Jihong Yu
Professor Paul McNelis
Machine Learning in Finance
12/15/2021

Which company will survive from S&P 500

1) Introduction

S&P 500 is the most famous and crucial stock index tracking the performance of 500 companies S&P picked up. If S&P picks up stock, the pick gives a positive sign to the stock market, and S&P should buy the stocks so that the stock price will increase. If a stock is delisted in the index, the exclusion presents a negative sign to the market, and S&P should sell the stocks. As a result of delisting, the stock price will decrease. Because enlisting and delisting from the S&P 500 can influence the stock price, predicting which company will be delisted from the index makes investors avoid loss. A stock price generally decreases about 15% in the long term after the stock is delisted from the index.

2) Research purposes

a) What characteristics affect delisting from S&P 500
With Random Forest and Neural Network models, I will investigate what characteristics of a company influence the delisting from the index.

b) Which stocks will be delisted from the index
Among current stocks in the index, I will figure out which stocks are most exposed to the risk of delisting.

3) Methods of analysis

A) Description of characteristics

I constructed five aspects of characteristics: Balance sheet aspect, Profit and Loss aspect, Cash flow aspect, Business aspect, and Trading ratio aspect.

i) Balance sheet aspect: Total Asset, Total Enterprise Value, Total Debt, and Total Equity
ii) Profit and Loss aspect: Total sales, Operating income, Net income, EBITDA, Dividend rate, and historical growth rate
iii) Business aspect: How old the company is, Whether the company is in the Tech industry (Tech = 1), Whether the company originated from the U.S. (the U.S. = 1), The number of total employees, and Market Capital.
iv) Cash flow: Free Cash Flow to Firm (FCFF), and Capital Expenditure (CAPEX)
v) Trading Ratio: PER, PSR, PBR, and EV/EBITDA

- Because characteristics' level of values is various among companies and how relatively high or low is a crucial factor for survival, all variables are scaled by MinMaxScaler function to alleviate the risk that the level of variable could distort the results.
- All models use training sets as 70% of total data, and 30% of total data is used as test sets.

B) Analyzing difference between time series

i) Stock list in 2001: Average companies resist 20.7 years in the index, so to investigate what characteristics influenced survival in the long-term aspect, I chose S&P 500 list in Dec 31, 2001. In the list, companies with all characteristics are 386, and 282 companies were delisted by 2021. (73.06%)
ii) Stock list in 2009: To figure out what characteristics effect in mid-term, I chose S&P 500 list in Dec 31, 2009, after 2008's crisis. In the list, companies with all characteristics are 405, and 139 companies were delisted by 2021. (34.32%)
iii) Stock list in 2019: To investigate what characteristics influenced survival in the short-term aspect, I chose S&P 500 list in Dec 31, 2019. In the list, companies with all characteristics are 441, and only 22 companies were delisted by 2021. (4.99%) Dec 31, 2019 was just before the COVID-19,

through this list, I can infer what features can help companies to survive from S&P 500 in the post COVID-19 era.

C) I analyzed the sensitivity test to find the best hyper-parameters for the RF model with 2001 data.
   i) I conducted the sensitivity test with the RF model, changing the number of trees. [Figure 1] As a result of the test, the accuracy is the highest, 0.7741, when the number of trees is 90. To conduct further sensitivity tests with other hyper-parameters, I choose five candidates of trees (90, 97, 94, 93 and 91) based on the cross-validation accuracy scores. [Figure 2]
   ii) Based on the highest accuracy with the 'GridSearchCV' algorithm, 0.7778 [Figure 4], among candidates for hyper-parameters [Figure 3], the best RF model has the 'Gini' criterion, 25 maximum depth, 35 maximum leaf nodes, and 94 trees.
   iv) Confusion matrix based on the best RF model with total data shows that the accuracy performance is 88.86%, False positive is 1% and False negative is 10%. [Figure 5]
   v) The best five features based on importance scores are 'Dividend rate', Net Income', 'EBITDA', 'FCFF', and 'CAPEX' respectively. [Figure 6]
   vi) The hyper-parameter adjusted RF model with 2001 data predicts that 61 stocks will be excluded after 20 years. [Figure 7]

D) I analyzed the sensitivity test to find the best hyper-parameters for the NN model with 2001 data.
   i) I conducted the sensitivity test with the NN model, changing the number of hidden layers and maximum iterations. [Figure 8] When the 1,000 iterations, 219 hidden layers show 74.07% accuracy. With the 2,000 iterations, 222 hidden layers shows 74.07% accuracy. With the 5,000 iterations, 222 hidden layers shows 74.44% accuracy. With the 10,000 iterations, 246 hidden layers shows 73.70% accuracy.
   ii) Therefore, I confirm that the NN model should have222 hidden layers with 5,000 maximum iterations to present the best performance. [Figure 9]
   iii) Confusion matrix based on the hyper-parameters adjusted NN model with total data shows that the accuracy performance is 80.05%, False positive is 4.4% and False negative is 16%. [Figure10]
   iv) The importance chart based on the best NN model does not present meaningful information about which variables influence survival because of the enormous numbers of iterations and hidden layers. [Figure 11]
   v) The hyper-parameter adjusted NN model with 2001 data predicts that 18 stocks will be excluded after 20 years. [Figure 12]

E) I analyzed the sensitivity test to find the best hyper-parameters for the RF model with 2009 data, just as I did with the Random Forest model including 2001 data.
   i) [Figure 14] As a result of the test, the accuracy is the highest, 0.7458, when the number of trees is 87. To conduct further sensitivity tests with other hyper-parameters [Figure 15, 16]
   ii) Based on the highest accuracy with the 'GridSearchCV' algorithm, 0.7493 [Figure 17], the best RF model has the 'Gini' criterion, 10 maximum depth, 25 maximum leaf nodes, and 89 trees.
   iv) Confusion matrix based on the best RF model with total data shows that the accuracy performance is 87.16%, False positive is 11% and False negative is 2%. [Figure 18]
   v) The best five features based on importance scores are 'Market Capital', Total Enterprise Value', 'EV/EBITDA', 'EBITDA', and 'Operating Income' respectively. [Figure 19]
   vi) The hyper-parameter adjusted RF model with 2009 data predicts that 14 stocks will be excluded after 12 years. [Figure 20]

F) I analyzed the sensitivity test to find the best hyper-parameters for the NN model with 2009 data just as I did with the Neural Network including 2001 data.

i) Changing the number of hidden layers and maximum iterations, I confirmed that the NN model should have 236 hidden layers with 10,000 maximum iterations to present the best performance. [Figure 21]
ii) Confusion matrix based on the hyper-parameters adjusted NN model with total data shows that the accuracy performance is 77.53%, False positive is 15% and False negative is 7.2%. [Figure22]
iii) Starting from 2009 list, Neural Network models predict every stock in 2021's list will survive within 12 years.

G) I analyzed the sensitivity test to find the best hyper-parameters for the RF model with 2019 data, just as I did with the Random Forest model including 2001 data.
i) [Figure 23, 24, 25, 26] As a result of the sensitivity test, the best RF model has the 'Entropy' criterion, 20 maximum depth, 30 maximum leaf nodes, and 13 trees.
iv) Confusion matrix based on the best RF model with total data shows that the accuracy performance is 98.41%, False positive is 1.4% and False negative is 0.23%. [Figure 27]
v) The best five features based on importance scores are 'Market Capital', Total Enterprise Value', 'EV/EBITDA', 'EBITDA', and 'Operating Income' respectively. [Figure 19]
vi) The hyper-parameter adjusted RF model with 2019 data predicts every stock in 2021's list can survive within 2 years.

H) I analyzed the sensitivity test to find the best hyper-parameters for the NN model with 2019 data just as I did with the Neural Network including 2001 data.
iv) Changing the number of hidden layers and maximum iterations, I confirmed that the NN model should have 202 hidden layers with 10,000 maximum iterations to present the best performance. [Figure 29]
v) Confusion matrix based on the hyper-parameters adjusted NN model with total data shows that the accuracy performance is 97.28%, False positive is 2% and False negative is 0.68%. [Figure30]

Conclusion and inference

A) Prediction conclusion
i) The RF and NN models with 2001 data predict that 8 companies in the 2021 list can not survive after 20 years. [Figure 13]
ii) Two RF models with 2001, and 2009 data and NN model with 2001 data anticipated that 6 stocks will be excluded from the S&P 500 within 20 years. [Figure 30]
iii) If an investor seeks long-term profit from the stock market, the models recommend that do not put your money on the stocks in the [Figure 13] and [Figure 30]. The investor can avoid expected loss by the models, not investing in the stocks.
iv) The models with 2019 data predict that every stock in the 2021 list will survive within 2 years. Because some companies will be delisted every year, this result is clearly wrong, but I can infer that S&P 500 cannot find the critical problems with current companies comprising S&P 500 based on the current public data.

B) Importance [Figure 31]
i) In the long-term the Dividend Rate and Net income are the most crucial features for survival. Since investors expect dividends from long-term investment and a large dividend rate and net income illustrate the company's stable operation, it is reasonable that these features are the most significant.
ii) Relatively short-term aspects, current Market capital is the most important for survival. It can be interpreted as the S&P 500 considering Market capital the best business result, believing efficient market theory.
iii) Among the balance sheet aspects, Total enterprise value outperformed the other factors, total equity, total asset, and total debt. Total EV presents the consolidated value of the business,

considering asset, debt, and equity value. Hence, valuation experts regard total EV as the best criteria for evaluating the target company's value. Therefore, the result, outperforming EV, well illustrates valuation reality.

iv) Along EV, EV/EBITDA is usually an important factor by valuation experts. Furthermore, not as much as FCFF but EBITDA is manipulated as the approximated cash flow value by adding noncash flow expenses such as depreciation and amortization. When evaluating the target company's value, CF is more important than profit and loss. Therefore, the results, EBITDA and EV/EBITDA, are highly ranked, also represent valuation reality.

v) Capital Expenses, CAPEX, is an investment for future growth. Despite the common sense that professional investors do not care about future growth, CAPEX played a significant role in the classification, so we can infer those professional investors also consider future expansion.

vi) Some researchers claim that S&P 500 highly concentrates on the Tech industry on purpose. However, whether the company is in the Tech industry does nothing in classification. Therefore, I can infer that financial data is more crucial for survival than a company's industry, and Tech companies are included because of their performance, not industry. Furthermore, whether the company originated from the US is never a crucial factor in survival.

Limitation and Further Research

A) Limitation

i) Insufficient delisting data in 2019's list: Only 5% of total companies are delisted by 2021, so classification models can not sufficiently learn what characteristics affect the delisting. Although the performance of accuracy in RF and NN models with 2019 is higher than 97%, it is wrong that the models predict every company in the 2021 list will survive.

ii) The number of hidden layers in Neural network models: I confirmed that with 100~150 hidden layers, the model's performance was 73.6%, while with 200~250 hidden layers, that of the model was 74.4%. Based on the result, I can anticipate that the performance will increase as the number of hidden layers increases. However, compiling the model with more than 200 layers took more than 20 minutes at once, so I could not compile with larger numbers.

iii) Limitation of data access: several factors can affect the survival, such as other industry categories, CEO's background information, the difference between financial analysts' forecasts and current price, and text data.

B) Further Research

i) Which companies will be included in the S&P 500: Forecasting delisting is a passive investment strategy, avoiding expected loss, whereas predicting enlisting is an active investment strategy since the stock price will soar when the information that S&P will include the stock in the index is released. Therefore, if an investor predicts which stock will be included, the investor can profit when soring.

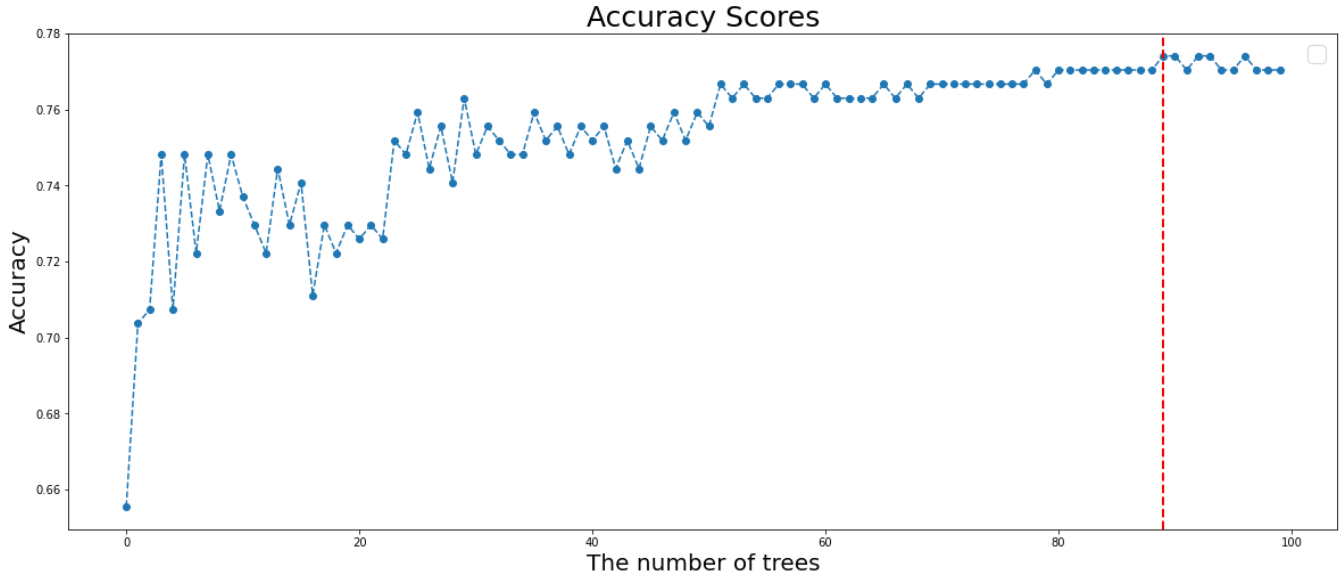
ii) Collecting more restricted data above and training with the data again

[**Appendix**]

A. Results with 2001 data

[Figure 1]

The Accuracy Scores analysis in Random Forest model with 2001 S&P 500 data



[Figure 2]

Accuracy Rank	The number of trees	Accuracy
1	90	0.7741
2	97	0.7741
3	94	0.7741
4	93	0.7741
5	91	0.7741

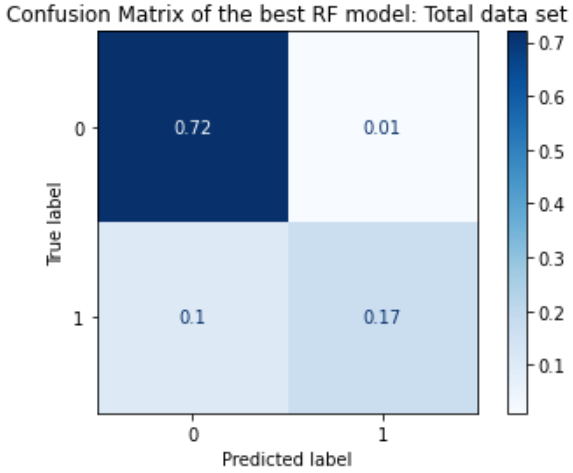
[Figure 3]

Hyper-parameters	Candidates for variables
The number of trees	97, 94, 93, 91, 90
Maximum Depth	10, 15, 20, 25
Maximum leaf nodes	25, 30, 35
Criterion	‘Gini’, ‘Entropy’

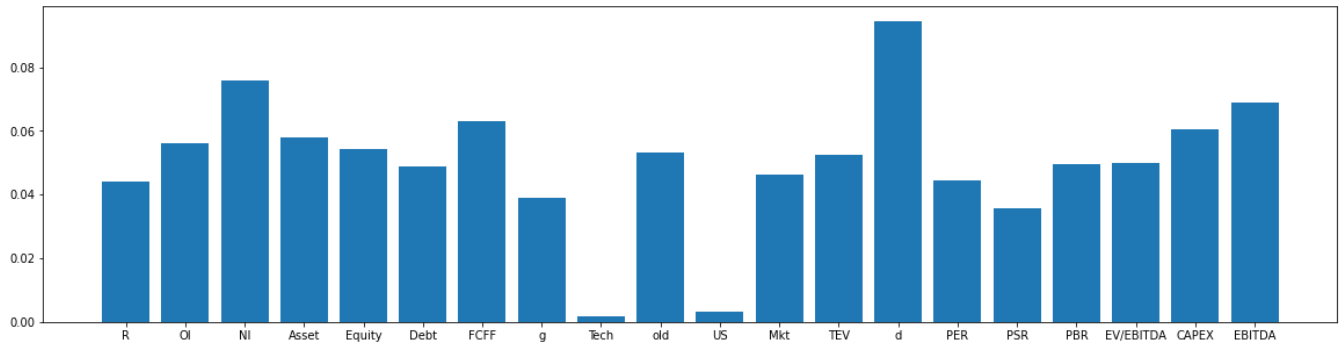
[Figure 4]
The accuracy analysis result with hyper-parameters candidates (Random Forest model with 2001 S&P 500 data)

	params	mean_test_score
56	{'criterion': 'gini', 'max_depth': 25, 'max_le...	0.777778
109	{'criterion': 'entropy', 'max_depth': 25, 'max...	0.777778
116	{'criterion': 'entropy', 'max_depth': 25, 'max...	0.774074
4	{'criterion': 'gini', 'max_depth': 10, 'max_le...	0.774074
89	{'criterion': 'entropy', 'max_depth': 15, 'max...	0.774074
..	...	...
94	{'criterion': 'entropy', 'max_depth': 20, 'max...	0.744444
30	{'criterion': 'gini', 'max_depth': 20, 'max_le...	0.740741
21	{'criterion': 'gini', 'max_depth': 15, 'max_le...	0.740741
104	{'criterion': 'entropy', 'max_depth': 20, 'max...	0.740741
88	{'criterion': 'entropy', 'max_depth': 15, 'max...	0.737037

[Figure 5]
Confusion matrix based on the adjusted Random Forest model with 2001 S&P 500 data



[Figure 6]
The importance of Features in the adjusted Random Forest model with 2001 S&P 500 data



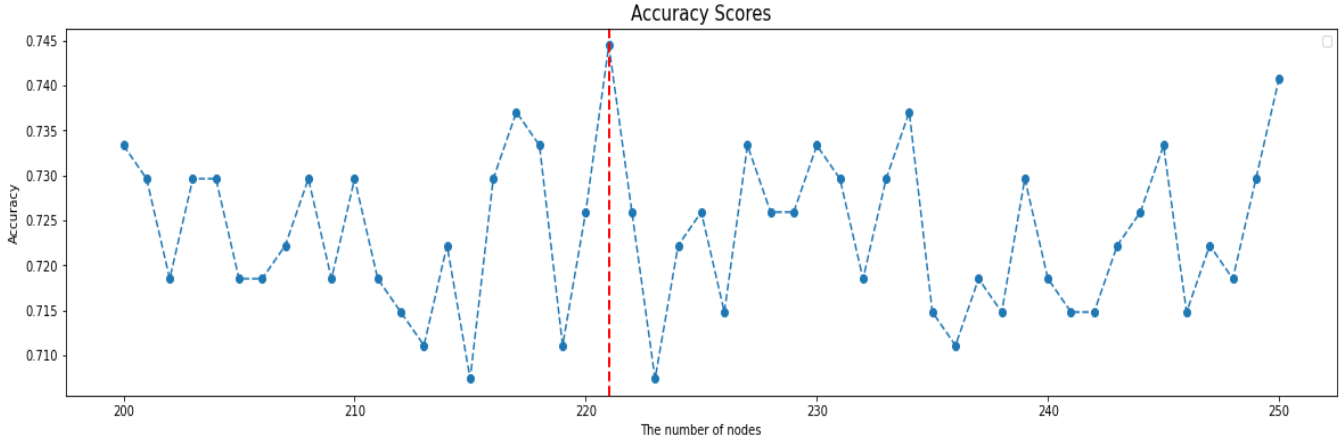
[Figure 7]
Full list of 61 companies which will be excluded from S&P 500 after 20 years, predicted by adjusted Random Forest model with 2001 data

◆ Boeing Co/The ◆ Walt Disney Co/The ◆ FleetCor Technologies Inc ◆ Exxon Mobil Corp ◆ Phillips 66 ◆ General Electric Co ◆ General Motors Co ◆ Under Armour Inc ◆ Royal Caribbean Cruises Ltd ◆ Ceridian HCM Holding Inc ◆ Berkshire Hathaway Inc ◆ Carnival Corp ◆ Lumen Technologies Inc ◆ IPG Photonics Corp ◆ Caesars Entertainment Inc ◆ Dexcom Inc ◆ Catalent Inc	◆ Hanesbrands Inc ◆ Viatris Inc ◆ Occidental Petroleum Corp ◆ PPL Corp ◆ Public Service Enterprise Group Inc ◆ Southwest Airlines Co ◆ Ulta Beauty Inc ◆ Marathon Oil Corp ◆ VF Corp ◆ Vornado Realty Trust ◆ Molson Coors Beverage Co ◆ Marriott International Inc/MD ◆ American Airlines Group Inc	◆ Host Hotels & Resorts Inc ◆ Incyte Corp ◆ Twitter Inc ◆ Waters Corp ◆ DXC Technology Co ◆ Valero Energy Corp ◆ Intuitive Surgical Inc ◆ Booking Holdings Inc ◆ Under Armour Inc ◆ CarMax Inc ◆ Wynn Resorts Ltd ◆ Live Nation Entertainment Inc ◆ Expedia Group Inc ◆ Discovery Inc ◆ Moderna Inc ◆ MGM Resorts International ◆ Mohawk Industries Inc	◆ Amcor PLC ◆ T-Mobile US Inc ◆ Delta Air Lines Inc ◆ United Airlines Holdings Inc ◆ DISH Network Corp ◆ Penn National Gaming Inc ◆ Aptiv PLC ◆ Illumina Inc ◆ Las Vegas Sands Corp ◆ Discovery Inc ◆ Biogen Inc ◆ Fox Corp ◆ Norwegian Cruise Line Holdings Ltd ◆ Alaska Air Group Inc

[Figure 8]

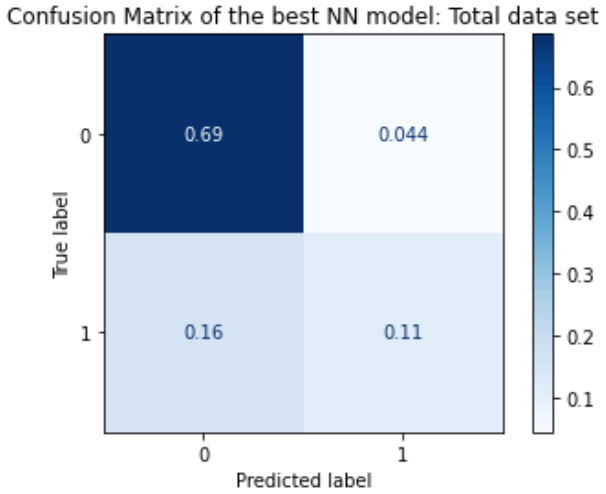
Hyper-parameters	Candidates for variables
The number of hidden layers	200~250
Maximum iterations	1,000, 2,000, 5,000, 10,000

[Figure 9]
The Accuracy Scores analysis in Neural Network model with 2001 S&P 500 data (5,000 maximum iterations)

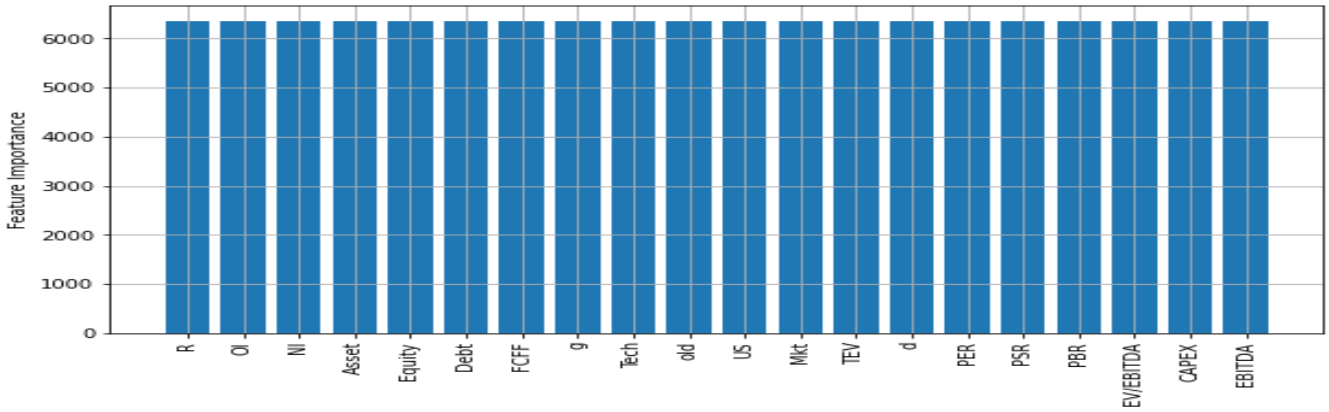


The highest performance is 0.7444 with 222 hidden layers and 5,000 maximum iterations

[Figure 10]
Confusion matrix based on the adjusted Neural Network model with 2001 S&P 500 data



[Figure11]
The importance of Features in the adjusted Neural Network model S&P 500 data



[Figure 12]
Full list of 18 companies which will be excluded from S&P 500 after 20 years, predicted by adjusted Neural Network model with 2001 data

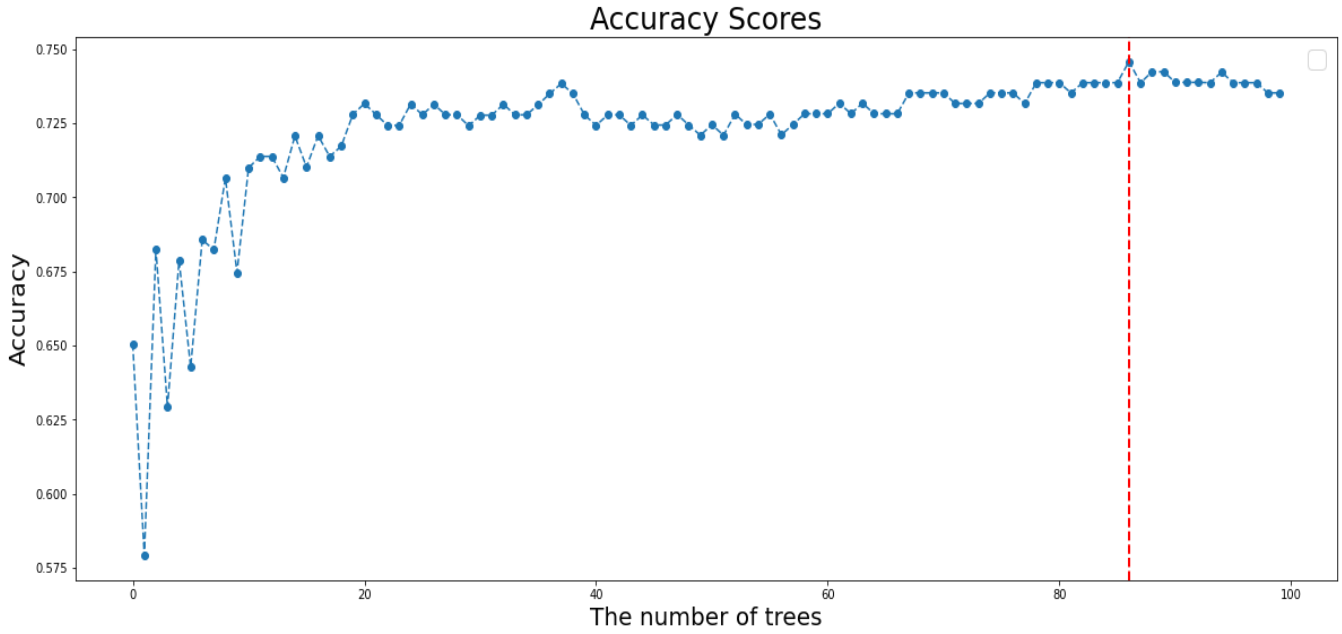
◆ Iron Mountain Inc ◆ American Tower Corp ◆ Boston Properties Inc ◆ TransDigm Group Inc ◆ Yum! Brands Inc ◆ SBA Communications Corp ◆ Wynn Resorts Ltd ◆ Penn National Gaming Inc ◆ Crown Castle International Corp ◆ Las Vegas Sands Corp	◆ McDonald's Corp ◆ Royal Caribbean Cruises Ltd ◆ Carnival Corp ◆ Caesars Entertainment Inc ◆ Fox Corp ◆ Norwegian Cruise Line Holdings Ltd ◆ Simon Property Group Inc ◆ Domino's Pizza Inc

[Figure 13]
Full list of 8 companies which will be excluded from S&P 500 after 20 years, predicted by both adjusted Random Forest model and adjusted Neural Network model with 2001 data

- Royal Caribbean Cruises Ltd
- Carnival Corp
- Caesars Entertainment Inc
- Fox Corp
- Norwegian Cruise Line Holdings Ltd
- Wynn Resorts Ltd
- Penn National Gaming Inc
- Las Vegas Sands Corp

B. Results with 2009 data

[Figure 14]
The Accuracy Scores analysis in Random Forest model with 2009 S&P 500 data



[Figure 15]

Accuracy Rank	The number of trees	Accuracy
1	87	0.7458
2	89	0.7424
3	90	0.7424
4	95	0.7422
5	93	0.7388

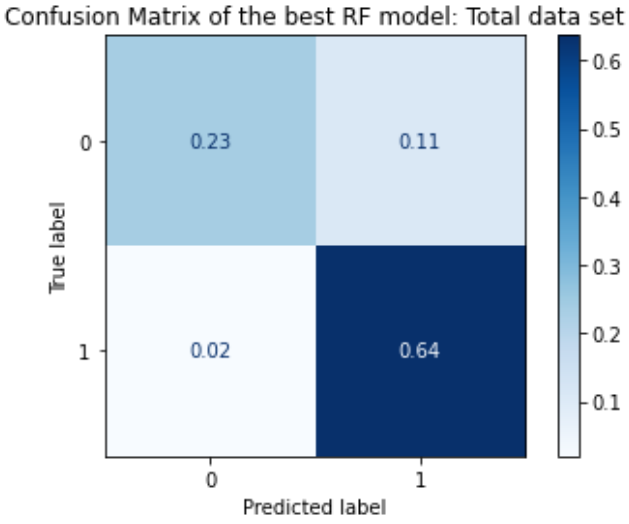
[Figure 16]

Hyper-parameters	Candidates for variables
The number of trees	87, 89, 90, 95, 93
Maximum Depth	10, 15, 20, 25
Maximum leaf nodes	25, 30, 35
Criterion	‘Gini’, ‘Entropy’

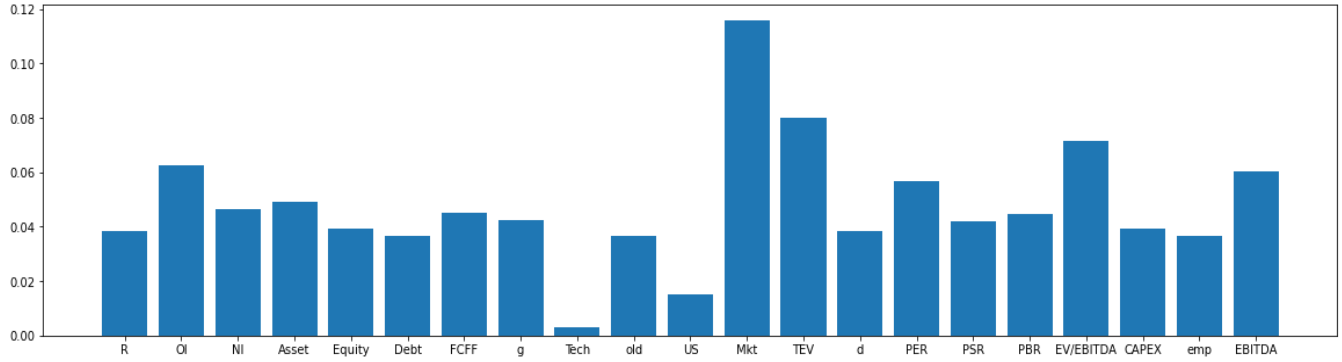
[Figure 17]
The accuracy analysis result with hyper-parameters candidates (Random Forest model with 2009 S&P 500 data)

	params	mean_test_score
1	{'criterion': 'gini', 'max_depth': 10, 'max_le...	0.749261
52	{'criterion': 'gini', 'max_depth': 25, 'max_le...	0.745936
91	{'criterion': 'entropy', 'max_depth': 20, 'max...	0.745813
109	{'criterion': 'entropy', 'max_depth': 25, 'max...	0.745813
54	{'criterion': 'gini', 'max_depth': 25, 'max_le...	0.745690
..	...	...
67	{'criterion': 'entropy', 'max_depth': 10, 'max...	0.717241
94	{'criterion': 'entropy', 'max_depth': 20, 'max...	0.713793
65	{'criterion': 'entropy', 'max_depth': 10, 'max...	0.713793
44	{'criterion': 'gini', 'max_depth': 20, 'max_le...	0.713670
88	{'criterion': 'entropy', 'max_depth': 15, 'max...	0.710222

[Figure 18]
Confusion matrix based on the adjusted Random Forest model with 2009 S&P 500 data



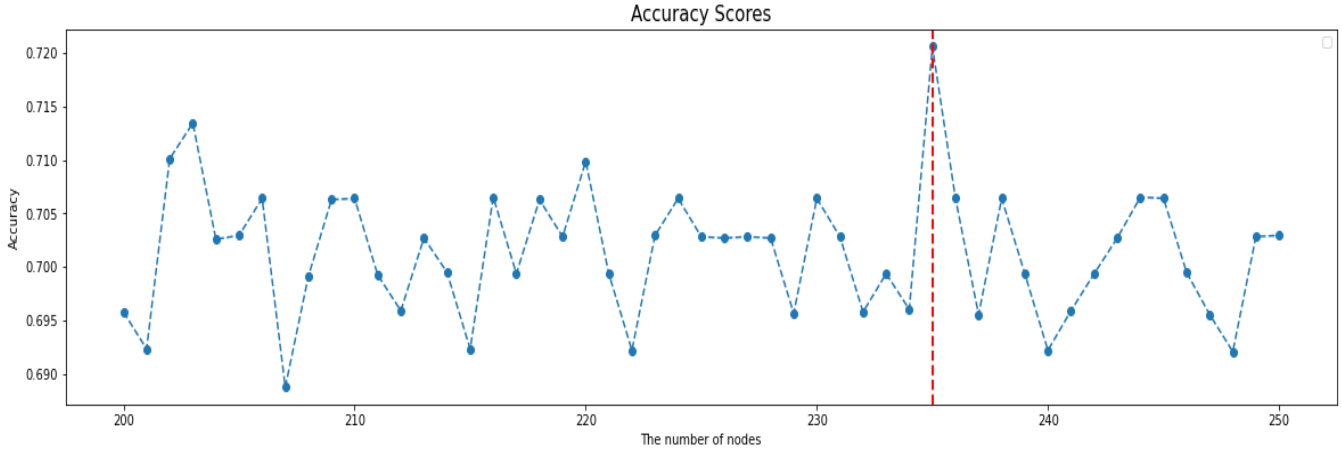
[Figure 19]
The importance of Features in the adjusted Random Forest model with 2009 S&P 500 data



[Figure 20]
Full list of 14 companies which will be excluded from S&P 500 after 12 years, predicted by adjusted Random Forest model with 2009 data

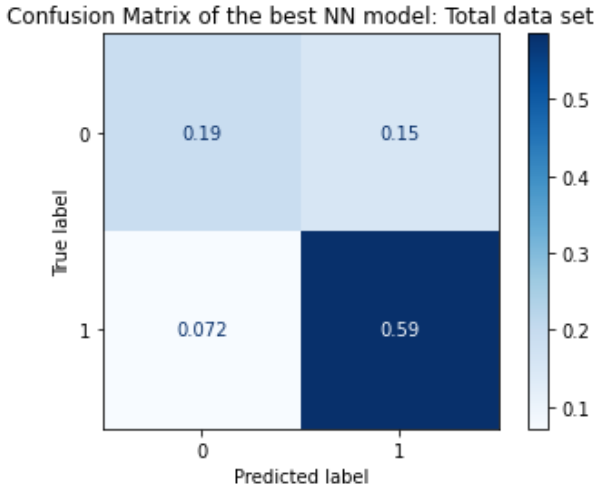
◆ Boeing Co/The ◆ Phillips 66 ◆ Royal Caribbean Cruises Ltd ◆ Carnival Corp ◆ American Airlines Group Inc ◆ Fox Corp ◆ Norwegian Cruise Line Holdings Ltd	◆ Expedia Group Inc ◆ Delta Air Lines Inc ◆ United Airlines Holdings Inc ◆ Las Vegas Sands Corp ◆ Host Hotels & Resorts Inc ◆ Wynn Resorts Ltd ◆ Live Nation Entertainment Inc

[Figure 21]
The Accuracy Scores analysis in Neural Network model with 2009 S&P 500 data (10,000 maximum iterations)



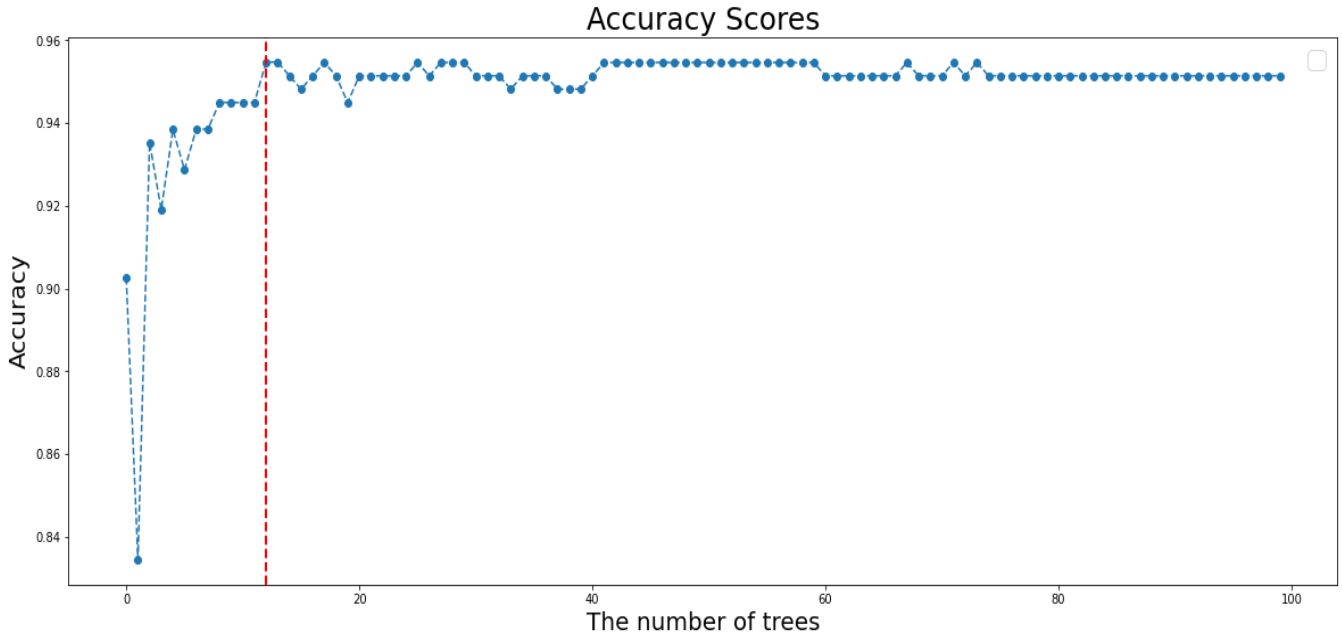
The highest performance is 0.7206 with 222 hidden layers and 10,000 maximum iterations

[Figure 22]
Confusion matrix based on the adjusted Neural Network model with 2009 S&P 500 data



C. Results with 2019 data

[Figure 23]
The Accuracy Scores analysis in Random Forest model with 2019 S&P 500 data



[Figure 24]

Accuracy Rank	The number of trees	Accuracy
1	13	0.9547
2	14	0.9547
3	50	0.9546
4	74	0.9546
5	30	0.9546

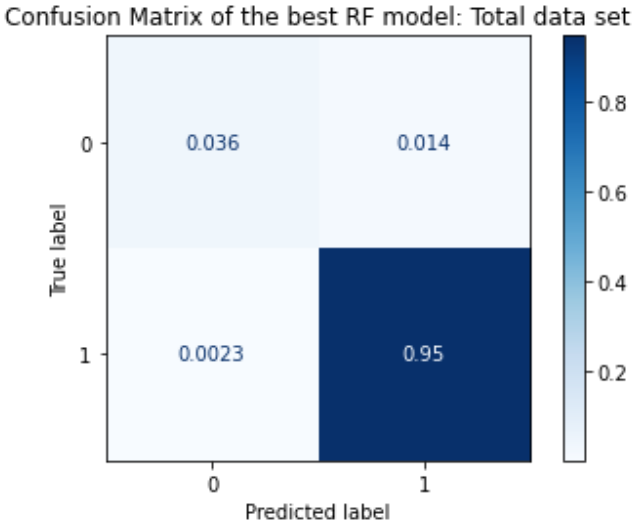
[Figure 25]

Hyper-parameters	Candidates for variables
The number of trees	13, 14, 50, 74, 30
Maximum Depth	10, 15, 20, 25
Maximum leaf nodes	25, 30, 35
Criterion	‘Gini’, ‘Entropy’

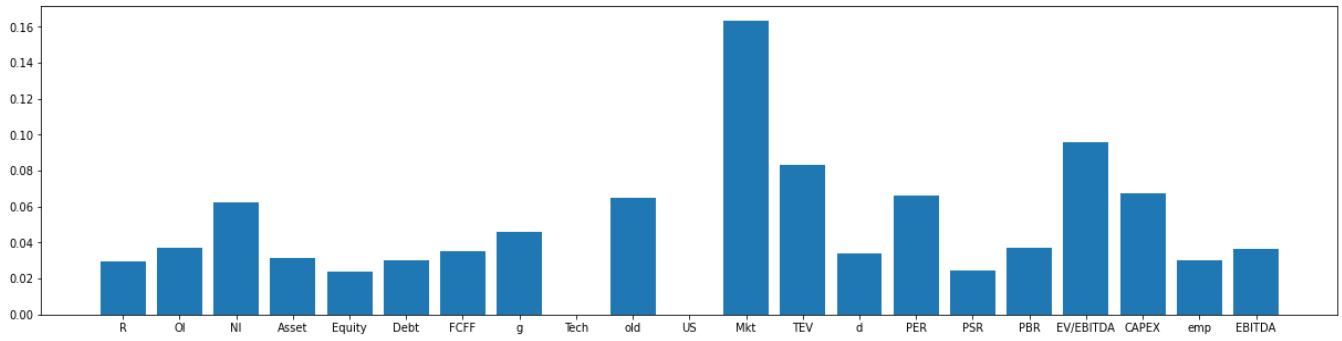
[Figure 26]
The accuracy analysis result with hyper-parameters candidates (Random Forest model with 2019 S&P 500 data)

	params	mean_test_score
95	{'criterion': 'entropy', 'max_depth': 20, 'max...	0.961183
37	{'criterion': 'gini', 'max_depth': 20, 'max_le...	0.957957
74	{'criterion': 'entropy', 'max_depth': 10, 'max...	0.957957
112	{'criterion': 'entropy', 'max_depth': 25, 'max...	0.957849
119	{'criterion': 'entropy', 'max_depth': 25, 'max...	0.954731
..	...	...
41	{'criterion': 'gini', 'max_depth': 20, 'max_le...	0.941720
5	{'criterion': 'gini', 'max_depth': 10, 'max_le...	0.941720
51	{'criterion': 'gini', 'max_depth': 25, 'max_le...	0.941720
6	{'criterion': 'gini', 'max_depth': 10, 'max_le...	0.941613
91	{'criterion': 'entropy', 'max_depth': 20, 'max...	0.938387

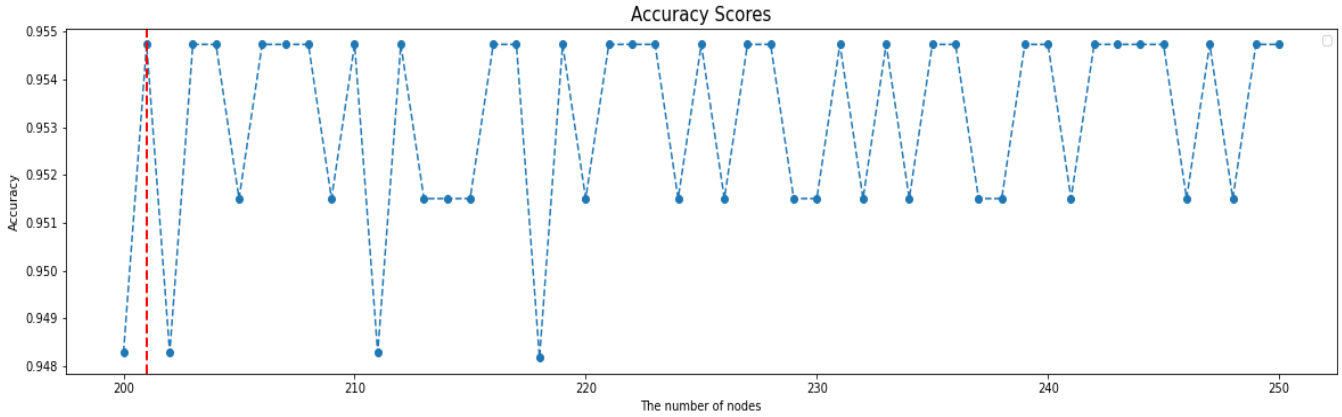
[Figure 27]
Confusion matrix based on the adjusted Random Forest model with 2019 S&P 500 data



[Figure 28]
The importance of Features in the adjusted Random Forest model with 2019 S&P 500 data

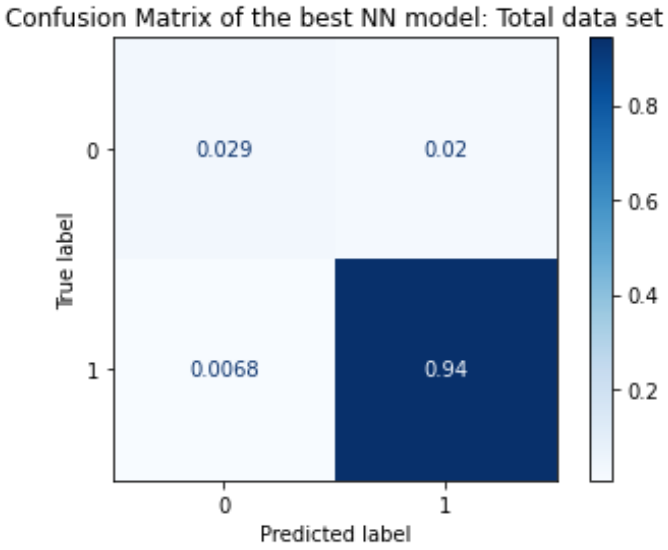


[Figure 29]
The Accuracy Scores analysis in Neural Network model with 2019 S&P 500 data (10,000 maximum iterations)



The highest performance is 0.9547 with 202 hidden layers and 10,000 maximum iterations

[Figure 30]
Confusion matrix based on the adjusted Neural Network model with 2019 S&P 500 data



[Figure 31]
Anticipated by three models, Random Forest models with 2001 and 2009 data and Neural Network model with 2001 data, The prediction that 6 stocks will be excluded from S&P 500 within 20 years

Companies	RF2001	NN2001	RF2009	Risk
Boeing Co/The	0	1	0	Risky
Phillips 66	0	1	0	Risky
Royal Caribbean Cruises Ltd	0	0	0	More Risky
Carnival Corp	0	0	0	More Risky
American Airlines Group Inc	0	1	0	Risky
Fox Corp	0	0	0	More Risky
Norwegian Cruise Line Holdings Ltd	0	0	0	More Risky
Host Hotels & Resorts Inc	0	1	0	Risky
Wynn Resorts Ltd	0	0	0	More Risky
Live Nation Entertainment Inc	0	1	0	Risky
Expedia Group Inc	0	1	0	Risky
Delta Air Lines Inc	0	1	0	Risky
United Airlines Holdings Inc	0	1	0	Risky
Las Vegas Sands Corp	0	0	0	More Risky

[Figure 31]
Five highest importance from Random Forest Models with 2001, 2009, and 2019 data

Variables	RF2001	RF2009	RF2019
Dividend rate	**1**		
Net Income	**2**		
EBITDA	3	4	
FCFF	4		
CAPEX	**5**		**4**
Market Capital		**1**	**1**
Total EV		**2**	**3**
EV/EBITDA		**3**	**2**
Operating Income		5	
PER			5