directors were or are.

Figure 4-8

Screenshot: Identified Historical and Present Identities



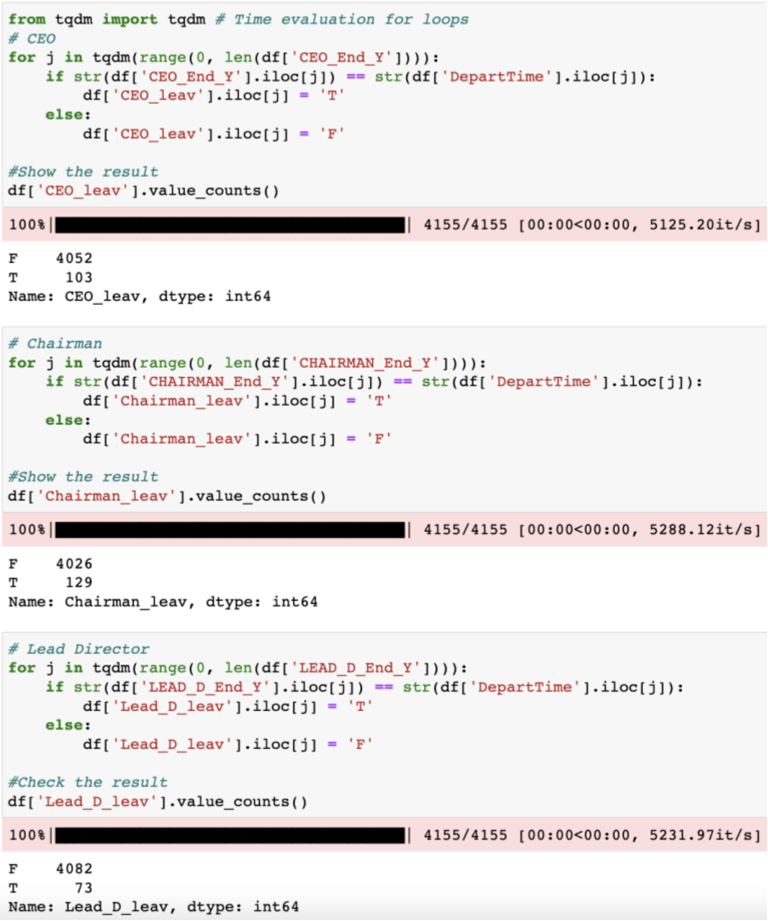
```
print(df_Identity['HistIdentity_CEO'].value_counts())
print(df_Identity['HistIdentity_Chairman'].value_counts())
print(df_Identity['HistIdentity_LeadDirector'].value_counts())
```

```
F    22553
T     2268
Name: HistIdentity_CEO, dtype: int64
F    22509
T     2312
Name: HistIdentity_Chairman, dtype: int64
F    24089
T      732
Name: HistIdentity_LeadDirector, dtype: int64
```

During the analysis, we discovered misrepresentation related to the CEO, CHAIRMAN, and LEAD_DIRECTOR columns with binary values True and False. Those columns were representing as True only those directors who are still on the board. This means that all other directors who also were CEO/Chairman/Lead Directors but are not currently on the board are indicated as False. Thus, we extracted the year values of each three positions ending periods (CEO_End_Y, CHAIRMAN_End_Y, LEAD_D_End_Y) which then were compared to the departing year (DepartTime). So, if the directors left the position in the same year they departed from the company, we assign T (True) for the new column (CEO_leav, Chairman_leav, Lead_D_leav), if not, or if they are still on the board, then F (False). This will help us to analyze whether the director was occupying the position while leaving the board, as our goal is to discover the factors that help to predict the director’s departure.

Figure 4-9

Screenshots: Loop for Identifying the Director's position while leaving the board



```
from tqdm import tqdm # Time evaluation for loops
# CEO
for j in tqdm(range(0, len(df['CEO_End_Y']))):
    if str(df['CEO_End_Y'].iloc[j]) == str(df['DepartTime'].iloc[j]):
        df['CEO_leav'].iloc[j] = 'T'
    else:
        df['CEO_leav'].iloc[j] = 'F'

#Show the result
df['CEO_leav'].value_counts()
```

```
100%|██████████████████████████████| 4155/4155 [00:00<00:00, 5125.20it/s]

F    4052
T     103
Name: CEO_leav, dtype: int64
```

```
# Chairman
for j in tqdm(range(0, len(df['CHAIRMAN_End_Y']))):
    if str(df['CHAIRMAN_End_Y'].iloc[j]) == str(df['DepartTime'].iloc[j]):
        df['Chairman_leav'].iloc[j] = 'T'
    else:
        df['Chairman_leav'].iloc[j] = 'F'

#Show the result
df['Chairman_leav'].value_counts()
```

```
100%|██████████████████████████████| 4155/4155 [00:00<00:00, 5288.12it/s]

F    4026
T     129
Name: Chairman_leav, dtype: int64
```

```
# Lead Director
for j in tqdm(range(0, len(df['LEAD_D_End_Y']))):
    if str(df['LEAD_D_End_Y'].iloc[j]) == str(df['DepartTime'].iloc[j]):
        df['Lead_D_leav'].iloc[j] = 'T'
    else:
        df['Lead_D_leav'].iloc[j] = 'F'

#Check the result
df['Lead_D_leav'].value_counts()
```

```
100%|██████████████████████████████| 4155/4155 [00:00<00:00, 5231.97it/s]

F    4082
T      73
Name: Lead_D_leav, dtype: int64
```

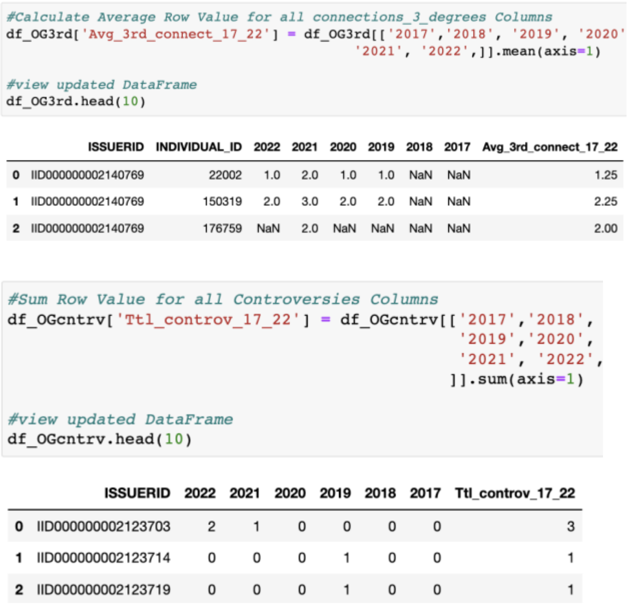
Note. Please see details in the attached Jupyter Notebook file, <2. Data Processing.ipynb>

IPS, Third-Degree Connection, and Controversies

The data of the influence percentage score (IPS), and the third-degree connection for each director as well as the controversies for each company are recorded at different time points. In the descriptive analysis and regression analysis, we can have only one feature for each of them. As a result, we calculate the average value for IPS and third-degree connection, as well as the total value for the controversies over the overlapping period of 2017 - 2022 to simplify our analysis. Even though it may cost us some insights into the influence of their changes, it is the most efficient way to cover them in the following processes.

Figure 4-10

Screenshots: Average Third-Degree Connections and Number of Total Involved Controversies



```
#Calculate Average Row Value for all connections_3_degrees Columns
df_OG3rd['Avg_3rd_connect_17_22'] = df_OG3rd[['2017','2018', '2019', '2020'
                                           '2021', '2022',]].mean(axis=1)

#view updated DataFrame
df_OG3rd.head(10)
```

	ISSUERID	INDIVIDUAL_ID	2022	2021	2020	2019	2018	2017	Avg_3rd_connect_17_22
0	IID000000002140769	22002	1.0	2.0	1.0	1.0	NaN	NaN	1.25
1	IID000000002140769	150319	2.0	3.0	2.0	2.0	NaN	NaN	2.25
2	IID000000002140769	176759	NaN	2.0	NaN	NaN	NaN	NaN	2.00

```
#Sum Row Value for all Controversies Columns
df_OGcntrv['Ttl_controv_17_22'] = df_OGcntrv[['2017','2018',
                                               '2019','2020',
                                               '2021', '2022',
                                              ]].sum(axis=1)

#view updated DataFrame
df_OGcntrv.head(10)
```

	ISSUERID	2022	2021	2020	2019	2018	2017	Ttl_controv_17_22
0	IID000000002123703	2	1	0	0	0	0	3
1	IID000000002123714	0	0	0	1	0	0	1
2	IID000000002123719	0	0	0	1	0	0	1

Note. Please see details in the attached Jupyter Notebook file, <2. Data Processing.ipynb>

On the other hand, we also decided to apply the secondary approach to deal with the IPS feature to compare its effect of exact value on the departure results. IPS, as mentioned before, is the attribute created by our sponsor to tell "who should be responsible?". At this point, we have found out who took the consequences of controversial events (departure results from target variable identification) and we applied the average value to evaluate IPSs of individual directors at different time points. However, even though this method is efficient, it could cost us some insights of how exact IPS values when directors were departed influence the departure results. Thus, we raised the second data structure to deal with the IPS data.

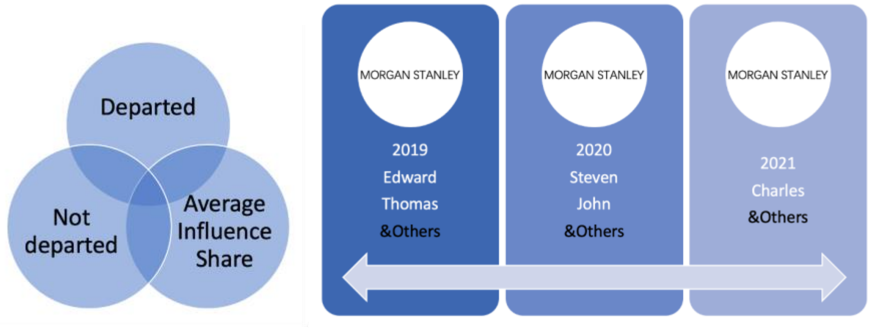
The logic of second approach of dealing with IPS data is that:

We take each observations of individual directors on different company boards as multal exclusive cases. For example, if director A was identified as "departed" in 2020, we extract its IPS value and IPS values of other "non-departed" directors on the same board in 2020 (IPS here is yearly average value that we processed before in the Data Integration section) as a group of observations; Similarly, if director B was identified as "departed" in

2019, we extract its IPS value and IPS values of other "non-departed" directors in 2019 as another group of observations.

Figure 4-11

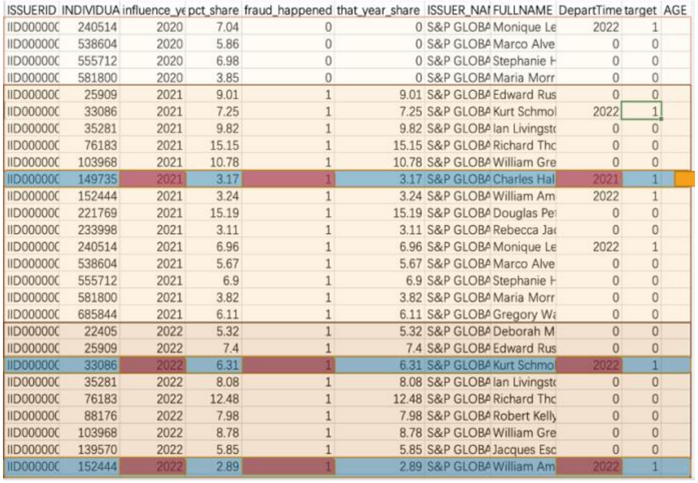
Logic of Second Approach to Deal with IPS Data:



Take S&P Global data for example (Table 4-1). In 2019, two directors, Kurt and William were identified as "departed" from S&P Global. Thus, we extracted the 2019 IPS data of directors on the board of S&P Global; while in 2022, another director, Monique, in S&P Global was identified as "departed", thus, we extracted the 2022 IPS data of directors on the board for another group of observations.

Table 4-1

Sample Data of S&P Global Directors in the Second Approach of IPS Data



ISSUERID	INDIVIDUA	influence_ye	pct_share	fraud_happened	that_year_share	ISSUER_NAI	FULLNAME	DepartTime	target	AGE
IID00000C	240514	2020	7.04	0	0	S&P GLOBA	Monique Le	2022	1	
IID00000C	538604	2020	5.86	0	0	S&P GLOBA	Marco Alve	0	0	
IID00000C	555712	2020	6.98	0	0	S&P GLOBA	Stephanie H	0	0	
IID00000C	581800	2020	3.85	0	0	S&P GLOBA	Maria Morr	0	0	
IID00000C	25909	2021	9.01	1	9.01	S&P GLOBA	Edward Rus	0	0	
IID00000C	33086	2021	7.25	1	7.25	S&P GLOBA	Kurt Schmol	2022	1	
IID00000C	35281	2021	9.82	1	9.82	S&P GLOBA	Ian Livingstc	0	0	
IID00000C	76183	2021	15.15	1	15.15	S&P GLOBA	Richard Thc	0	0	
IID00000C	103968	2021	10.78	1	10.78	S&P GLOBA	William Gre	0	0	
IID00000C	149735	2021	3.17	1	3.17	S&P GLOBA	Charles Hal	2021	1	
IID00000C	152444	2021	3.24	1	3.24	S&P GLOBA	William Am	2022	1	
IID00000C	221769	2021	15.19	1	15.19	S&P GLOBA	Douglas Pe	0	0	
IID00000C	233998	2021	3.11	1	3.11	S&P GLOBA	Rebecca Ja	0	0	
IID00000C	240514	2021	6.96	1	6.96	S&P GLOBA	Monique Le	2022	1	
IID00000C	538604	2021	5.67	1	5.67	S&P GLOBA	Marco Alve	0	0	
IID00000C	555712	2021	6.9	1	6.9	S&P GLOBA	Stephanie H	0	0	
IID00000C	581800	2021	3.82	1	3.82	S&P GLOBA	Maria Morr	0	0	
IID00000C	685844	2021	6.11	1	6.11	S&P GLOBA	Gregory Wa	0	0	
IID00000C	22405	2022	5.32	1	5.32	S&P GLOBA	Deborah M	0	0	
IID00000C	25909	2022	7.4	1	7.4	S&P GLOBA	Edward Rus	0	0	
IID00000C	33086	2022	6.31	1	6.31	S&P GLOBA	Kurt Schmol	2022	1	
IID00000C	35281	2022	8.08	1	8.08	S&P GLOBA	Ian Livingstc	0	0	
IID00000C	76183	2022	12.48	1	12.48	S&P GLOBA	Richard Thc	0	0	
IID00000C	88176	2022	7.98	1	7.98	S&P GLOBA	Robert Kelly	0	0	
IID00000C	103968	2022	8.78	1	8.78	S&P GLOBA	William Gre	0	0	
IID00000C	139570	2022	5.85	1	5.85	S&P GLOBA	Jacques Esc	0	0	
IID00000C	152444	2022	2.89	1	2.89	S&P GLOBA	William Am	2022	1	

Data Cleaning

After integrating all the needed data sources, now we have over 156,507 observations of individual directors along with their companies. However, due to the limits of the controversy records scope, only 31,085 directors were identified as “departed” or “non-departed”; on the other hand, there are a lot of missing values in features like age, genders, identities, and so on and some categorical features is extremely imbalanced. Thus, this section will demonstrate the data-cleaning process.

Scope Narrowing

As mentioned before, the data scope for the final analysis and model building should be the directors whose companies have controversial event records; also, controversial data range from 2008 to 2022, and director data ranges from “prehistory” to 2023 – but the data of influence percentage score only contains the records from 2017 to 2022, which means that only the data overlapped period can offer us all the data we need to test our hypothesis and build models. Thus, to obtain the data set that we can put into our analysis and models, we need to narrow down the scope of previously processed data with these two conditions.

Narrowing by the first condition is easy: we only need to filter all the records with Nan values in our target variable because they were not involved in the target-generating process which means these observations are the directors whose companies do not have controversial event records. On the other hand, the second condition is more complex than the previous one. However, with our interesting finding: every director got departed only ONCE, we can conclude that directors who were departed before 2017 are not our concern: they cannot be back in the business (at least in the same company he/she got departed). Is that all? Yes! This is because the directors who were departed from companies before 2017 also do not have influence percentage values. Thus, we only need to filter the observation who got departed before 2017 and the rest of the director observations are either always on the boards or were departed from 2017 to 2022.

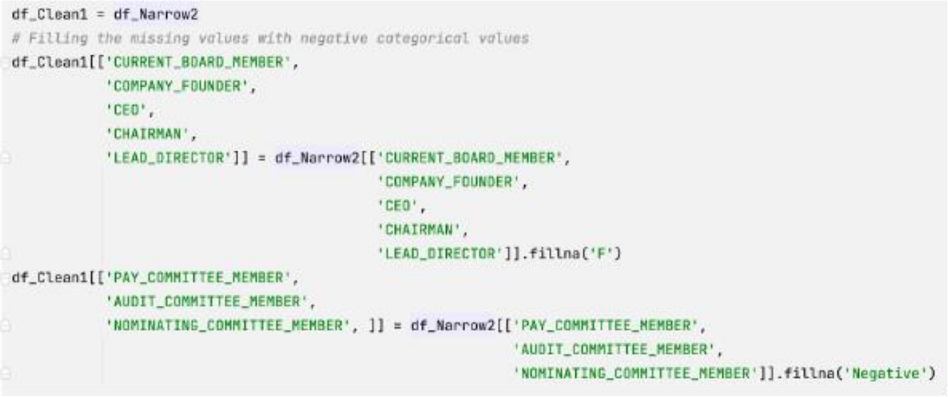
Thus, after narrowing down our data scope, we have 24,812 observations of directors, 2,033 of these directors got departed due to the companies' controversial events from 2017 to 2022.

Missing Values

Missing Values for Categorical Variables. The missing values mainly exist in the categorical variables, such as “CURRENT_BOARD_MEMBER”, “CEO (or not)”, “PAY_COMMITTEE_MEMBER”, and so on. These values are missed simply because these directors are not in these positions, so these values are left blank in the original data. Thus, we replace these missing values with assigned categorical values: “F” or “Negative” (Figure 4-9). On the other hand, the missing values in “GENDER” may be because directors are not willing to identify their genders. Thus, for the consideration of simplifying our research without manually checking the missing genders, we will assign the missing genders the categorical value, “Unknown” (Figure 4-10).

Figure 4-9

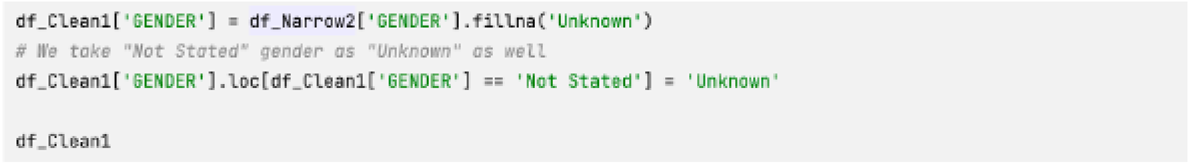
Screenshot: Replacing the Missing Values in Position Features



```
df_Clean1 = df_Narrow2
# Filling the missing values with negative categorical values
df_Clean1[['CURRENT_BOARD_MEMBER',
           'COMPANY_FOUNDER',
           'CEO',
           'CHAIRMAN',
           'LEAD_DIRECTOR']] = df_Narrow2[['CURRENT_BOARD_MEMBER',
                                           'COMPANY_FOUNDER',
                                           'CEO',
                                           'CHAIRMAN',
                                           'LEAD_DIRECTOR']].fillna('F')
df_Clean1[['PAY_COMMITTEE_MEMBER',
           'AUDIT_COMMITTEE_MEMBER',
           'NOMINATING_COMMITTEE_MEMBER', ]] = df_Narrow2[['PAY_COMMITTEE_MEMBER',
                                                            'AUDIT_COMMITTEE_MEMBER',
                                                            'NOMINATING_COMMITTEE_MEMBER']].fillna('Negative')
```

Figure 4-10

Screenshot: Replacing the Missing Values in Genders



```
df_Clean1['GENDER'] = df_Narrow2['GENDER'].fillna('Unknown')
# We take "Not Stated" gender as "Unknown" as well
df_Clean1['GENDER'].loc[df_Clean1['GENDER'] == 'Not Stated'] = 'Unknown'

df_Clean1
```

Missing Values for Ages. The feature “AGE” contains a lot of zero values which are obviously mistakenly recorded values. Thus, we take these zero values as missing values, too. And considering different genders may have different distributions of ages, we first apply a box plot (Figure 4-11) to check the distribution of ages grouped by different genders. Because the distributions all have outliers and skewness, we use median values to separately replace the missing values in “AGE” (Figure 4-12).

Figure 4-11

Boxplot: Age Distributions Grouped by Genders

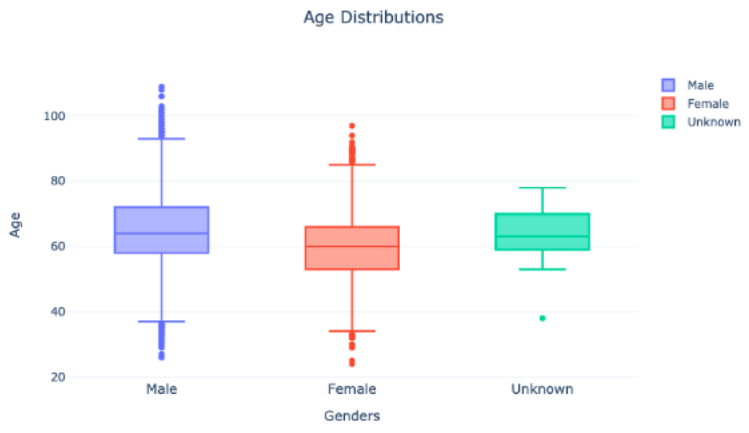
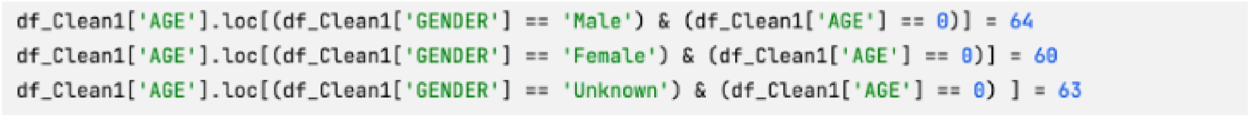


Figure 4-12

Screenshot: Replacing the Missing Values in Ages of Different Genders



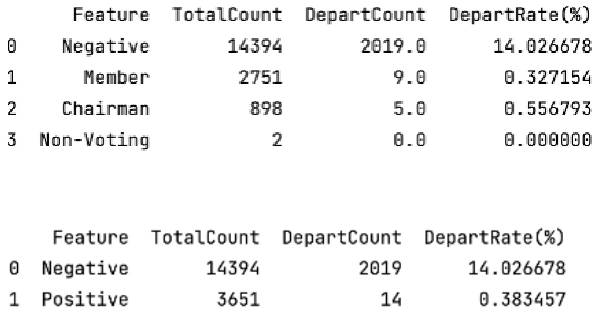
```
df_Clean1['AGE'].loc[(df_Clean1['GENDER'] == 'Male') & (df_Clean1['AGE'] == 0)] = 64
df_Clean1['AGE'].loc[(df_Clean1['GENDER'] == 'Female') & (df_Clean1['AGE'] == 0)] = 60
df_Clean1['AGE'].loc[(df_Clean1['GENDER'] == 'Unknown') & (df_Clean1['AGE'] == 0) ] = 63
```

Imbalanced Categorical Features

Positions in Committees: As shown in Figure 4-13, most directors are not in the pay, audit, or nominating committees, and some positions in these committees only have a few directors. Thus, it is better to convert the committee features into binary categories: “negative” and “positive” to reduce the imbalance.

Figure 4-13

Screenshot: Distribution of Original & Transformed Audit Committee Member Positions

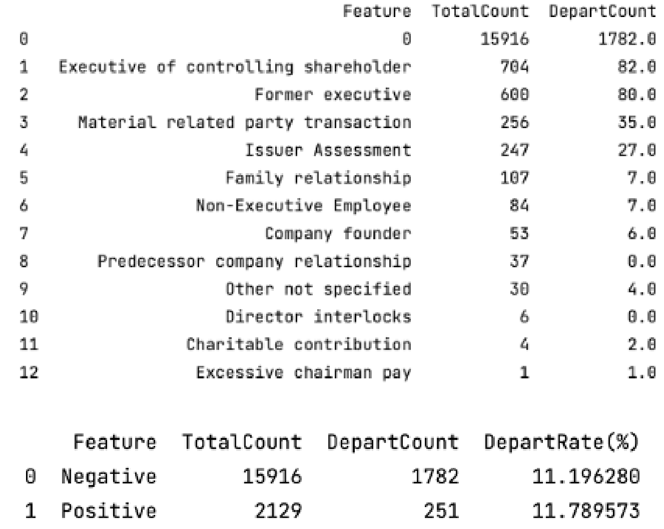


```
     Feature  TotalCount  DepartCount  DepartRate(%)
0   Negative       14394       2019.0      14.026678
1     Member        2751          9.0       0.327154
2   Chairman         898          5.0       0.556793
3  Non-Voting          2          0.0       0.000000


    Feature  TotalCount  DepartCount  DepartRate(%)
0  Negative       14394         2019      14.026678
1  Positive        3651           14       0.383457
```

Outside-Related Reasons: similar to the committee features, the features of outside-related reasons are also extremely imbalanced. Here we also convert the two features into binary categories to reduce the imbalance (Figure 4-14).

Figure 4-14

Screenshot: Distributions of Original and Transformed Outside-Related Reasons

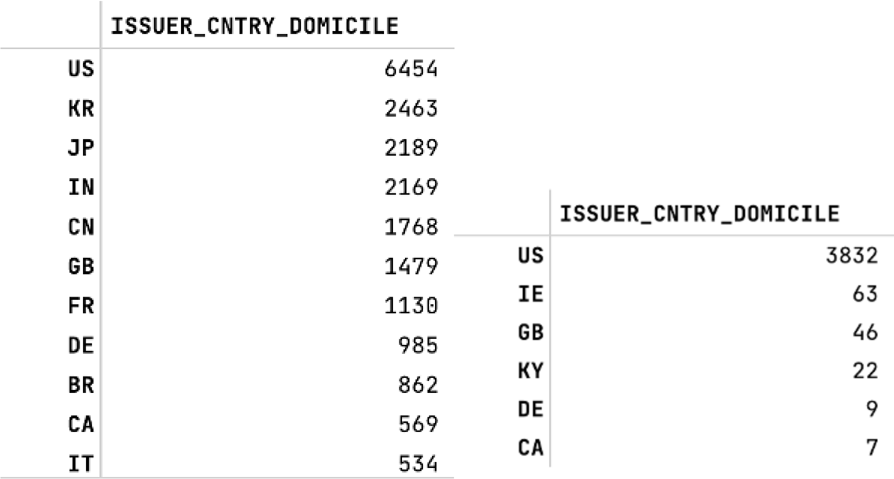


```
                               Feature  TotalCount  DepartCount
0                                    0       15916       1782.0
1   Executive of controlling shareholder        704         82.0
2                     Former executive         600         80.0
3     Material related party transaction        256         35.0
4                    Issuer Assessment         247         27.0
5                  Family relationship         107          7.0
6               Non-Executive Employee          84          7.0
7                      Company founder          53          6.0
8        Predecessor company relationship       37          0.0
9                  Other not specified          30          4.0
10                  Director interlocks          6          0.0
11              Charitable contribution          4          2.0
12               Excessive chairman pay          1          1.0

    Feature  TotalCount  DepartCount  DepartRate(%)
0  Negative       15916         1782      11.196280
1  Positive        2129          251      11.789573
```

US-Issued Companies: Even though we have narrowed down our research scope by the overlapped period of integrated data sources, there are some issues with this data: 1) the countries in which these companies were issued are extremely imbalanced (the US has 6,454 companies, taking up about one-fourth of all observations, while the total number of Korean (2,463), Japanese(2,189), and Indian (2,169) companies is less than 8,000 and many countries are even less than 1,000). 2) IPS and the weighted connectivity data are mainly collected for the US companies (Figure 4-9). Thus, the following analysis will focus on the observations whose companies were issued in the US.

Figure 4-15

Screenshot: Observation Frequencies of Issued Countries and IPS



	ISSUER_CNTRY_DOMICILE
US	6454
KR	2463
JP	2189
IN	2169
CN	1768
GB	1479
FR	1130
DE	985
BR	862
CA	569
IT	534

	ISSUER_CNTRY_DOMICILE
US	3832
IE	63
GB	46
KY	22
DE	9
CA	7

In the end, we obtained 4,155 unique directors in different companies which are issued in the United States.

Note. For the Descriptive Analysis and Predictive Models Part 1, the observations are the 4,155 directors in US-Issued companies; however, for the Predictive Models Part 2, the observations increases to 35,215 as we applied a different approach to deal with IPS data in order to obtain the very IPS data when directors were departed.

Descriptive Analysis

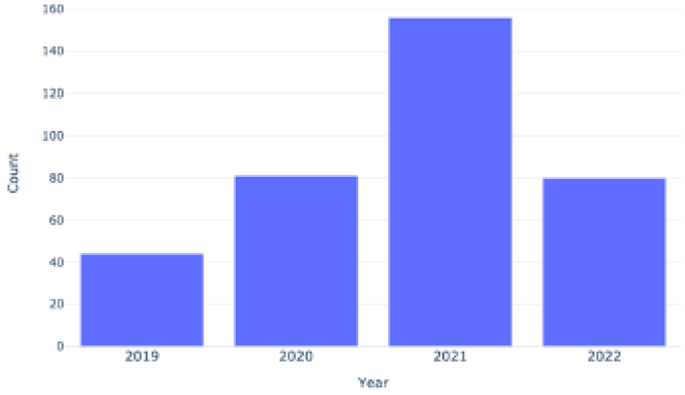
This section will mainly apply statistical visualizations to demonstrate the departure distributions grouped by different features, offering references for the following modeling process.

Historical Departure Frequency

The total number of departed directors from 2007 to 2022 is shown in Figure 5-1. We have the following findings: 1) there are no departed directors before 2019 which is because the original controversial event count data have no records from 2014 to 2018 and also, we remove the observations in which directors were departed before 2017; 2) from 2019 to 2021, the number of departures in US-issued companies is in the inclining trend, and even though we lack a part of data for 2022, we can expect that the departure amount in 2022 would keep increasing.

Figure 5-1

Bar Chart: Distribution of Yearly Total Counts of Departures in US-Issued Companies.

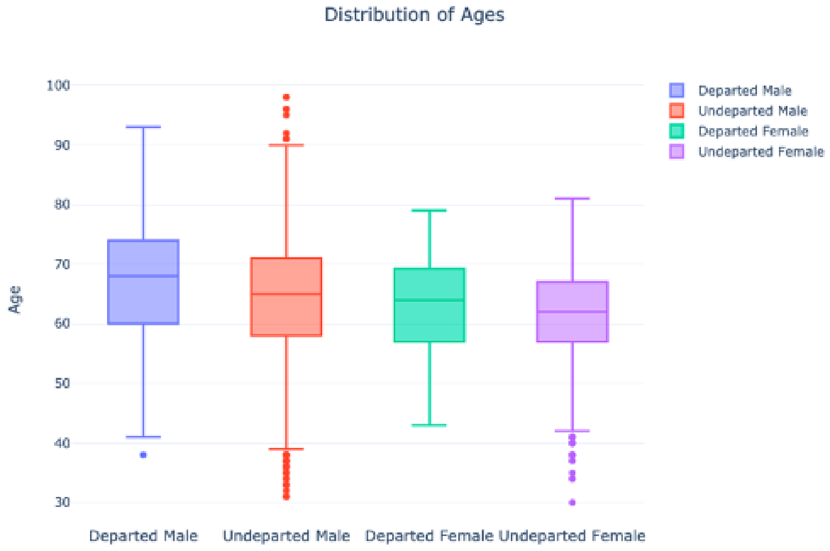


Departure with Age & Genders

From the box plot (Figure 5-2), we can conclude that 1) for both males and females, departures happened more frequently in the older group (median departed age – male 68, female 64; median non-departed age – male 65, female 62); 2) the age distribution of male is "older" than the age distribution of female; 3) the outliers mainly exist in the young-age and undeparted distribution (most of them are under 40) while the departed group hardly have the directors who are under 40. This indicates that directors under 40 were very rare to be departed from US-issued companies from 2019 to 2022.

Figure 5-2

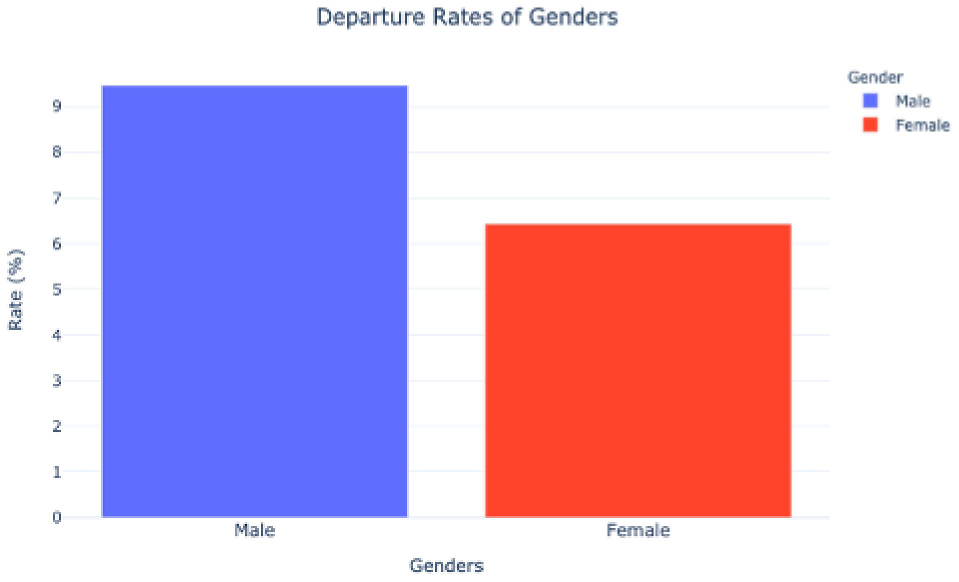
Box Plot: Distributions of Directors Ages Grouped by Genders & Departure



On the other hand, male directors have higher departure rates (9.47%) than female directors (6.43%) in the US-issued companies from 2019 to 2022 (Figure 5-3).

Figure 5-3

Bar Chart: Departure Rates of Different Genders



Departure with Committee Members

To evaluate if the departures during the controversial events are influenced by the committee positions of directors, such as pay committee members or not, and so on, we applied the bar chart to visualize the departure rates of different committee positions (Chart 5-1). According to the distributions of departure rates for different positions, we can conclude that in US-issued companies, directors who are NOT members of the pay committee, audit committee, or nominating committee are much more likely to be departed when controversial events happen.

Table 5-1

Departure Rates Grouped by Pay, Audit, Nominating Committees

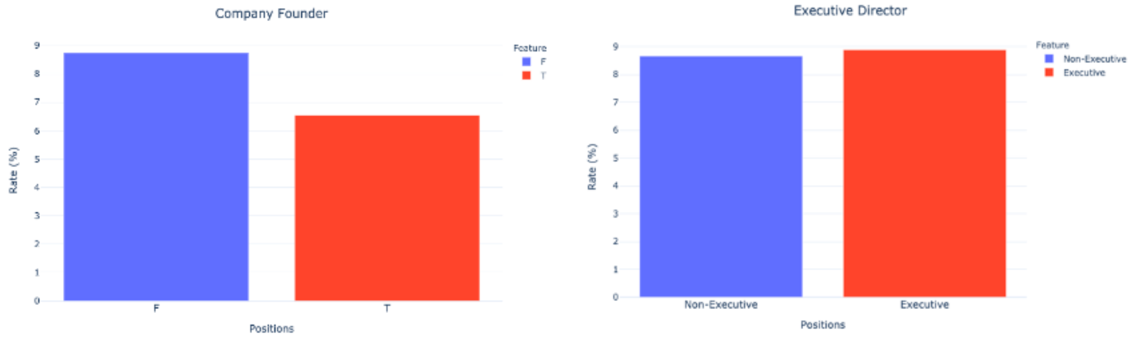
	Departure Rates	
	Yes	No
Pay Committee	0.10%	11.48%
Audit Committee	0%	11.69%
Nominating Committee	0.39%	11.35%

Departure with Identities

The identities are considered the capital of directors on the board, such as the company founder, CEO, executive director, and so on. From the bar charts of departure rates (Figure 5-3), we can conclude that directors who are company founders are less likely to be departed when controversial events happen; however, executive directors and non-executive directors (have close departure rates (8.89% and 8.65% respectively).

Figure 5-3

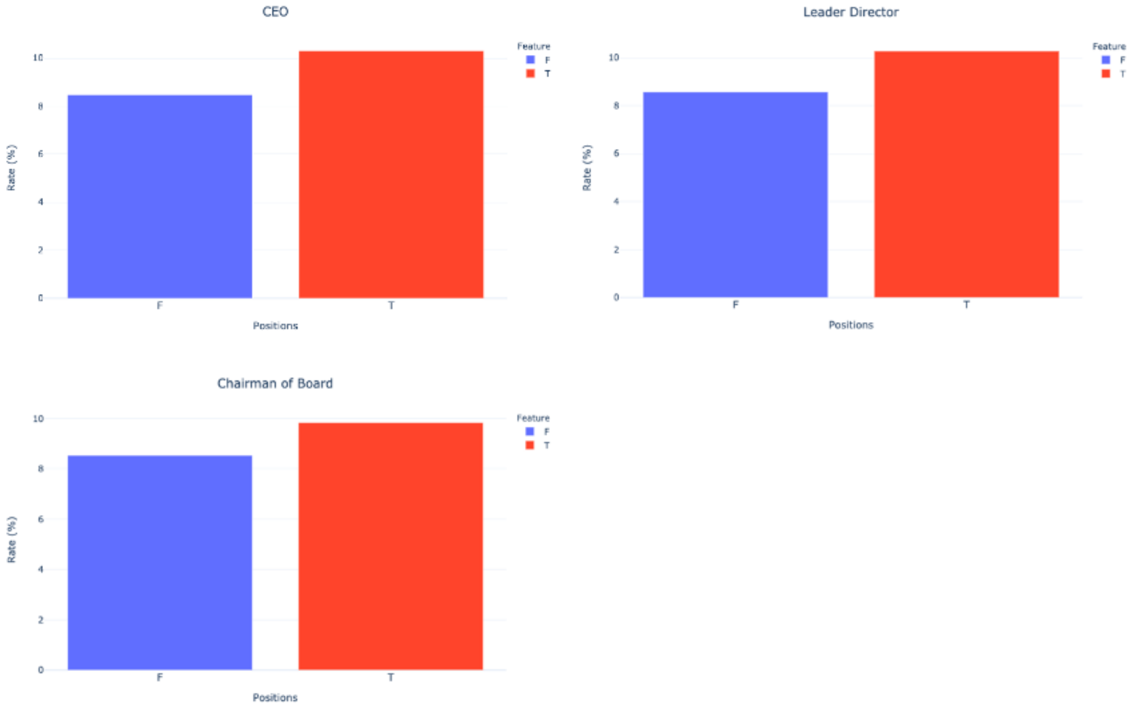
Bar Charts: Departure Rate Distributions of Company Founders and Executive Directors



On the other hand, the directors who were or are CEO, lead directors, or chairmen of boards, are more likely to be departed than the directors who have never been in these positions during the controversial events (Figure 5-4). This can be explained that these directors are the people who have or had the main responsibility for the management of their companies. And when controversial events were disclosed to the public, they are the first to be blamed and regarded as “unfulfilling their duties”.

Figure 5-4

Bar Charts: Departure Rate Distributions of CEO, Lead Director, and Chairman of Board

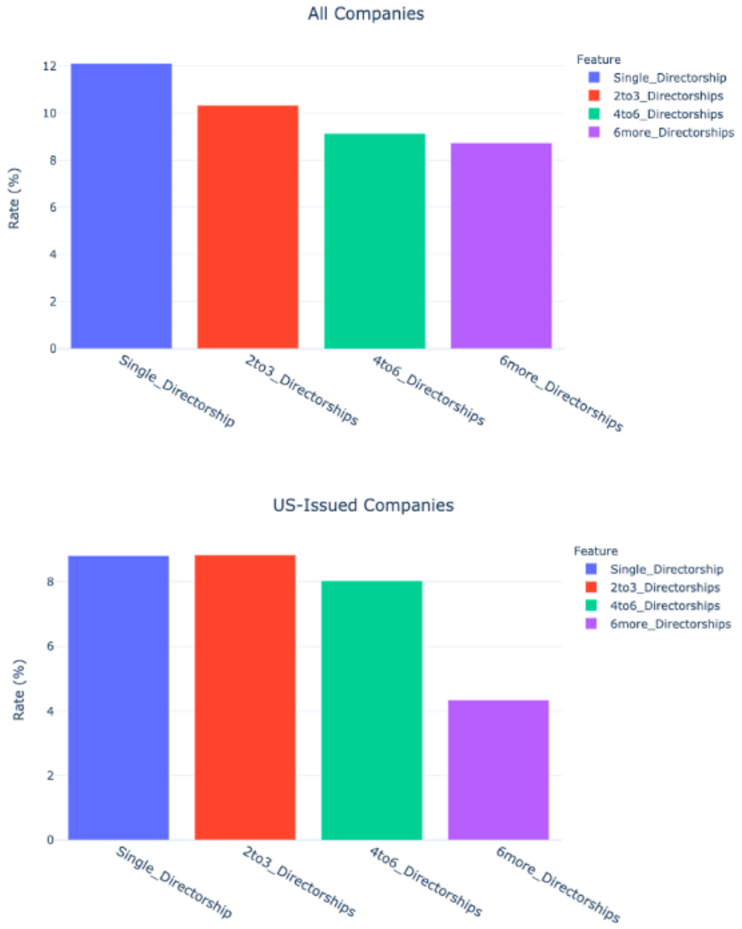


Departure with Number of Directorships

Another capital of board directors is the number of directorships across different companies: the more directorships they have, the more resources they own. From the bar charts of the departure rates of different levels of directorship counts (Figure 5-5), we can observe that for all companies (not limited to the US-issued), the more directorships, the less chance to be departed; for the US-issued companies, even though the distribution is not in an exact declining trend, we can still say that directors with more than 3 directorships are less likely to be departed than the directors with less than or equal to 3 directorships, especially for the directors who have more than 6 directorships across different companies.

Figure 5-5

Bar Charts: Departure Rate Distributions Grouped by Directorship Count Levels

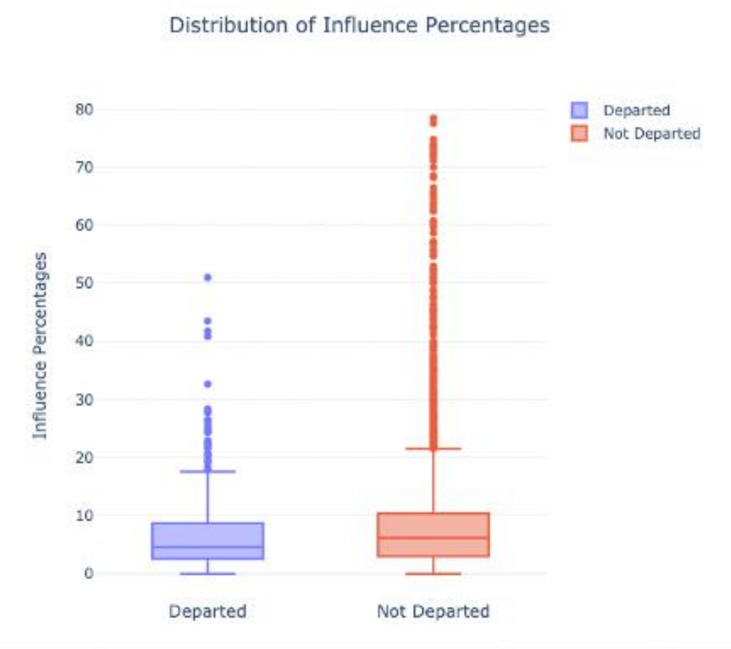


Departure with Influence Percentage

The Influence percentage score is the attribute generated by our sponsor, which quantitively measures how influential a director is on the company board. From the box plot (Figure 5-6), we can observe that the distribution of non-departed directors' influence percentage scores is higher than the departed directors'; on the other hand, both groups are having a significant number of outliers outside the right tail and the non-departed group is more “severe”, indicating that the directors with abnormally high influence percentage scores are less likely to be departed when the controversial events happen.

Figure 5-6

Box Plot: Influence Percentage Score Grouped by Departures

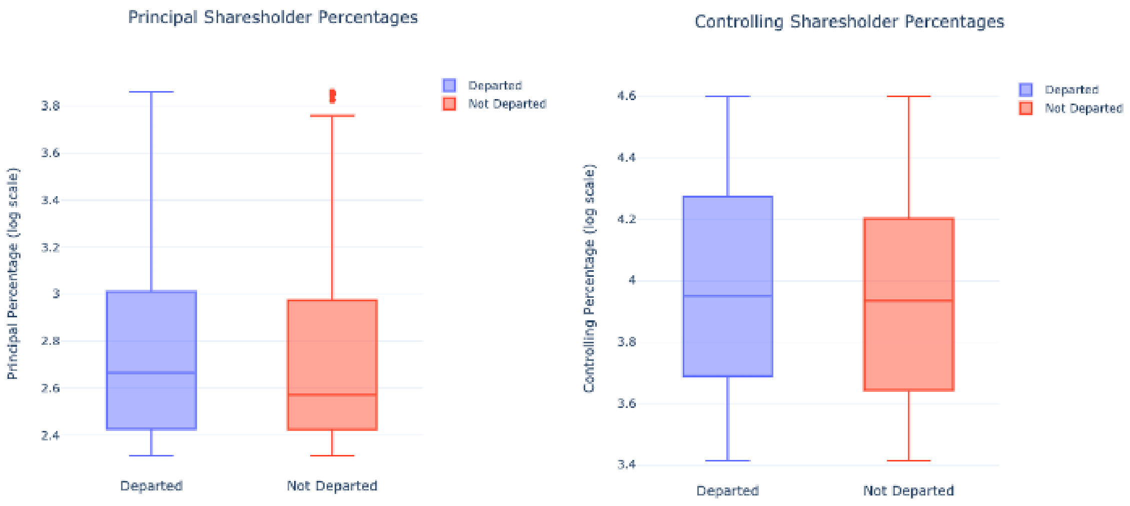


Departure with Shareholder Percentages

Shareholder percentages are another quantitive method to evaluate the influence of directors. From the box plots (Figure 5-7), we can say that directors with higher principal shareholder percentages are more likely to be departed. On the other hand, the distributions of controlling shareholder percentages of departed and non-departed groups are very close and do not have much difference.

Figure 5-7

Box Plots: Distributions of Principal & Controlling Shareholder Percentages

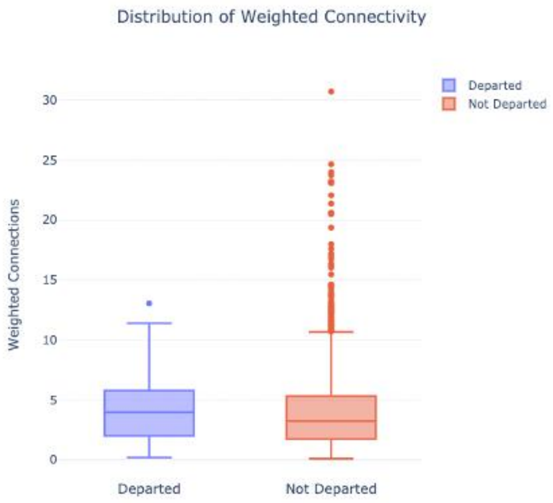


Departure with Weighted Network Connectivity

The last capital of the directors in the research is network connectivity. Similar to the number of directorships, the more connectivity a director has, the more resources he/she may own for the company. Here, we apply the weighted network connectivity values to compare the distribution difference between the departed group and the non-departed group (Figure 5-10). The most interesting finding is from the outliers: even though the box body of the departed group is higher than that of the non-departed group (higher connectivity with more chance of departure), however, the significant number of outliers in the non-departed group shows that directors with abnormally higher weighted connectivity are less to be departed.

Figure 5-8

Box Plots: Distributions Weighted Network Connectivity



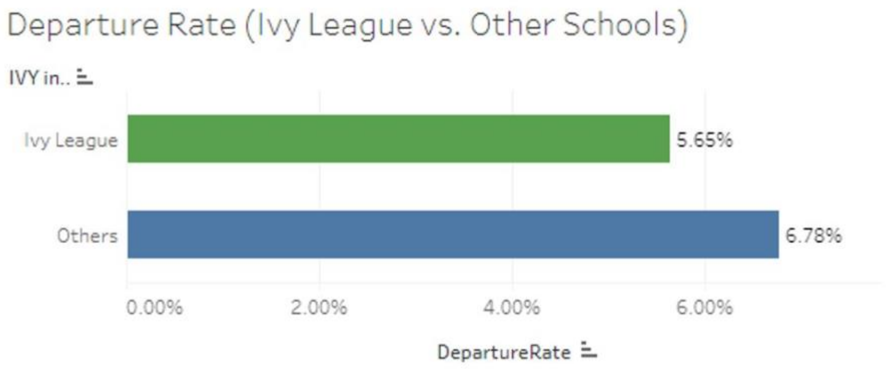
Departure with IVY League Backgrounds

The last descriptive analysis is the study of how different educational backgrounds can affect an individual's departure from the company. Because there is no perfectly weighted rank of college in the education system, it is not possible to compare the quality of education and level of educational background. However, Ivy League schools are publicly recognized elite education institutions. Therefore, we decided to use Ivy League schools as our indicator for higher education levels. We compare the difference between people who graduated from Ivy League schools and those who graduated from other schools and find the difference in the rate of departure.

From the bar plot (Figure 5-9), of 151 people who were departed from the company, 38 graduated from Ivy League schools and 113 graduated from other schools. Compared to the total graduates from each group, Ivy Leagues school graduates have a slightly lower departure rate (5.65%) compared to other school graduates' departure rate (6.78%).

Figure 5-9

Bar Plot: Departure Rates of Graduates from Ivy League and Other Schools

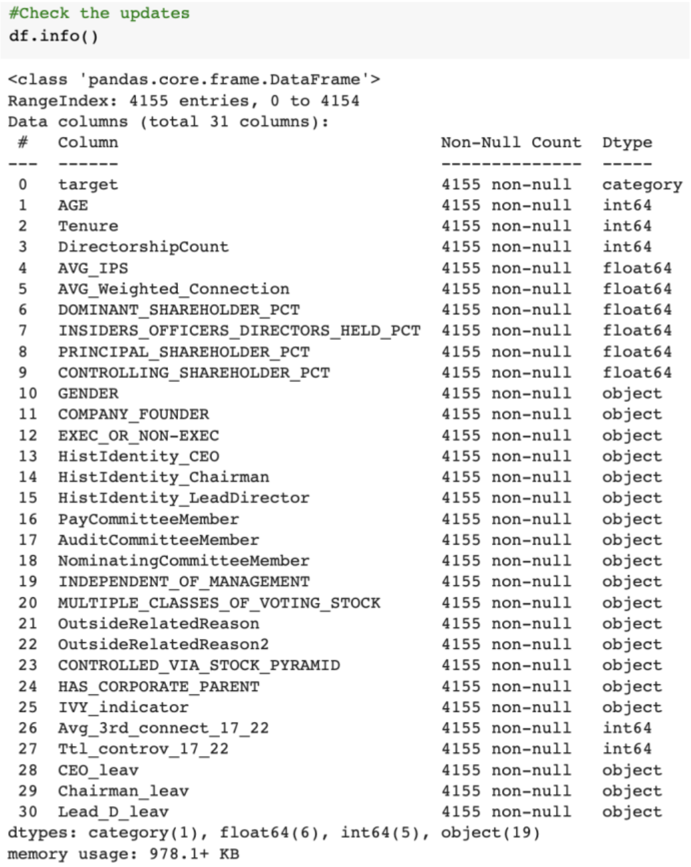


Predictive Models Part 1

This section represents the first approach of modeling using the same dataset presented throughout the analysis. This includes two methods, building multiple models using Auto Machine Learning (ML) technique, and regression analysis by the logistics regression model. For both methods, unnecessary features were removed. This includes anything related to IDs, names, education, and any other features related to the company level. Moreover, models were built using 30 features, the target variable, and 4,155 records (Figure 6-1). Out of 30 features, 11 of them were numerical, and 19 were categorical.

Figure 6-1

Screenshot: Cleaned and Ready for Modeling Dataset Information



```
#Check the updates
df.info()
```

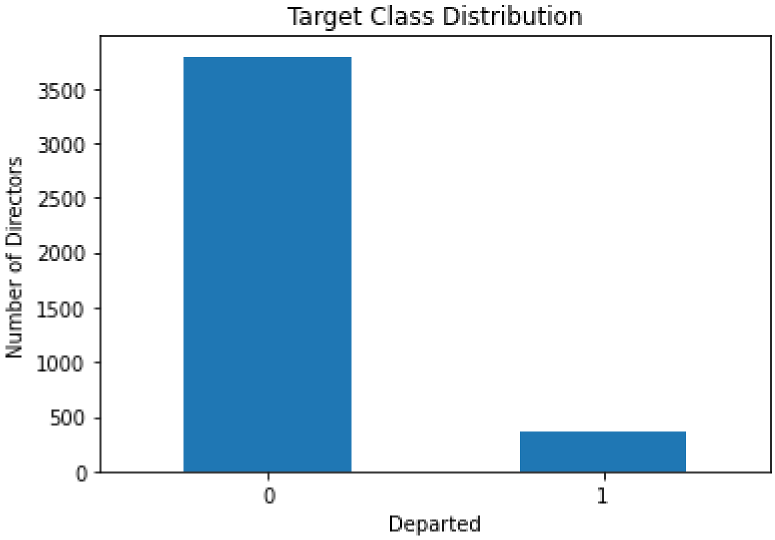
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4155 entries, 0 to 4154
Data columns (total 31 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   target                             4155 non-null   category
 1   AGE                                4155 non-null   int64
 2   Tenure                             4155 non-null   int64
 3   DirectorshipCount                  4155 non-null   int64
 4   AVG_IPS                            4155 non-null   float64
 5   AVG_Weighted_Connection            4155 non-null   float64
 6   DOMINANT_SHAREHOLDER_PCT           4155 non-null   float64
 7   INSIDERS_OFFICERS_DIRECTORS_HELD_PCT  4155 non-null   float64
 8   PRINCIPAL_SHAREHOLDER_PCT          4155 non-null   float64
 9   CONTROLLING_SHAREHOLDER_PCT        4155 non-null   float64
 10  GENDER                             4155 non-null   object
 11  COMPANY_FOUNDER                    4155 non-null   object
 12  EXEC_OR_NON-EXEC                   4155 non-null   object
 13  HistIdentity_CEO                   4155 non-null   object
 14  HistIdentity_Chairman              4155 non-null   object
 15  HistIdentity_LeadDirector          4155 non-null   object
 16  PayCommitteeMember                 4155 non-null   object
 17  AuditCommitteeMember               4155 non-null   object
 18  NominatingCommitteeMember          4155 non-null   object
 19  INDEPENDENT_OF_MANAGEMENT          4155 non-null   object
 20  MULTIPLE_CLASSES_OF_VOTING_STOCK   4155 non-null   object
 21  OutsideRelatedReason               4155 non-null   object
 22  OutsideRelatedReason2              4155 non-null   object
 23  CONTROLLED_VIA_STOCK_PYRAMID       4155 non-null   object
 24  HAS_CORPORATE_PARENT               4155 non-null   object
 25  IVY_indicator                      4155 non-null   object
 26  Avg_3rd_connect_17_22              4155 non-null   int64
 27  Ttl_controv_17_22                  4155 non-null   int64
 28  CEO_leav                           4155 non-null   object
 29  Chairman_leav                      4155 non-null   object
 30  Lead_D_leav                        4155 non-null   object
dtypes: category(1), float64(6), int64(5), object(19)
memory usage: 978.1+ KB
```

Note. Please see details in the attached Jupyter Notebook file, <4. Predictive Modeling Part 1 (LR).ipynb>

Furthermore, the target variable has imbalanced class distribution, having 3,794 directors who have not departed, and only 361 directors who have departed (Figure 6-2). Thus, the problem was fixed using SMOTE oversampling technique.

Figure 6-2

Bar Plot of Target Class Distribution



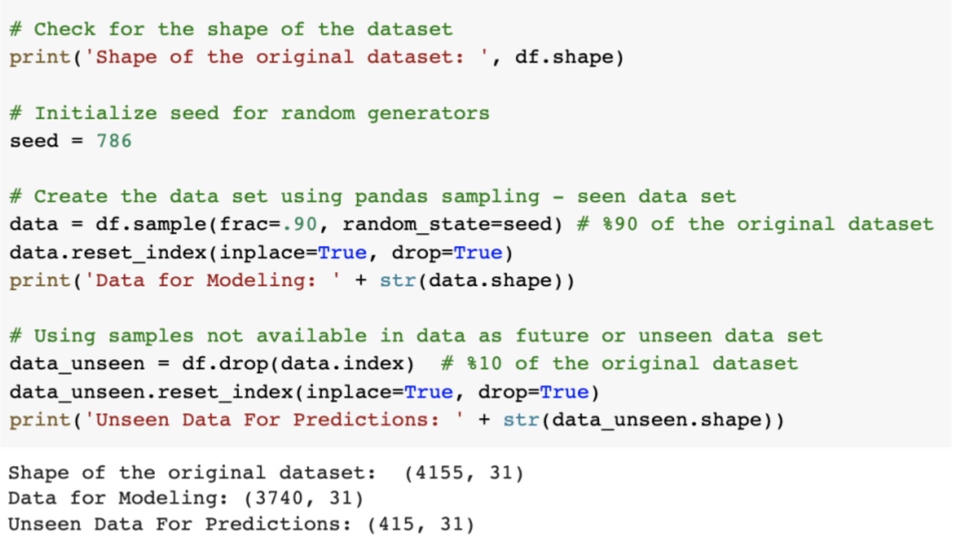
Predictive Models and Performance Evaluation

As the sponsor is looking for the best model to predict who is going to depart the company due to fraud, we need to build multiple models. Instead of building each model manually, we decided to use the Auto ML technique that allows automation of the process by quickly building multiple ML models. In this case, we used the Pycaret library for the classification problem, as it allows us to build multiple models and proceed with the best one, showing the best parameters that work for the data. Thus, this allows not only to speed up the process but also to increase performance accuracy by reducing possible errors or biases made by humans, and instead focusing on the actual problem. Moreover, it has a setup function that allows one to set and change various parameters as needed, such as removing multicollinearity, normalizing, imputing, encoding, fixing class imbalance, and even specifying the method.

For this, we first divided dataset of 4,155 records into 90/10 where 90% is seen data to be used for Auto ML, and 10% is unseen data to be used for prediction (Figure 6-2). This will help us to evaluate prediction of Auto ML technique. As a result, we ended up building models with 90% of the original dataset which is 3,740 records.

Figure 6-2

Screenshot: Splitting the Original Dataset into Modeling and Prediction Datasets



```
# Check for the shape of the dataset
print('Shape of the original dataset: ', df.shape)

# Initialize seed for random generators
seed = 786

# Create the data set using pandas sampling - seen data set
data = df.sample(frac=.90, random_state=seed) # %90 of the original dataset
data.reset_index(inplace=True, drop=True)
print('Data for Modeling: ' + str(data.shape))

# Using samples not available in data as future or unseen data set
data_unseen = df.drop(data.index)  # %10 of the original dataset
data_unseen.reset_index(inplace=True, drop=True)
print('Unseen Data For Predictions: ' + str(data_unseen.shape))
```

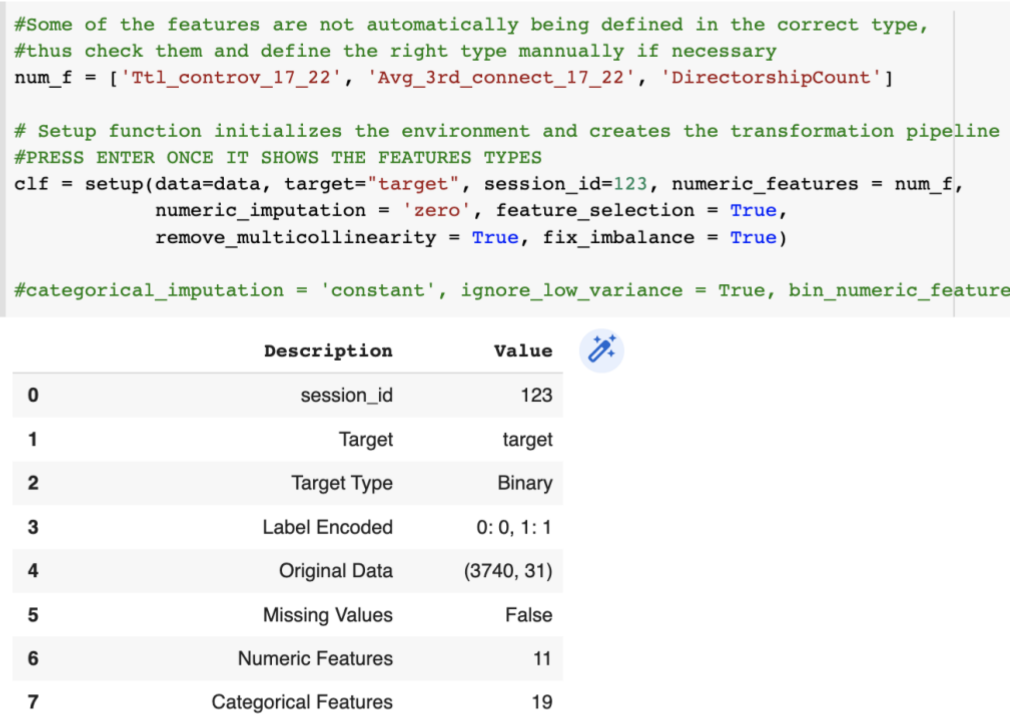
```
Shape of the original dataset:  (4155, 31)
Data for Modeling: (3740, 31)
Unseen Data For Predictions: (415, 31)
```

Note. Please see details in the attached Jupyter Notebook file, <3. Predictive Modeling Part 1 (Auto ML).ipynb>

Next, we set up the environment in PyCaret by specifying the parameters we need and keeping the rest as default (Figure 6-3). The setup function splits the data to test and train sets as 70/30 by default. Here, we set the parameters to remove multicollinearity with a threshold of 0.9, and fix imbalance with the default method of SMOTE (over-sampling), as well as indicated the target variable. Moreover, the feature will be selected based on its important score using sklearn's SelectFromModel. As for the numeric features, missing values will be replaced with 0 as this is the most accurate way to represent data. The code outcome shows all the parameters used for the modeling, target variable, and size of the dataset.

Figure 6-3

Screenshot: Setting up Environment in PyCaret



```
#Some of the features are not automatically being defined in the correct type,
#thus check them and define the right type manually if necessary
num_f = ['Ttl_controv_17_22', 'Avg_3rd_connect_17_22', 'DirectorshipCount']

# Setup function initializes the environment and creates the transformation pipeline
#PRESS ENTER ONCE IT SHOWS THE FEATURES TYPES
clf = setup(data=data, target="target", session_id=123, numeric_features = num_f,
            numeric_imputation = 'zero', feature_selection = True,
            remove_multicollinearity = True, fix_imbalance = True)

#categorical_imputation = 'constant', ignore_low_variance = True, bin_numeric_feature
```

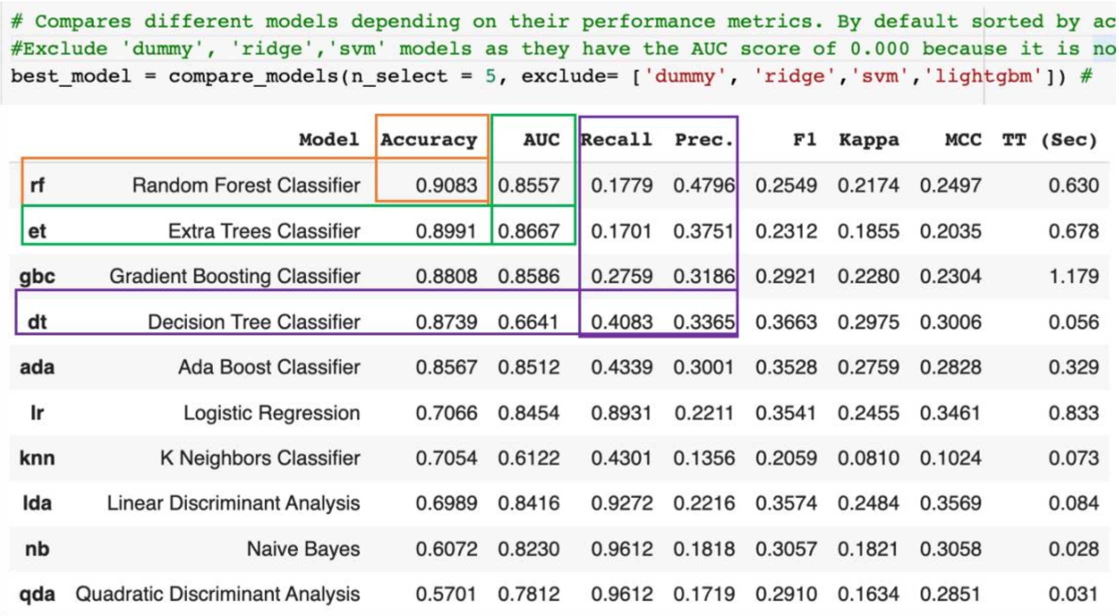
	Description	Value
0	session_id	123
1	Target	target
2	Target Type	Binary
3	Label Encoded	0: 0, 1: 1
4	Original Data	(3740, 31)
5	Missing Values	False
6	Numeric Features	11
7	Categorical Features	19

Note. Please see details in the attached Jupyter Notebook file, <3. Predictive Modeling Part 1 (Auto ML).ipynb>

PyCaret first built 14 different models for classification problems, but we excluded 4 models with 0 AUC, then generated a table sorted by Accuracy score. This helps compare different models based on their performance metrics. According to Figure 6-4, the Random Forest model performed the best with the highest accuracy of 90.83%, while the Extra Trees Classifier performed best in distinguishing departed and not-departed groups having the highest AUC of 86.67%. While from the perspective of best-balanced recall and precision value, the Decision Tree model performed the best. Thus, each model has its own advantages and disadvantages, thus should be chosen based on the project goal, and problem to be solved.

Figure 6-4

Screenshot: Model Performances by Auto ML



```
# Compares different models depending on their performance metrics. By default sorted by ac
#Exclude 'dummy', 'ridge','svm' models as they have the AUC score of 0.000 because it is no
best_model = compare_models(n_select = 5, exclude= ['dummy', 'ridge','svm','lightgbm']) #
```

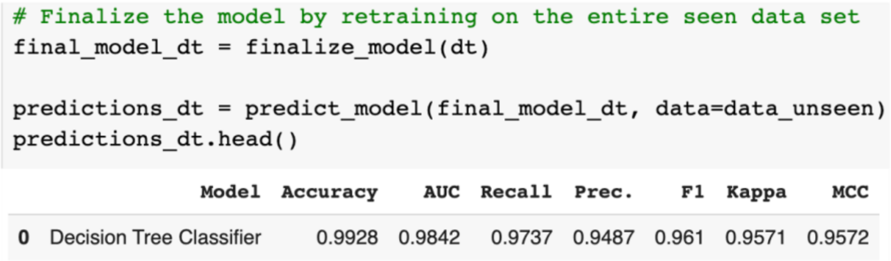
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.9083	0.8557	0.1779	0.4796	0.2549	0.2174	0.2497	0.630
et	Extra Trees Classifier	0.8991	0.8667	0.1701	0.3751	0.2312	0.1855	0.2035	0.678
gbc	Gradient Boosting Classifier	0.8808	0.8586	0.2759	0.3186	0.2921	0.2280	0.2304	1.179
dt	Decision Tree Classifier	0.8739	0.6641	0.4083	0.3365	0.3663	0.2975	0.3006	0.056
ada	Ada Boost Classifier	0.8567	0.8512	0.4339	0.3001	0.3528	0.2759	0.2828	0.329
lr	Logistic Regression	0.7066	0.8454	0.8931	0.2211	0.3541	0.2455	0.3461	0.833
knn	K Neighbors Classifier	0.7054	0.6122	0.4301	0.1356	0.2059	0.0810	0.1024	0.073
lda	Linear Discriminant Analysis	0.6989	0.8416	0.9272	0.2216	0.3574	0.2484	0.3569	0.084
nb	Naive Bayes	0.6072	0.8230	0.9612	0.1818	0.3057	0.1821	0.3058	0.028
qda	Quadratic Discriminant Analysis	0.5701	0.7812	0.9612	0.1719	0.2910	0.1634	0.2851	0.031

Note. Please see details in the attached Jupyter Notebook file, <3. Predictive Modeling Part 1 (Auto ML).ipynb>

This method allows diving deeper into each of the demonstrated models separately. In this case, we extracted Decision Tree Model and evaluated the model prediction by retraining it on unseen data (10% of the original dataset). Figure 6-5 shows the results accuracy of 99.28% and above 95% of precision and recall which means that the prediction of the Auto ML technique was accurate. The code file has more details on models, including confusion matrix, feature importance, etc. which also allows extracting any specific model generated in Figure 6-4.

Figure 6-5

Screenshot: Evaluating Decision Tree Model Prediction on Unseen Data



```
# Finalize the model by retraining on the entire seen data set
final_model_dt = finalize_model(dt)

predictions_dt = predict_model(final_model_dt, data=data_unseen)
predictions_dt.head()
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Decision Tree Classifier	0.9928	0.9842	0.9737	0.9487	0.961	0.9571	0.9572

Regression Analysis

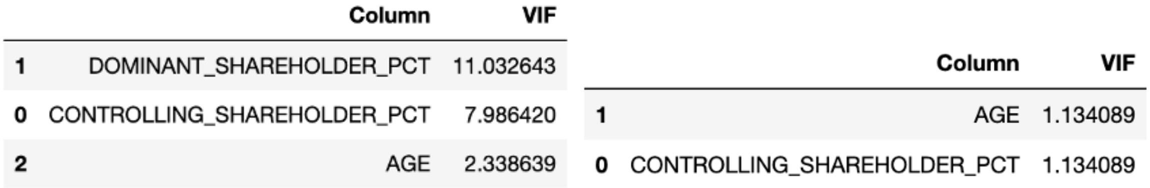
Next, we have built Logistic Regression (LR) with the statsmodels library as it allows us to generate a summary table where we can define each feature’s constant effect on the director's departure. LR is straightforward to be interpreated and unlike the other models with higher accuracies, like random forest and gradient boosting, we apply LR to quantitively analyze how the characteristics (features) influence the target variable within our model. This will also help to respond sponsor’s request to find characteristics of the directors that increases the probability of being departed from the board if controversies happen.

Data Preparation

After separating independent features (X), and target variables (y), we encoded categorical ones in the first to dummy variables using get_dummies() function and dropping the first category. Meaning that if we have a categorical variable with True/False values, it will keep a column with only True values, where 1 for True, and 0 for False. This allows us to remove redundant columns, thus avoid multicollinearity. Furthermore, we checked correlation, and VIF score of the features to remove multicollinearity. The results show that DOMINANT_SHAREHOLDER_PCT variable has the highest VIF score which is greater than the cutoff score of 5 (Figure 6-6). Thus, removing the variable helped to remove VIF score of the rest features.

Figure 6-6

Screenshots: VIF Score of Independent Variables



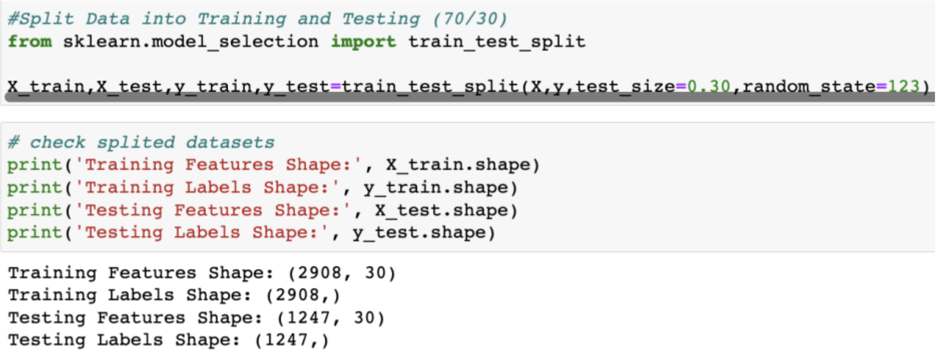
	Column	VIF
1	DOMINANT_SHAREHOLDER_PCT	11.032643
0	CONTROLLING_SHAREHOLDER_PCT	7.986420
2	AGE	2.338639

	Column	VIF
1	AGE	1.134089
0	CONTROLLING_SHAREHOLDER_PCT	1.134089

Next, we split the data into train and test sets with the ratio of 70/30, and end up with 2,908 records for the train, 1,247 for the test sets and 30 features (Figure 6-7).

Figure 6-7

Screenshot: Splitting data into test and train sets



```
#Split Data into Training and Testing (70/30)
from sklearn.model_selection import train_test_split

X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.30,random_state=123)
```

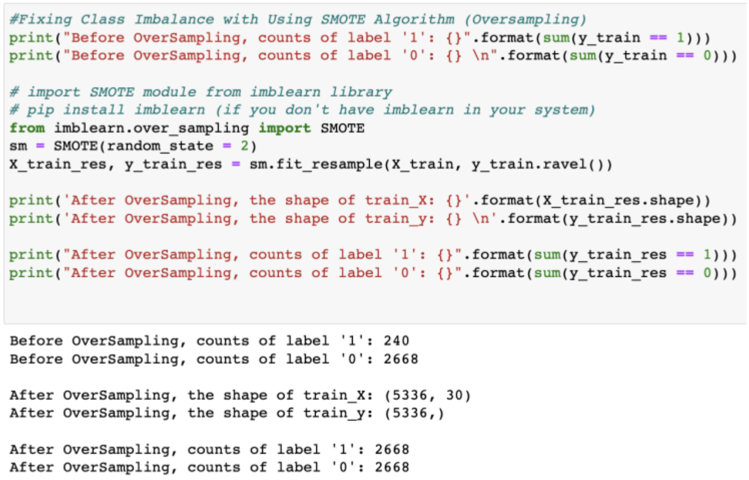
```
# check splited datasets
print('Training Features Shape:', X_train.shape)
print('Training Labels Shape:', y_train.shape)
print('Testing Features Shape:', X_test.shape)
print('Testing Labels Shape:', y_test.shape)
```

```
Training Features Shape: (2908, 30)
Training Labels Shape: (2908,)
Testing Features Shape: (1247, 30)
Testing Labels Shape: (1247,)
```

However, as we ended up with imbalanced data distribution having very few departed classes in the target variable, we used SMOTE (oversampling) algorithm for the train sets to solve this problem. It helps to balance the class distribution increasing minority class by creating new synthetic minority samples. Figure 6-8 demonstrates the final result, where the algorithm oversampled the minority class making them equal to the majority class. Before, we had only 240 records for class 1 (departed directors), and 2,668 records for class 0 (not-departed directors). As a result, we ended up having equal 2,668 records in both classes.

Figure 6-8

Screenshot: SMOTE oversampling for the imbalanced data in train sets



```
#Fixing Class Imbalance with Using SMOTE Algorithm (Oversampling)
print("Before OverSampling, counts of label '1': {}".format(sum(y_train == 1)))
print("Before OverSampling, counts of label '0': {} \n".format(sum(y_train == 0)))

# import SMOTE module from imblearn library
# pip install imblearn (if you don't have imblearn in your system)
from imblearn.over_sampling import SMOTE
sm = SMOTE(random_state = 2)
X_train_res, y_train_res = sm.fit_resample(X_train, y_train.ravel())

print('After OverSampling, the shape of train_X: {}'.format(X_train_res.shape))
print('After OverSampling, the shape of train_y: {} \n'.format(y_train_res.shape))

print("After OverSampling, counts of label '1': {}".format(sum(y_train_res == 1)))
print("After OverSampling, counts of label '0': {}".format(sum(y_train_res == 0)))
```

```
Before OverSampling, counts of label '1': 240
Before OverSampling, counts of label '0': 2668

After OverSampling, the shape of train_X: (5336, 30)
After OverSampling, the shape of train_y: (5336,)

After OverSampling, counts of label '1': 2668
After OverSampling, counts of label '0': 2668
```

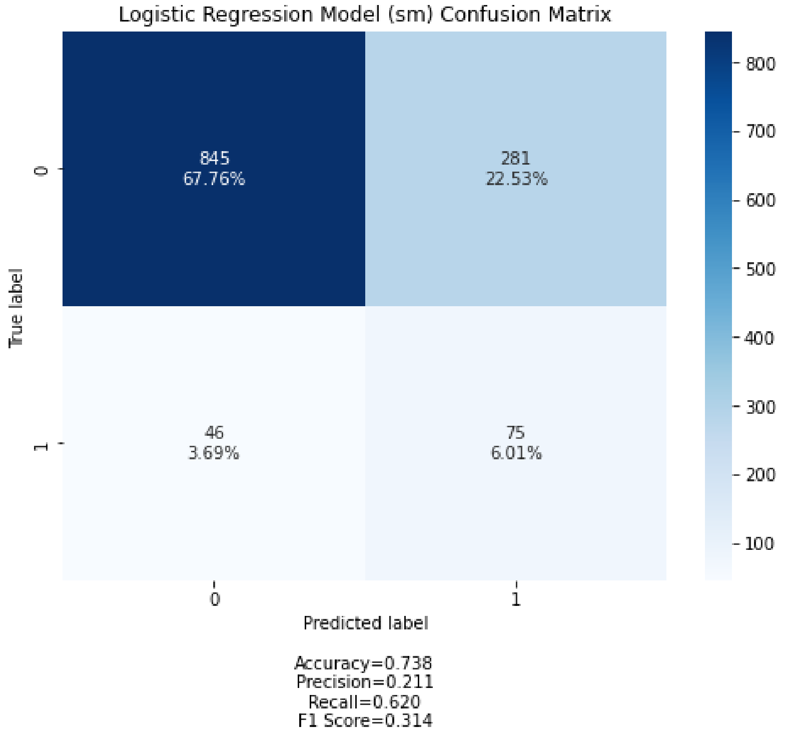
Note. Please see details in the attached Jupyter Notebook file, <4. Predictive Modeling Part 1 (LR).ipynb>

Model Evaluation

Next, we built a Logistic Regression Model using two packages, sklearn, and statsmodels, as the latest one allows us to get the summary table for analyzing the feature importance. Both of them had almost the same performance results. Figure 6-9 demonstrates the Confusion Matrix and Performance metrics of the model which resulted in almost 74% of accuracy. Although this is not the highest accuracy level we would want it to be, it is still a relatively good performance considering not much data is available on the director's level. Furthermore, we can see that the model predicted 73.77% of all data correctly, and 26.22% incorrectly, resulting in more False Positive values than True Positives. For the true prediction, out of 1,247 records in the test set, the departure of 845 directors was predicted correctly as not-departed (TN), and only 75 directors as departed (TP). As for the false prediction, 46 directors were predicted incorrectly as not-departed (FN), while they departed, and 281 directors departed while they did not (FP).

Figure 6-9

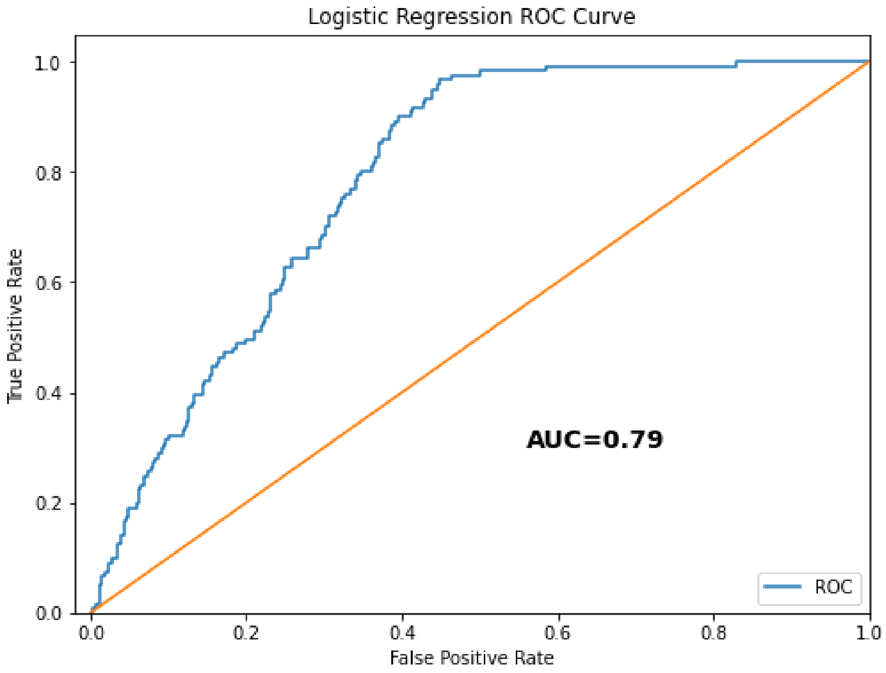
Confusion Matrix and Performance Metrics of Logistic Regression



Then, we plotted the ROC curve to graph demonstrate the summary of classifier performance. Ideally, if the classifiers show the curve closer to the top-left starting from one then the model considers performing a perfect prediction. We can see in Figure 6-10 that our plot looks relatively good but not ideal. Here, the Area Under the ROC curve helps us to identify the quality of model classification. Ideally, the AUC = 1 indicates a perfect classifier. In this case, the AUC is 0.79 which means that the whole area under the ROC curve - a blue line in the plot - is 79% and thus might be considered a good classifier. Thus, our model classifies almost 79% of values correctly.

Figure 6-10

ROC Curve and AUC for Logistic Regression

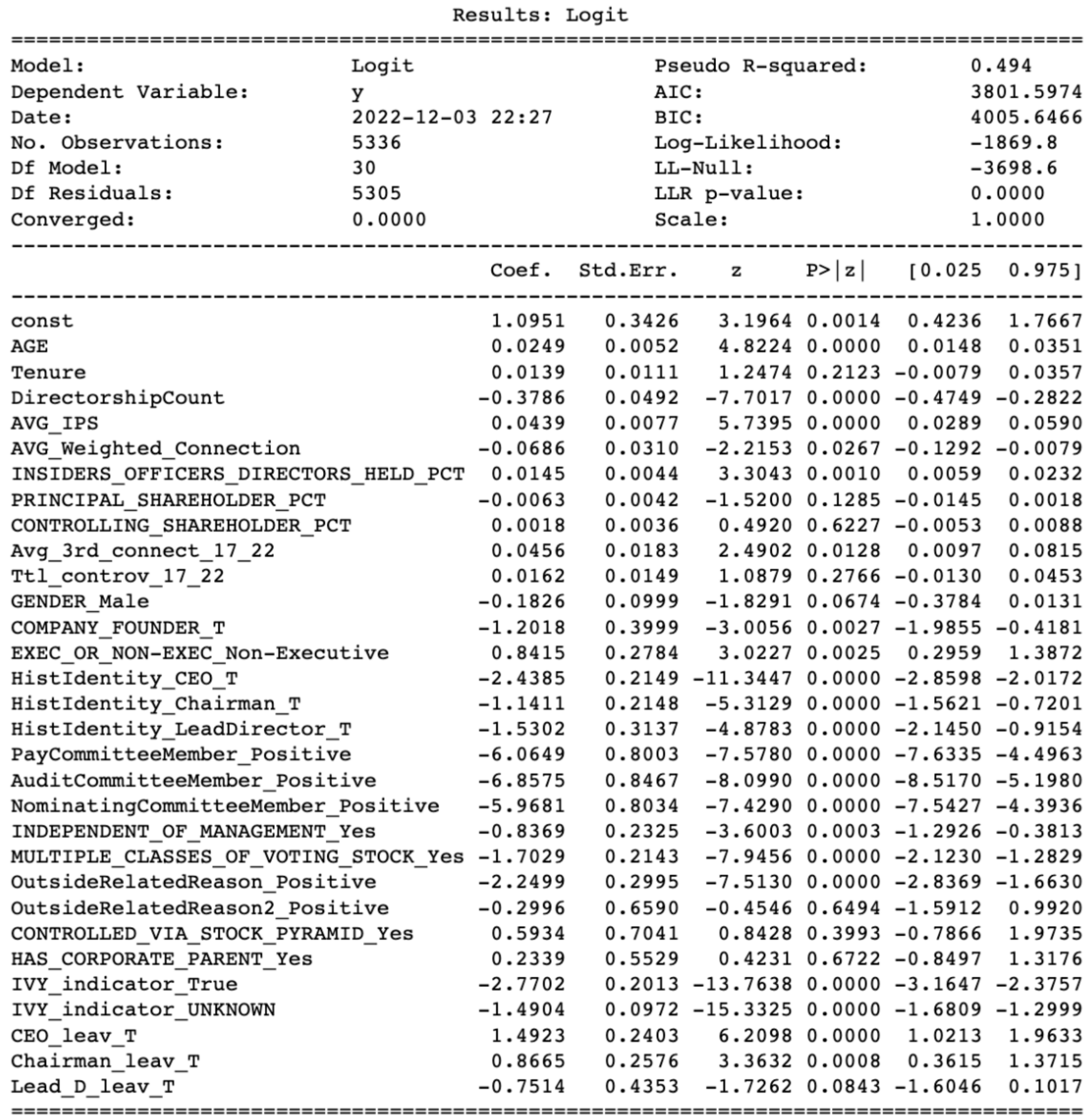


Model Interpretation

Figure 6-11 represents the summary of the Logistic Regression model with 30 features in total. Out of all the features, only 21 of them are important, having p-value less than 0.05, thus have significant impact on the target variable.

Figure 6-11

Logistic Regression Summary Table

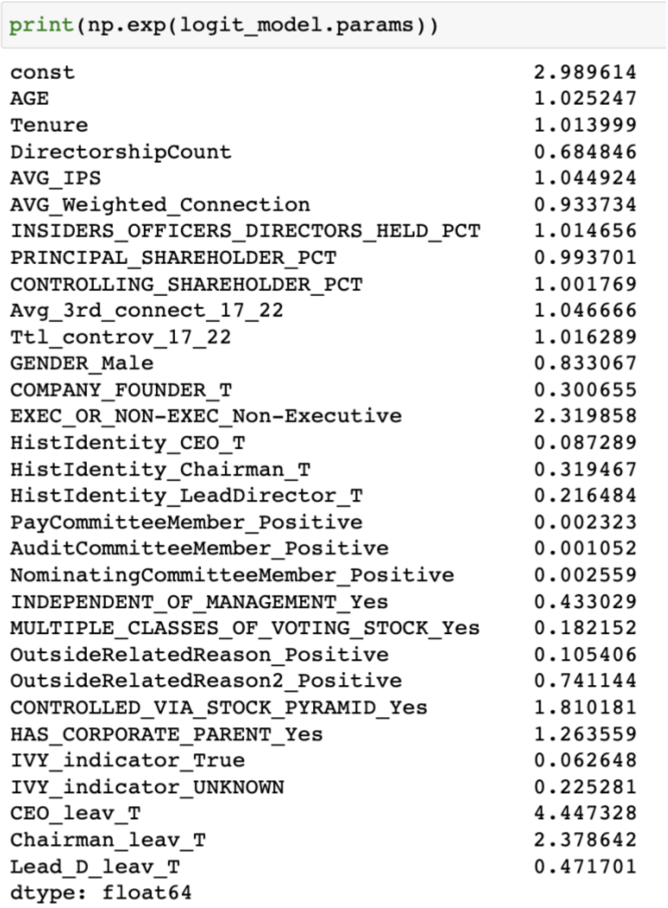


```
                                   Results: Logit
=====================================================================================
Model:                     Logit                   Pseudo R-squared:        0.494
Dependent Variable:        y                       AIC:                     3801.5974
Date:                      2022-12-03 22:27        BIC:                     4005.6466
No. Observations:          5336                    Log-Likelihood:          -1869.8
Df Model:                  30                      LL-Null:                 -3698.6
Df Residuals:              5305                    LLR p-value:             0.0000
Converged:                 0.0000                  Scale:                   1.0000
-------------------------------------------------------------------------------------
                                       Coef. Std.Err.    z     P>|z|   [0.025  0.975]
-------------------------------------------------------------------------------------
const                                 1.0951   0.3426   3.1964 0.0014  0.4236  1.7667
AGE                                   0.0249   0.0052   4.8224 0.0000  0.0148  0.0351
Tenure                                0.0139   0.0111   1.2474 0.2123 -0.0079  0.0357
DirectorshipCount                    -0.3786   0.0492  -7.7017 0.0000 -0.4749 -0.2822
AVG_IPS                               0.0439   0.0077   5.7395 0.0000  0.0289  0.0590
AVG_Weighted_Connection              -0.0686   0.0310  -2.2153 0.0267 -0.1292 -0.0079
INSIDERS_OFFICERS_DIRECTORS_HELD_PCT  0.0145   0.0044   3.3043 0.0010  0.0059  0.0232
PRINCIPAL_SHAREHOLDER_PCT            -0.0063   0.0042  -1.5200 0.1285 -0.0145  0.0018
CONTROLLING_SHAREHOLDER_PCT           0.0018   0.0036   0.4920 0.6227 -0.0053  0.0088
Avg_3rd_connect_17_22                 0.0456   0.0183   2.4902 0.0128  0.0097  0.0815
Ttl_controv_17_22                     0.0162   0.0149   1.0879 0.2766 -0.0130  0.0453
GENDER_Male                          -0.1826   0.0999  -1.8291 0.0674 -0.3784  0.0131
COMPANY_FOUNDER_T                    -1.2018   0.3999  -3.0056 0.0027 -1.9855 -0.4181
EXEC_OR_NON-EXEC_Non-Executive        0.8415   0.2784   3.0227 0.0025  0.2959  1.3872
HistIdentity_CEO_T                   -2.4385   0.2149 -11.3447 0.0000 -2.8598 -2.0172
HistIdentity_Chairman_T              -1.1411   0.2148  -5.3129 0.0000 -1.5621 -0.7201
HistIdentity_LeadDirector_T          -1.5302   0.3137  -4.8783 0.0000 -2.1450 -0.9154
PayCommitteeMember_Positive          -6.0649   0.8003  -7.5780 0.0000 -7.6335 -4.4963
AuditCommitteeMember_Positive        -6.8575   0.8467  -8.0990 0.0000 -8.5170 -5.1980
NominatingCommitteeMember_Positive   -5.9681   0.8034  -7.4290 0.0000 -7.5427 -4.3936
INDEPENDENT_OF_MANAGEMENT_Yes        -0.8369   0.2325  -3.6003 0.0003 -1.2926 -0.3813
MULTIPLE_CLASSES_OF_VOTING_STOCK_Yes -1.7029   0.2143  -7.9456 0.0000 -2.1230 -1.2829
OutsideRelatedReason_Positive        -2.2499   0.2995  -7.5130 0.0000 -2.8369 -1.6630
OutsideRelatedReason2_Positive       -0.2996   0.6590  -0.4546 0.6494 -1.5912  0.9920
CONTROLLED_VIA_STOCK_PYRAMID_Yes      0.5934   0.7041   0.8428 0.3993 -0.7866  1.9735
HAS_CORPORATE_PARENT_Yes              0.2339   0.5529   0.4231 0.6722 -0.8497  1.3176
IVY_indicator_True                   -2.7702   0.2013 -13.7638 0.0000 -3.1647 -2.3757
IVY_indicator_UNKNOWN                -1.4904   0.0972 -15.3325 0.0000 -1.6809 -1.2999
CEO_leav_T                            1.4923   0.2403   6.2098 0.0000  1.0213  1.9633
Chairman_leav_T                       0.8665   0.2576   3.3632 0.0008  0.3615  1.3715
Lead_D_leav_T                        -0.7514   0.4353  -1.7262 0.0843 -1.6046  0.1017
=====================================================================================
```

Further, got the odds ratios of features by taking the exponent of the coefficients, as it is easier to explain and understand (Figure 6-12). This allows us to analyze and see the constant effect of every X predictor on the target variable (departure). In this case, we are going to focus only important 21 features. Important to mention we got high intercept odds ratio of 2.98 which means that, as a basis, the likelihood of the director that does not have any of these mentioned features, increases their departure likelihood to almost 3 times.

Figure 6-12

Screenshot: Getting Odd Ratios of the Features



```
print(np.exp(logit_model.params))
```

```
const                                    2.989614
AGE                                      1.025247
Tenure                                   1.013999
DirectorshipCount                        0.684846
AVG_IPS                                  1.044924
AVG_Weighted_Connection                  0.933734
INSIDERS_OFFICERS_DIRECTORS_HELD_PCT     1.014656
PRINCIPAL_SHAREHOLDER_PCT                0.993701
CONTROLLING_SHAREHOLDER_PCT              1.001769
Avg_3rd_connect_17_22                    1.046666
Ttl_controv_17_22                        1.016289
GENDER_Male                              0.833067
COMPANY_FOUNDER_T                        0.300655
EXEC_OR_NON-EXEC_Non-Executive           2.319858
HistIdentity_CEO_T                       0.087289
HistIdentity_Chairman_T                  0.319467
HistIdentity_LeadDirector_T              0.216484
PayCommitteeMember_Positive              0.002323
AuditCommitteeMember_Positive            0.001052
NominatingCommitteeMember_Positive       0.002559
INDEPENDENT_OF_MANAGEMENT_Yes            0.433029
MULTIPLE_CLASSES_OF_VOTING_STOCK_Yes     0.182152
OutsideRelatedReason_Positive            0.105406
OutsideRelatedReason2_Positive           0.741144
CONTROLLED_VIA_STOCK_PYRAMID_Yes         1.810181
HAS_CORPORATE_PARENT_Yes                 1.263559
IVY_indicator_True                       0.062648
IVY_indicator_UNKNOWN                    0.225281
CEO_leav_T                               4.447328
Chairman_leav_T                          2.378642
Lead_D_leav_T                            0.471701
dtype: float64
```

We ended up having 6 numeric and 15 categorical significant features. Here the dark blue cells with an odds ratio greater than 1 represent a positive relationship, meaning a higher risk to depart; while light blues with an odds ratio less than 1 describe a negative relationship, meaning a lower risk to depart. Figure 6-13 shows numeric features sorted by the odds ratio. Here, for each 1 unit increase in the predictor, the likelihood of being departed increased for the value of the odds ratio. For a positive relationship, when a controversial event happens, the odds of departing increase by a factor of 1-1.046 = 0.046 for each 1 more 3rd-degree connection, holding all other predictors constant. In other words, the likelihood of departing increases by 4.6% for each 1 more 3rd-degree connection. Same for average influence share %, the likelihood of departing increases by 4.4% for each 1 % increase in influence share. And the likelihood of departing increases by 2% for each 1-year increase in the age of director. Although, it seems that some features do not really have a strong relationship with the target, it still matters. For example, in case there is a 10-year difference in the age of the directors, the likelihood to be departed increases by 20% for each 10 more years increase in age. As for the negative relationship, the likelihood of departing decrease by 7% for each 1 more average weighted connection and decrease by 32% (100%-68%) for each 1 more directorship count.

Figure 6-13

Important Numeric Features with Odds Ratio

Numeric Features	
Name	Odds ratio
Avg 3rd degree connection	1.046
Avg Influence share %	1.044
Age	1.02
Insiders Officers Directors Held %	1.01
Avg Weighted Connection (edge)	0.93
Directorship Count	0.68

Figure 6-14 shows categorical features sorted by the odds ratio. We will explain starting with the positive relationship, directors who are a CEO or Chairmen of the boards during controversial events have higher departure likelihoods of 2 to 4 times compared to the directors who are not in the positions. Moreover, if a director is not a member of management (non-executive), their departure likelihood increases over 2.31 times. On the contrary, those directors with the characteristics shown in blue cells are at lower risk to depart. For example, for directors who are independent of management or used to be CEO/Chairman/Company founder, their departure likelihoods are reduced to less than 30% to 40%. Important to mention the last three features, the odds of depart are very low if a director is a member of the nominating committee, and/or the pay committee, and/or the audit committee. In other words, there is a 1 in 10 chance of departing a director who is a member of the audit committee.

Figure 6-14

Important Categorical Features with Odds Ratio

Categorical Features	
Feature	Odds ratio
CEO while leaving – True *	4.44
Chairman while leaving – True *	2.37

Non-Executive	2.31
Independent of Management - Yes	0.43
Was a Chairman while on the board - True	0.31
Company Founder - True	0.30
Was a Lead Director while on the board - True	0.21
Multiple Classes Of Voting Stock - Yes	0.18
Outside Related Reason - Positive	0.10
Was a CEO while on the board - True	0.08
Ivy League Alumni - True	0.06
Nominating Committee Member – Positive	0.0025
Pay Committee Member - Positive	0.0023
Audit Committee Member - Positive	0.001

Summing up, having more of 3rd-degree connections, and/or influence share %, and/or Age, and/or Insider Officers Directors Held %, increases directors' likelihood to be departed. At the same time, the having more of weighted connection, and/or directorship count decreases directors' likelihood to be departed. Moreover, a director who is also a CEO and/or a chairman of the company at the moment when a controversial event happens, and/or not a member of the management, is at higher risk to depart. At the same time, a director who is independent of management, and/or was a chairman/lead director/CEO at any time while on the board, and/or is a company founder, and/or has Multiple Classes Of Voting Stock, and/or has Outside Related Reason, and/or Ivy League Alumni is at lower risk to depart. As for the director who is a member of the nominating/pay /audit committee, they have a very low risk to depart.

Predictive Models Part 2

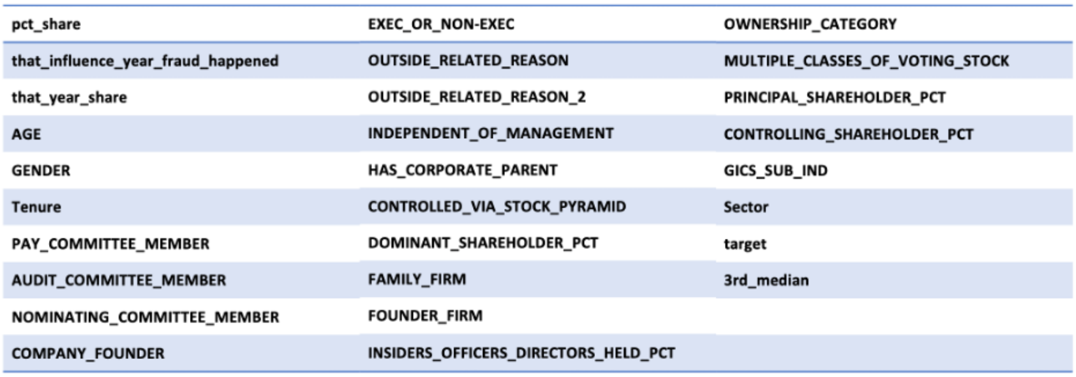
This part will apply the second approach of IPS data and a different data structure from the previous descriptive analysis and predictive models to check if the model performance can get improved by a larger sample size. In the former studies, one of the big problems we encountered was that the proportion of the target variable in the total data was relatively low, and such deviation made the model inaccurate. Based on this problem, we applied SMOTE technique to balance the dataset. 'Synthetic Minority Oversampling Technique' generated similar target variables and enlarge the datasets.

Data Preparation

First, we selected 28 features for the modeling. The variables that were chosen here are the features the characteristics of individual directors we want to test for hypotheses and some company-level information, such as percentage share, age, gender, role on the board, and so on. Secondly, we encode the categorical features and normalized the data. The data that was prepared for the modeling has 30380 rows and 28 columns, among them, 16495 rows were original data, and 13885 were enlarged data.

Table 7-1

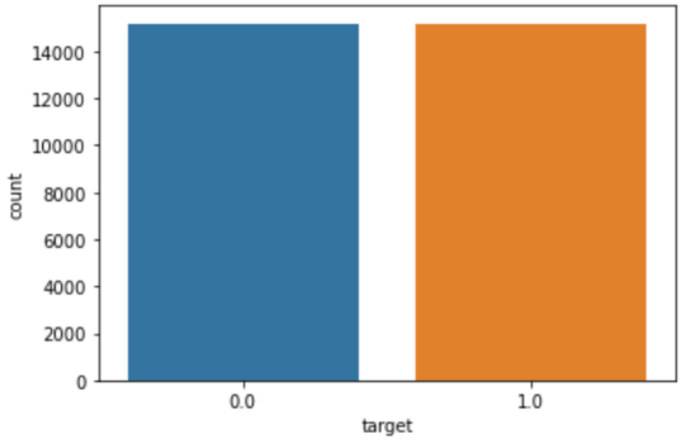
Feature Selection for Modeling:



pct_share	EXEC_OR_NON-EXEC	OWNERSHIP_CATEGORY
that_influence_year_fraud_happened	OUTSIDE_RELATED_REASON	MULTIPLE_CLASSES_OF_VOTING_STOCK
that_year_share	OUTSIDE_RELATED_REASON_2	PRINCIPAL_SHAREHOLDER_PCT
AGE	INDEPENDENT_OF_MANAGEMENT	CONTROLLING_SHAREHOLDER_PCT
GENDER	HAS_CORPORATE_PARENT	GICS_SUB_IND
Tenure	CONTROLLED_VIA_STOCK_PYRAMID	Sector
PAY_COMMITTEE_MEMBER	DOMINANT_SHAREHOLDER_PCT	target
AUDIT_COMMITTEE_MEMBER	FAMILY_FIRM	3rd_median
NOMINATING_COMMITTEE_MEMBER	FOUNDER_FIRM	
COMPANY_FOUNDER	INSIDERS_OFFICERS_DIRECTORS_HELD_PCT	

Figure 7-2

Balanced Dataset after SMOTE:



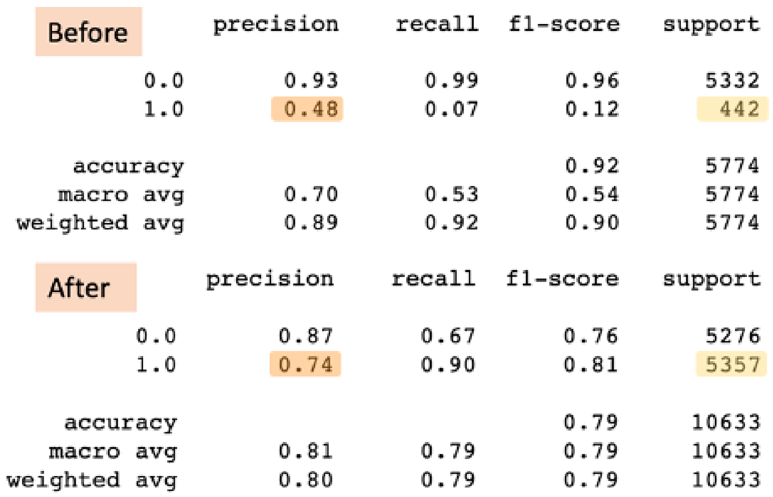
The last step before modeling is split train and set data. Here we applied train test split through sklearn.model_selection package. The test set takes 35% of the dataset, train set takes 65% of the dataset. After splitting the data, the train data has 19747 records, and the test set has 10633 records.

Model Performance

First and foremost, we redid the logistic regression model. Before SMOTE, the accuracy of logistic regression accuracy is 92%, and the accuracy of predicting departed people is 48%. There were only 442 support target variable data points, which is 10% of the dataset (Figure 7-3). The data that supports the target variable after SMOTE has 5357 records, which is 50% of the dataset. In this case, the overall accuracy of the model dropped from 93% to 79% after SMOTE. On the other hand, the precision of predicting target variables increased from 48% to 74%, which is a huge leap. Based on this change, we think SMOTE is an appropriate approach for this scenario.

Figure 7-3

Screenshots: Performance of Logistic Regression Model before/after the Oversampling

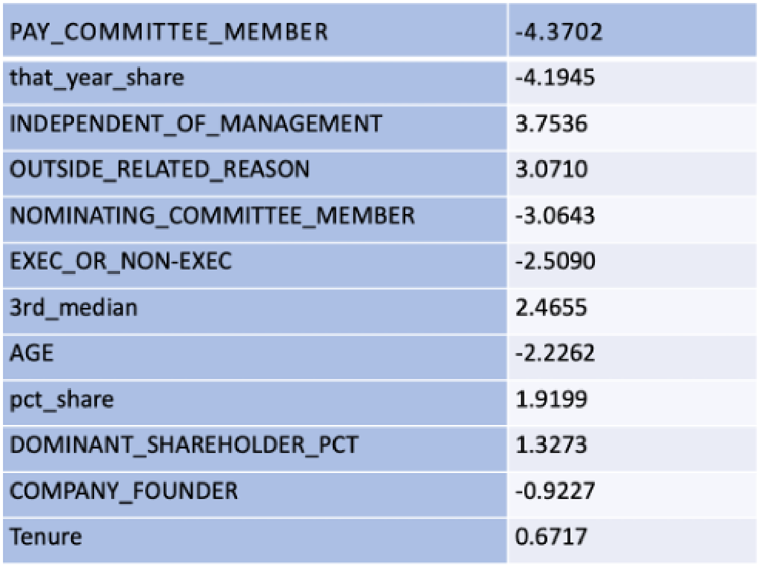


```
  Before        precision    recall  f1-score   support

         0.0       0.93      0.99      0.96      5332
         1.0       0.48      0.07      0.12       442

    accuracy                           0.92      5774
   macro avg       0.70      0.53      0.54      5774
weighted avg       0.89      0.92      0.90      5774
  After        precision    recall  f1-score   support

         0.0       0.87      0.67      0.76      5276
         1.0       0.74      0.90      0.81      5357

    accuracy                           0.79     10633
   macro avg       0.81      0.79      0.79     10633
weighted avg       0.80      0.79      0.79     10633
```

Next, let’s have a closer look at the model's performance (Table 7-2). The top three features are pay_committee_member, that_year_share, independent_of_management. Being a member of the pay committee has significant negative impact on the departure results while having option of independent of management has significant positive impact on the departure results.

Table 7-2

Coefficients of Logistic Regression Model

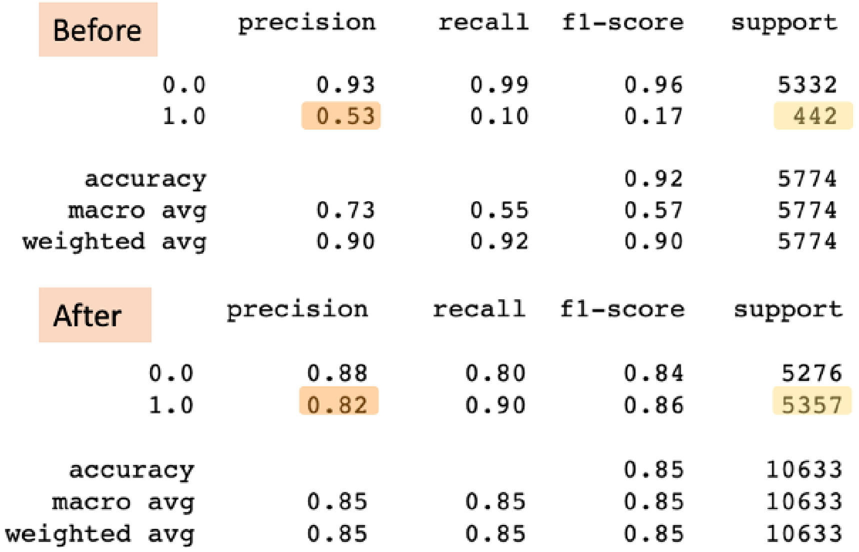


PAY_COMMITTEE_MEMBER	-4.3702
that_year_share	-4.1945
INDEPENDENT_OF_MANAGEMENT	3.7536
OUTSIDE_RELATED_REASON	3.0710
NOMINATING_COMMITTEE_MEMBER	-3.0643
EXEC_OR_NON-EXEC	-2.5090
3rd_median	2.4655
AGE	-2.2262
pct_share	1.9199
DOMINANT_SHAREHOLDER_PCT	1.3273
COMPANY_FOUNDER	-0.9227
Tenure	0.6717

Coming up with a decision tree classifier model. The accuracy of the model decreased from 92% to 85% after SMOTE. On the other hand, the precision of predicting departed members increased from 53% to 82%.

Figure 7-4

Screenshots: Performances of Decision Tree Model before/after Oversampling

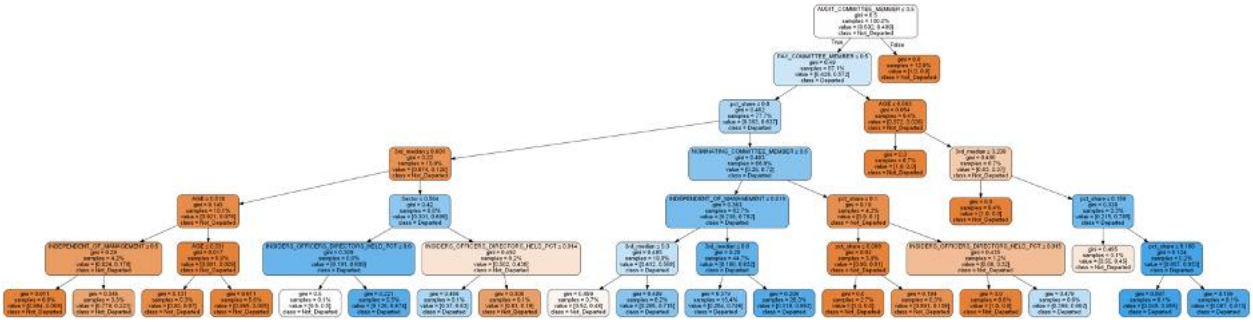


```
Before      precision    recall  f1-score   support

         0.0       0.93      0.99      0.96      5332
         1.0       0.53      0.10      0.17       442

    accuracy                           0.92      5774
   macro avg       0.73      0.55      0.57      5774
weighted avg       0.90      0.92      0.90      5774

After       precision    recall  f1-score   support

         0.0       0.88      0.80      0.84      5276
         1.0       0.82      0.90      0.86      5357

    accuracy                           0.85     10633
   macro avg       0.85      0.85      0.85     10633
weighted avg       0.85      0.85      0.85     10633
```

Figure 7-5 is the visualization of the decision tree classifier. Through the visualization, we can easily identify the most influential feature on the target variable from top to bottom. For example, being an audit/pay committee member, influence percentage score, and third-degree connections have the most impact in our classification model.

Figure 7-5

Visualization of Decision Tree Classifier


AUDIT_COMMITTEE_MEMBER ≤ 0.5
True
False
PAY_COMMITTEE_MEMBER ≤ 0.5
NOMINATING_COMMITTEE_MEMBER ≤ 0.5
INDEPENDENT_OF_MANAGEMENT ≤ 0.519
INSIDERS_OFFICERS_DIRECTORS_HELD_PCT ≤ 0.045
class = Departed
class = Not_Departed

We also applied Random Forest, SVC, KNN, and Naïve Bayes models. Among all, the random forest has the best performance which has 99% of accuracy, coming up with SVC with 86% accuracy, KNN with 82% accuracy, and Naïve Bayes with 76% accuracy.

Figure 7-6

Confusion Matrixes for Random Forest, SVC, KNN, and Naïve Bayes Models:

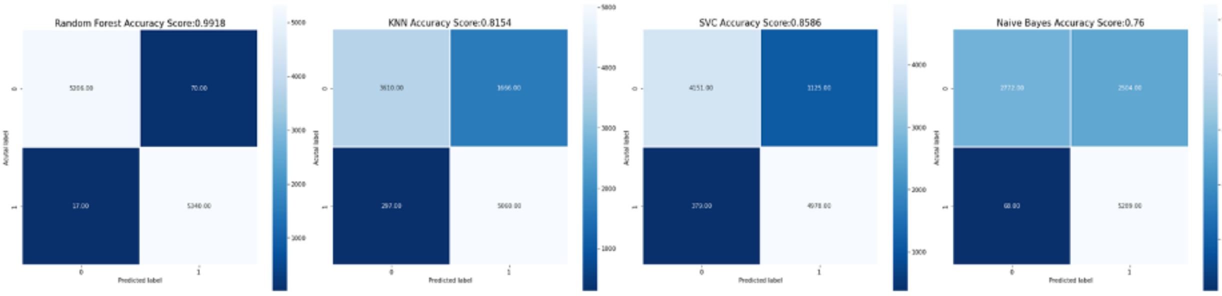
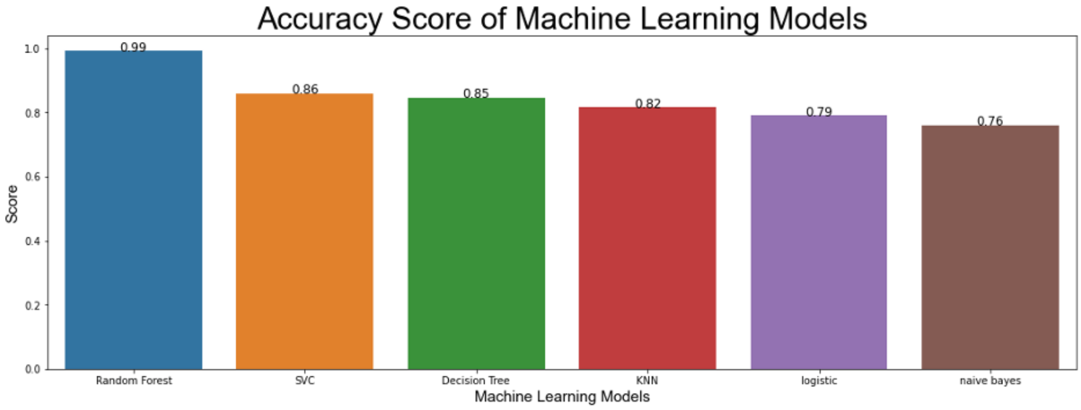


Figure 7-7

Bar Chart: Accuracies of Predictive Models



After all, the random forest has the best performance among all the models. We recommend using random forest, svc, and decision tree for this data. The second approach shows a higher accuracy overall. Especially in logistic regression, the influence percentage share of the year fraud happened has a high coefficient. The second approach gives us a different understanding of the data. In the end, we see a lot of similarities compared to the first approach. It's another way of thinking and processing that valid our modeling and research.

Conclusions

By now, we have raised our assumptions and hypotheses and completed the data processing and predictive modeling. In the last section, we will discuss our findings regarding the hypotheses testing and our predictive models.

Hypotheses Test Results

Quantitive Influence. H1.1 and H1.2 focus on the quantifiable influences of board directors. We did not observe obvious differences between the departed group and the non-departed group regarding these quantitive influences. However, the influence percentage score (IPS) was shown to have a significantly positive influence on the departure results, as well as the insider's officers directors held percentage – with 10 percentages increased in these two percentage values, the departure likelihoods of directors will increase about 50% and 10% respectively (at the alpha level of 0.05, same as follows). Thus, H1.1 and H1.2 are proven as false. This can be explained by the directors with relatively higher quantitive influence values are regarded as being responsible for the operations and decision-making of the companies: when their companies involve controversies, they should be the first to be blamed.

Identities. H2.1, H2.2, and H2.3 focus on the identities or positions that directors hold. From the descriptive analysis, we have observed that directors in the pay, audit, or nominating committees has obviously lower departure rates than directors who are not in these committees; also from the regression analysis, we can conclude that H2.1 is true: being the committees has significantly negative impacts on the departure results when companies involve controversies. However, regression analysis shows that being CEO or chairman has significantly positive impacts on the departure results while being a lead director has significantly negative impacts on the departure results. Thus, H2.2 is false. Lastly, being a company founder has significantly negative impacts while being the executive director has significantly positive impacts on the departure results. Thus, H2.3 is false. The results can be explained by the fact that the directors who are CEOs, chairmen of boards, and executive directors are the people who are directly responsible for the company operations. As a result, they are the first responsible for being blamed when controversial events are disclosed to the public.

Network & Resources. H3.1 to H3.4 focus on the resources of directors. The tenure length is shown as the insignificant feature in the regression analysis and predictive models; also, the average number of third-degree connections was proven to have significantly positive impacts on the departure results: thus, H3.2 and H3.4 are proven false. On the other hand, H3.1 (weighted connectivity) and H3.3 (number of directorships) are both proven as true through the descriptive analysis and regression analysis, which can be explained by companies that tend to keep the directors with more resources from which companies may benefit.

Directors' Options. H4.1 and H4.2 focus on the options of “independent of management” and “multiple classes of voting stocks”. Both hypotheses are proven as true: directors who are independent of management or have multiple classes of voting stocks are less likely to be departed compared to directors who do not have these options.

Others. H5, H6, and H7 focus on other characteristics of individual directors. Except for the H5 (gender), the other two hypotheses are both proven as true in the regression analysis: the departure likelihood of director A who is ten years older than director B is 20% more compared to director B; the departure likelihood of directors who graduated from IVY league schools is only 6% of the likelihood of directors who graduated from other schools.

Predictive Modeling

Even though we applied two different data structure approaches to build predictive models, both approaches have great accuracies in predicting the binary departure results.

Considering the various aspects of model performance (precision, recall, AUC, and prediction time), we recommend the Random Forest classification model. On the one hand, Random Forest classification has shown the best performance in prediction accuracy; on the other hand, it efficiently overcomes the over-fitting problem by assembling a number of decision trees (which already has great accuracy and balance performance between the precision and recall).

On the other hand, to further improve predictive model performance on departure results, there are several aspects to be considered:

Overcome the imbalance distributions of independent variables. Due to the limitations of boards of directors, only a few directors may have important positions, such as committee members, lead directors, etc. Even though we have applied the oversampling method to balance the target variable (departure results), the imbalance distributions in independent variables could also influence the model performance and the interpretations of the regression analysis.

Variable control. In our data processing, we have only applied the directors in the US-issued companies to avoid information loss and possible different board operation strategies across different countries. However, due to the limits of data scope, we have not separated the observations by different capital levels of the companies: it is very likely that compared to the companies worthy of millions, the companies worthy of billions have different strategies or behaviors to expel directors during controversial events, as well as for the different industries, like financial companies and manufacture companies. In other words, this issue should be regarded as the “outliers” problem in our research. With more observations or data, variable controlling of capital levels and sectors should be taken into account before building the model.

Assumption limitations. We set the assumptions to identify our target: “departed” or “non-departed”. From the descriptive analysis and predictive models, it works fine – but still, it can do better with additional data sources. The biggest limitation of the assumptions is we evaluate the departure results each year; however, the real scenario of controversies could have different time lengths – some of them could last for years. If we have the proximate time of how long each controversial event last, we can identify the departure results more precisely, obtaining a more balanced target variable distribution and reducing the misclassified numbers in predictive models.

References

D’Onza, G., & Rigolini, A. (2016). Does director capital influence board turnover after an incident of fraud? Evidence from Italian listed companies. Journal of Management &amp; Governance, 4, 993–1022. https://doi.org/10.1007/s10997-016-9372-2

Gao, Y., Kim, J.-B., Tsang, D., & Wu, H. (2016). Go before the whistle blows: an empirical analysis of director turnover and financial fraud. Review of Accounting Studies, 1, 320–360. https://doi.org/10.1007/s11142-016-9381-z

Marcel, J. J., & Cowen, A. P. (2013). Cleaning house or jumping ship? Understanding board upheaval following financial fraud. Strategic Management Journal, 6, 926–937. https://doi.org/10.1002/smj.2126

Martin, K. G. (2022, October 14). Why use odds ratios in logistic regression? The Analysis Factor. Retrieved December 5, 2022, from https://www.theanalysisfactor.com/why-use-odds-ratios/

Srinivasan, S. (2005). Consequences of Financial Reporting Failure for Outside Directors: Evidence from Accounting Restatements and Audit Committee Members. Journal of Accounting Research, 2, 291–334. https://doi.org/10.1111/j.1475-679x.2005.00172.x