Emmy Fortunato, Sean Jung, Stephen Thomas, and Qiyu Wang
ISOM 676 - Machine Learning II
Final Project
4/12/2021

Predicting Female Representation on Company Boards

Gender diversity within organizations, especially at the board level, has become an increasingly important factor for the long-term growth and success of businesses. Historically, females have comprised only a small fraction of total board members, and though the number of females on boards has generally increased over time, a large gender discrepancy still exists in most organizations. A statistic from 2017 revealed that women directors made up less than 20% of board members within companies on the Russell 1000 stock index[1]. Not only have organizations with gender-diverse boards proven to approach business problems more creatively, but they also tend to compensate females in lower positions more generously, resulting in narrower wage gaps between men and women. This lack of female board representation has motivated our team to investigate the factors that contribute to the percentage of females on a board, and ultimately build predictive models to estimate the female board composition for a number of companies across various industries. After predicting female board representation, we hope to identify specific industries that are lacking diverse board representation and the main features of organizations that perpetuate low female board representation.

We obtained three different data sets from Wharton Research Data Services (WRDS) that consisted of information about companies in the S&P 1500 related to 1) director qualities and compensations (*ExecuComp*), 2) year-end company-level financial ratio metrics (*Industry Financial Ratio*), and 3) annual metrics regarding directors' tenure on their respective boards and in their respective companies (*BoardEx - Organization summary analytics*).

[1] https://insights.diligent.com/board-diversity/the-importance-of-gender-diversity-in-the-boardroom

In the *Execucomp* dataset, there are 107 columns and 107,908 instances where each observation is specified by unique combination of company (2,162 unique) and year (2010-2021), and it contains information regarding a given director's annual compensation and stock awards in addition to general information about the director as it relates to that year (i.e., title, gender, age, etc.). The *Industry Financial Ratio* dataset containing key financial ratios has 480,739 observations and 76 columns. The ratios used comprise seven different categories of measurement: capitalization, efficiency, financial soundness/solvency, liquidity, profitability, valuation, and another category for ratios measuring what is not included in other categories. Each row corresponds to the various financial ratio metrics for a specific company, and industry information is also included in each observation. Overall, the data is fairly clean, but several columns contain an abundance of missing values. The *BoardEx* data related to organization summary analytics contains 757,192 observations and 60 columns, and it captures the overall size of each director's network, and the degree of nationality diversity of their board.

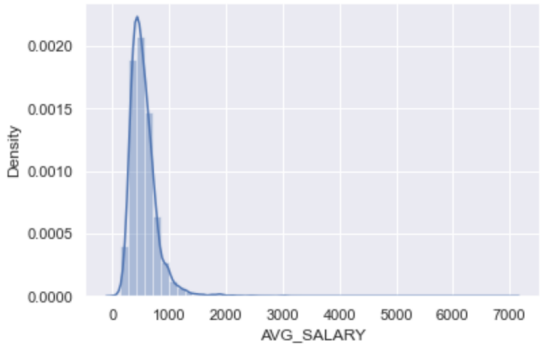
Our target variable, FEMALE_PCT, is the percentage of females on the board of a given company during a given year. This metric is calculated by dividing the number of women on the board by the total number of board members. The predictive attributes we used to perform the analysis are summarized as shown below:

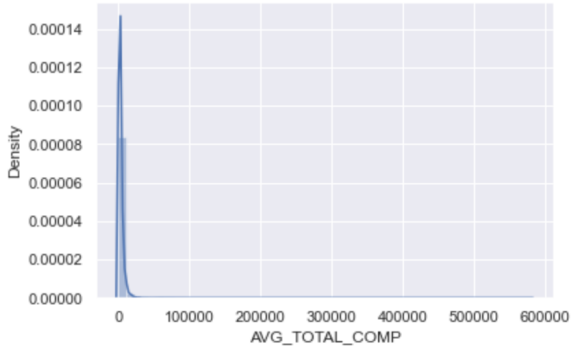
1) Board Characteristics: variables related to number of board members, average salary, age, role, network size, and other related details
2) Financial metrics of the firms (e.g., dividend payout ratio, return on assets)
3) Dummy variables indicating year from which the particular observation is seen (e.g., x0_2015)

4) Dummy variables indicating industry to which the firm belongs (e.g. x1_health)

5) Dummy variables indicating state in which the firm is located (e.g., x2_GA)

Exploratory Data Analysis





Next, as a part of data understanding, the variables, *'AVG_SALARY'* and *'AVG_TOTAL_COMP'* are log transformed such that the data resembles normally distributed data as shown above.

FEMALE_PCT_LAST	0.842876
CEO_FEMALE	0.341347
x1_General Retailers	0.123554
NetworkSize	0.083740
GProf	0.076943
x1_Utilities - Other	0.071347
x0_2019	0.067328
x2_MN	0.058776
at_turn	0.050807
adv_sale	0.050115

Then, correlation between a set of 2 numeric variables is computed to quickly and concisely comprehend the direction and magnitude of the relationships. Upon close observation, we notice that most collinearity happens between the different financial metrics rather than a set of variables that are different in nature.

```
Compute correlation between numeric features and target variables
FEMALE_PCT_LAST                 0.842876
CEO_FEMALE                      0.341347
x1_General Retailers            0.123554
NetworkSize                     0.083740
GProf                           0.076943
```

We have computed the correlation of each feature to the target variable, *'FEMALE_PCT'*. It can be seen from the figure above that the variables indicating female percentage of the last year and CEO being the female also have a considerable level of correlation to the female percentage in a given year for a given company.

To prepare the data for our modeling phase, we used PySpark and SparkSQL to merge the three datasets based on a common key; we used the *'TICKER'* column to merge the ExecuComp and financial ratio tables together, and then aggregated each instance by company for a specific year. Thus, each instance now represents a specific company and its financial metrics and number of board members for a given year. We dropped certain columns containing a large portion of missing values or including irrelevant information, as these would not provide any use in predicting female board concentration. To engineer the target variable (female percentage on the board), we divided the number of females on the company board by the total number of members on the board for a given year. Depending on the availability of data, some company boards have a female percentage value for 2010 thru 2019, while others have proportions for 2020 as well. After merging each data set, the final table contained 17,259 observations and 106 columns.

One of the main concerns with our merged data is how to handle the many missing values present in some of the columns. We decided to drop columns containing over 4,000 missing values, as these would not offer much predictive power in our models. Additionally, we normalized the variables since we are using methods such as *k*-NN and Neural Networks. We

didn't directly perform cox-box transformations on our independent variables given that we want to keep a clean interpretation whereby we can directly infer, for example, how a dollar difference in salary will impact the female percentage on the board. For other null values in the dataset, we filled in values with the average of a particular industry sector and year for each column. We handled missing values using yearly average at industry-level instead of doing it at company-level, because some companies don't have any data point for a particular financial index. Thus, the yearly average at industry-level would provide us the best information available to estimate the yearly approximate financial indexes for each company.

For ExecuComp and BoardEx, we both get the yearly average and yearly standard deviation for each metrics at company-level. Thus we can measure both the impact of the average level of those metrics on female percentage on the board and also how the deviation of them, such as variance in salary level or age among the board members, will impact our target variable. Moreover, we use the last month's financial metrics for each year and each company, since it's the year end records that present the snapshot or summary of companies' performances of the year.

In addition to summary statistics, we engineered multiple other features:

1. CEO_FEMALE: whether the CEO is female or not, to see whether female CEO will impact the gender equity situation in a company
2. EXECDIR_COUNT: Number of board members who served as executive director that year
3. EXEC_COUNT: Number of board members

4. REJOIN: Number of board members that are rejoined members, since the old days tend to have more male board members, having more rejoined members might indicate a higher percentage of male on the board
5. Length_term: how long the CEO has been taking the position
6. We transformed categorical features such as STATE, YEAR, SECTOR (Industry information) into dummy variables. Also we removed one dummy from each set (using k-1 in total after removal), to reduce the additional multicollinear that can be introduced by extra dummies.
7. We tried to Calculate male-to-female salary ratio, however, too many companies come in without a female board member, which makes it not an ideal predictor.
8. Last year's percentage of female representation for each company. This variable is important because our target variable is correlated to itself. To be more specific, since most board members will serve more than one year, the percentage of female representation is highly likely to create autocorrelation issues, which can be potentially solved by introducing last year's percentage of female representation as an independent variable into our prediction. For doing so, we also lost our first year's data, which leaves us with 16747 complete cases.

The target variable we are trying to predict is a numeric/fractional value, a percentage of female representation in each company for each year, thus we tested a variety of models that support numerical prediction, including regression (OLS, ridge, lasso, and elastic net), ensemble methods (Adaboost, bagging, and stacking), and other ensemble tree methods we didn't cover in class (gradient boosted regression trees, hist gradient boosted trees, and Extra Trees regression),.

We did an 80/20 train/test split to evaluate the final performance and used cross-validation with five folds for our training set to optimize the hyperparameters of each model. For each of the models adopted in this project, we have taken the ability of each model to meet the requirements of the task (such as generalization performance, comprehensibility, speed of learning and application, amount of data required, type of data, and missing values) into considerations. As each model has its own advantages and disadvantages, we have tried different modeling techniques and compared the nature and performance of each model. We discuss only our top performing models in the report, but in the **(Appendix 2)**, results of each methodology we examined are displayed.

First model category we want to look at is the linear models, which train fast and provide clear interpretations of the impacts of each individual predictors on the percentage of female representation. From there, we introduced regularized linear models to perform automatic feature selection for us. As a result, Ridge regression presents as the top performer among the linear category, which we later used to interpret how companies management decision and financial status can influence gender equity among board members.

To serve the other goal of our project to produce future predictions of percentage of female representation on board, we further proceed with ensemble methods to better fit the nonlinear patterns in our dataset. We first went over the ones we discussed in class, such as Random Forest, Boosting, and Stacking, then we tried some more advanced and new ensemble methods such as Extra trees or Gradient descent trees. The outcomes show that Boosting performs better than Decision Tree but surpassed by Bagging, Extra Trees and Random Forest, which might suggest the base models (Decision tree and OLS) have a lower bias but higher

variance, since Boosting tends to reduce bias and Bagging tends to reduce more variance than Boosting. However, the best performing ensemble method happens to be a Stacking model using Neural Nets, Decision Tree and Ridge Regression with a stacking method of linear regression. Unlike boosting or bagging, Stacking ensembles the predictions of the different learning algorithms with the original training set to improve predictions using a meta-model. With imputing so many null values in the original dataset, we might include a lot of variance and bias into our data, where stacking can reduce both bias and variance in some cases and potentially produce an even better result.

Our goal for this project was to train a model that would predict (with minimal error) the female proportion of directors on a Board of Directors. This model can be used to better understand which elements of a business can tell us most about female representation on Boards, whether it be how their executives are compensated, the industry that they belong to, the amount of debt that they've incurred over a year, or even the state that the company is headquartered in. For this reason, we've decided to evaluate the "goodness" of our model using a combination of the R^2 value, indicating how well the model explains the variance in the percentage of female board members, Root Mean Squared Error, indicating the standard deviation of the prediction error, and Mean Absolute Error, indicating the average difference between the value of the target variable predicted by the model and the actual value of the target variable. The result of our model should be evaluated using these measures because, as previously mentioned, our target variable is a continuous numeric value, so these measures help us to better understand the accuracy with which our model makes predictions. By minimizing the error in our predictions

and maximizing the R^2 score of the model, we can also ensure that the features used in our model can genuinely be attributed to predicting our target variable.

A naive method of predicting the percentage of female board members at a given company in a given year would rely simply on the percentage from the previous year, assuming no change year over year. A generalized linear regression model that used only the percentage from the previous year resulted in a R^2 value of .710, RMSE of .072, and MAE of .045.

The R^2 value of our final model came in at .856, indicating that the model is able to explain 85.6% of the variance in our target variable, the RMSE is .050, and the MAE of our final model is .029, indicating that on average, the model's predictions are off by a value of about 3% of directors that are female. Given that the majority of the boards observed in our dataset contained fewer than 10 directors, a 3% miss in prediction would likely still give an accurate estimate of the number of women who sit on the board.

While we don't have one specific firm for which we are exploring gender diversity, we built this model with the intention of it being of value to *any* company as they benchmark themselves against competitors in their industry and across industries in North America. Companies are under a lot of pressure to introduce diversity into their boards, especially those companies that have a history of heterogeneity. Studies have shown that in addition to achieving increased creativity in problem-solving and closing the gender wage gap, companies and business units with higher gender diversity often demonstrate higher financial success than their counterparts with lower gender diversity.[2] Research from Mckinsey shows that company profits

[2] https://bit.ly/3uHrYDl

and share performance can be 50 percent higher when women are well represented at the top. Beyond that, senior-level women have a vast and meaningful impact on a company's culture[3].

This model can serve as an explanatory tool for companies to better understand which factors have the most impact on gender diversity across top companies nationwide. From the regression output of this model, we can draw some few conclusions regarding the variables with the most significant impact on gender diversity. One of the most predictive variables of the level of female representation is *'CEO_Female'*, a binary variable indicating the gender of the CEO of that company. Although CEOs are not always on their company's board of directors, companies with female CEOs are much more likely to see higher levels of female representation on their boards than those with male CEOs.

From the variables that capture information about the active members of the board in question, we learned that boards whose members have large Networks and boards with higher values for *'NationalityMix'* - the percentage of non-National board members - also tend to have higher values for percentage of females on their boards, indicating that boards with a broader reach are likely to have greater gender diversity and that gender diversity may be tied to overall board diversity. Variables like Average Age and Average Salary of Board Members are both negatively correlated with Female Percentage, implying that boards with an older average age and higher average salaries are less likely to see high female representation among their members; however, variables like Standard Deviation of Age and Standard Deviation of Time on

[3] https://mck.co/3g3ue42

the Board are positively correlated with Female Percentage, indicating that wider diversity in age and experience on a board is tied with higher female representation a board.

Of the year-end financial ratios, variables that are negatively correlated with percentage of board members that are female are Return on Assets, Pre-Tax Profit Margin, Current to Total Liabilities Ratio, and Long Term Debt over Total Liabilities while variables that positively correlated with our target variable include Receivables to Current Assets, Inventory to Current Assets, Common Equity over Invested Capital, and Total Debt over Invested Capital. From these variables, we can discern that companies that have high profits and low debt generally have lower female representation on their boards, and companies with lower invested capital values and lower asset values may have a higher percentage of female board members. These inferences are validated when we look at the impact of each industry sector on female percentage.

Some of the features that are most indicative of female representation are the dummy variables that represent each of the different industry sectors to which a company belongs. Industry sectors that are academic (Education, Publishing), consumer-facing (Consumer Services, General Retailers, Clothing and Personal Products), or arts-and-travel-based (Media and Entertainment, Leisure and Hotels) in nature generally have high female representation on boards. Meanwhile, industry sectors that center more closely around construction, fuel, manufacturing, and technology (including Construction & Building Materials, Steel & Other Metals, Oil & Gas, Electronic & Electrical Equipment, Automobiles & Parts, and Telecommunications Services) generally have low female representation on boards. Although the push for gender diversity in leadership is not limited to specific industries, the low representation

of women on boards of more industrial organizations illustrates a clear divide in the kinds of organizations that are actively appointing female board members and those that are slower to adapt.

To use this model for benchmarking, at the start of each fiscal year, companies can compare the model's prediction based on their past year's data with the actual percentage of females on the board at the start of that new year. If their actual percentage is significantly *lower* than the model's prediction, then the company ought to flag that the gender diversity on their board is below what it ought to be. They can treat the percentage predicted by the model as an expected value for the percentage of females on their board based on their industry, financial performance, and the current board demographic makeup. They can then establish this prediction as a minimal acceptable percentage of directors who are female, letting it operate as a baseline for achieving gender diversity on their board.

Given that this problem is rooted in Social Sciences and Human Resources, we would advise our clients to ensure that in their pursuit of gender diversity and understanding the elements that contribute to robust female representation in leadership, they do not accidentally end up discriminating against other groups of people. As mentioned before, average Age of board members in an indicator of low female representation. However, it could be seen as discrimination and ageism to classify average Age as a "negative" attribute of a board.

In closing, we found that stacking with base learners ridge regression, neural nets, and decision tree regressor and a stacking model of a generalized linear regression gave us the best performance when predicting the percentage of women on a given board of directors. While

modeling was a large part of our process, the majority of our time and effort was spent researching the problem to understand what data, attributes, and methods other researchers had found were correlated with gender diversity, then finding and engineering the appropriate data to more accurately make the predictions that we hoped to make. We learned from the model that diversity in age, salary, and tenure on boards is a good indicator of gender diversity on the board.

We also learned that companies that belong to more industrial sectors with that high-value assets like technology and oil generally don't have as many women on their boards as less industrial, more artistic, and more people-centric sectors. We recommend, that as more companies across more industries are urged or inspired to diversify the gender makeup of their boards, they should use annual predictions from our model to better understand the expected percentage of female directors on their board given their current board makeup and fiscal situation and use that as benchmark to maintain and even exceed as their boards grow and change over time.

Appendix A: Data Description

YEAR	year
CONAME	company name
TICKER	stock ticker code
NAICSDESC	Description
INDDESC	Industry Description
STATE	State
CFO	Is the CFO on the board?
CEO	Is the CEO on the board?
CEO_FEMALE	Is the CEO of the company female?
EXECDIR_COUNT	Number of board members who served as executive director that year
EXEC_COUNT	Number of board members
AVG_SALARY	Average base salary of members
AVG_TOTAL_COMP	Average total annual compensation of members
AVG_AGE	Average age of members
FEMALE_COUNT	Number of women on the board
FEMALE_PCT	Percentage of Board Members that are Female
CAPEI	Shiller's cyclically adjusted price-to-earnings ratio
bm	Book/Market
evm	Enterprise Value Multiple
pe_op_basic	Price/Operating Earnings (basic)
pe_op_dil	Price/Operating Earnings (diluted)
pe_exi	Price /Earnings, excluding Extraordinary Items
pe_inc	Price/Earnings, including Extraordinary Items
ps	Price/Sales
pcf	Price/Cash Flow

dpr	Dividend Payout Ratio
npm	Net Profit Margin
opmbd	Operating Profit Margin before depreciation
opmad	Operating Profit Margin after depreciation
gpm	Gross Profit Margin
ptpm	Pre-Tax Profit Margin
cfm	Cash Flow Margin
roa	Return on Assets
roe	Return on Equity
roce	Return on Capital Employed
efftax	Effective Tax Rate
aftret_eq	After-Tax Return on Average Common Equity
aftret_invcapx	After-Tax Return on Invested Capital
aftret_equity	After-Tax Return on Total Stockholders' Equity
pretret_noa	Pre-Tax Return on Net Operating Assets
pretret_earnat	Pre-Tax Return on Total Earning Assets
GProf	Gross Profit/Total Assets
equity_invcap	Common Equity/Invested Capital
debt_invcap	Long-Term Debt/Invested Capital
totdebt_invcap	Total Debt/Invested Capital
capital_ratio	Capitalization Ratio
int_debt	Interest/Average Long-Term Debt
int_totdebt	Interest/Average Total Debt
cash_lt	Cash Balance/Total Liabilities
invt_act	Inventory/Current Assets
rect_act	Receivables/Current Assets

debt_at	Total Debt/Total Assets
debt_ebitda	Total Debt/EBITDA
short_debt	Short-Term Debt/Total Debt
curr_debt	Current Liabilities/Total Liabilities
lt_debt	Long-Term Debt/Total Liabilities
profit_lct	Profit Before Depreciation/Current Liabilities
ocf_lct	Operating CF/Current
cash_debt	Cash Flow/Total Debt
fcf_ocf	Free Cash Flow/Operating Cash Flow
lt_ppent	Total Liabilities/Total Tangible Assets
dltt_be	Long-Term Debt/Book Equity
debt_assets	Total Debt/Total Assets
debt_capital	Total Debt/Capital
de_ratio	Total Debt/Equity
intcov	After-Tax Interest Coverage
intcov_ratio	Interest Coverage Ratio
cash_ratio	Cash Ratio
quick_ratio	Quick Ratio (Acid Test)
curr_ratio	Current Ratio
cash_conversion	Cash Conversion Cycle (Days)
inv_turn	Inventory Turnover
at_turn	Asset Turnover
rect_turn	Receivables Turnover
pay_turn	Payables Turnover
sale_invcap	Sales/Invested Capital
sale_equity	Sales/Stockholders Equity

sale_nwc	Sales/Working Capital
rd_sale	Research and Development/Sales
adv_sale	Advertising Expenses/Sales
staff_sale	Labor Expenses/Sales
accrual	Accruals/Average Assets
ptb	Price/Book
PEG_trailing	Trailing P/E to Growth Ratio
divyield	Dividend Yield
TimeRetirement	Average time to retirement
TimeRole	Average time in current role
TimeBRD	Average time on board
TimeInCo	Average time in company
TotNoLstd	Total Number of Listed Boards sat on (avg)
TotNoUnLstd	Total Number of Unlisted Boards sat on (avg)
TotNoOthLstdBrd	Total Number of Other Boards sat on (avg)
TotCurrNoLstdBrd	Total Number of Current Listed Boards sitting on (avg)
TotCurrNoUnLstd	Total Number of Current Unlisted Boards sitting on (avg)
TotCurrNoOthLstdBrd	Total Number of Current Other Boards sitting on (avg)
NoQuals	Number of Qualifications (avg)
Sector	Sector Name
Succession	Succession Factor
Attrition	Attrition Rate
NationalityMix	% of Board Members not White
STDEVTimeBrd	St Dev time on board
STDEVTimeInCo	St Dev time in company
STDEVTotNoLstdBrd	St Dev Total Number of Listed Boards sat on

STDEVNoQuals	St Dev Number of Qualifications
STDEVAge	St Dev Age of Board Members
NetworkSize	Average Network Size of Board Members

Appendix B: Table Summarizing Regression Output

Ordinary Least Squares Regressions

	OLS (statistical analysis)	OLS (machine learning)
R-squared	0.722	0.708
MAE	n/a	0.044
RMSE	n/a	0.005

Ridge, Lasso, Elastic Net Regresion

	Ridge (L2) Regression	Lasso (L1) Regression	Elastic Net (L1 +L2) Regression
R-squared	0.720	0.635	0.703
MAE	0.043	0.061	0.048
RMSE	0.005	0.005	0.005

Best Parameters Used:
Ridge Regression: {'alpha': 1}
Lasso Regression: {'alpha': 0.001}
Elastic Net Regression: {'alpha': 0.001, 'l1_ratio': 0.5}

Ensemble Regressions

	Stacking Regressor	AdaBoost Regressor	Bagging Regressor
R-squared	0.856	0.702	0.713
MAE	0.029	0.047	0.048
RMSE	0.002	0.005	0.005

Best Parameters Used:
Stacking: {Ridge('alpha': 1); Neural Nets('Training cycles': 200, 'learning rate': 0.01, 'momentum':0.9); DecisionTree('criterion': mse, 'max_depth':3, 'min_split':2)}
AdaBoost: {'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 150}
Bagging: {'bootstrap':False, 'max_features':100, 'max_samples':10000}

Tree-based Regressions

	Decision Tree Regression	Random Forest Regressor	ExtraTrees Regressor	Gradient Boosted Tree Regression	Hist Gradient Boosting Regression
R-squared	0.692	0.759	0.762	0.692	0.732
MAE	0.044	0.038	0.039	0.043	0.042
RMSE	0.005	0.004	0.004	0.005	0.005

Best Parameters Used:
Decision Tree Regression: {'criterion': 'mse', 'max_depth': 3, 'min_samples_split': 2, 'splitter': 'best'}
Random Forest Regression: {'bootstrap': True, 'max_depth': None, 'max_leaf_nodes': None, 'min_samples_split': 3}
Extra Trees Regression: {'bootstrap': True, 'max_depth': None, 'max_leaf_nodes': None, 'min_samples_split': 2, 'min_sample_leaf': 2}
Gradient Boosted Tree Regression: {'learning_rate': 0.1, 'max_depth': 3, 'min_samples_split': 3}
Hist Gradient Boosting Regression: {'l2_regularization': 5, 'learning_rate': 0.1, 'max_depth': None, 'max_iter': 150}

Appendix C: Regression Output - Final Stacking Model

```
Model Metrics Type: RegressionGLM
 Description: N/A
 model id: rm-h2o-model-generalized_linear_model_(2)-203
 frame id: rm-h2o-frame-generalized_linear_model_(2)-203
 MSE: 0.0025529657
 RMSE: 0.050526883
 R^2: 0.8555753
 mean residual deviance: 0.0025529657
 mean absolute error: 0.029992335
 root mean squared log error: 0.044321306
 null DOF: 16724.0
 residual DOF: 16525.0
 null deviance: 295.6444
 residual deviance: 42.698353
 AIC: -51991.11
GLM Model (summary):
   Family     Link Regularization Number of Predictors Total Number of Active Predictors
 gaussian identity           None                  203                            199
Scoring History:
            timestamp   duration iterations negative_log_likelihood objective
 2021-04-11 19:28:57  0.000 sec          0               295.64438   0.01768

H2O version: 3.30.0.1-rm9.8.1
```

Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value ↑
base_prediction2	0.879	0.107	0.009	93.752	0
base_prediction1	0.457	0.055	0.013	34.966	0
FEMALE_PCT_LA...	-0.039	-0.039	0.002	-25.762	0
Intercept	-0.032	0.097	0.001	-24.083	0
x0_2019	0.003	0.003	0.001	6.328	0.000
CEO_FEMALE	-0.002	-0.002	0.000	-4.184	0.000
NetworkSize	0.002	0.002	0.001	3.988	0.000
AVG_AGE	-0.002	-0.002	0.000	-3.341	0.001
capital_ratio	-0.003	-0.003	0.001	-2.912	0.004
x1_Clothing & Per...	0.002	0.002	0.001	2.896	0.004
AVG_SALARY	-0.002	-0.002	0.001	-2.776	0.006
cash_debt	-0.001	-0.001	0.001	-2.637	0.008
STD_SALARY	0.002	0.002	0.001	2.630	0.009
x0_2017	0.001	0.001	0.001	2.450	0.014
STDEVTimeInCo	0.004	0.004	0.002	2.383	0.017

staff_sale	-0.005	-0.005	0.002	-2.338	0.019
debt_ebitda	0.001	0.001	0.000	2.302	0.021
x0_2016	0.001	0.001	0.001	2.297	0.022
x1_General Retail...	0.002	0.002	0.001	2.276	0.023
ps	-0.002	-0.002	0.001	-2.119	0.034
x1_Engineering &...	0.002	0.002	0.001	2.097	0.036
x1_Forestry & Pa...	0.001	0.001	0.001	2.068	0.039
x2_NC	0.002	0.002	0.001	1.916	0.055
x1_Banks	0.002	0.002	0.001	1.902	0.057
TimeInCo	-0.003	-0.003	0.002	-1.885	0.059
x1_Construction ...	0.002	0.002	0.001	1.865	0.062
x1_Speciality & Ot...	0.002	0.002	0.001	1.860	0.063
x0_2018	0.001	0.001	0.001	1.857	0.063
rd_sale	0.003	0.003	0.002	1.789	0.074
x1_Transport	0.001	0.001	0.001	1.785	0.074
lt_debt	-0.002	-0.002	0.001	-1.772	0.076
cfm	-0.019	-0.019	0.011	-1.768	0.077
x2_VA	0.002	0.002	0.001	1.762	0.078
Length_term	-0.001	-0.001	0.000	-1.729	0.084
x2_HI	0.001	0.001	0.001	1.720	0.085
x1_Education	0.001	0.001	0.001	1.674	0.094
x1_other_sector	-0.001	-0.001	0.000	-1.646	0.100
x1_Software & Co...	0.002	0.002	0.001	1.644	0.100
x2_UNKOWN	0.001	0.001	0.001	1.602	0.109
x2_LA	0.001	0.001	0.001	1.593	0.111
x2_TX	0.003	0.003	0.002	1.582	0.114
x1_Insurance	0.001	0.001	0.001	1.569	0.117
x2_KY	0.001	0.001	0.001	1.566	0.117
STDEVTimeBrd	-0.002	-0.002	0.001	-1.543	0.123

x2_NY	0.002	0.002	0.002	1.528	0.127
x2_PA	0.002	0.002	0.001	1.470	0.142
x1_Household Pr...	0.001	0.001	0.001	1.425	0.154
x1_Business Ser...	0.001	0.001	0.001	1.420	0.156
x1_Containers & ...	0.001	0.001	0.001	1.414	0.157
x2_ME	0.001	0.001	0.000	1.400	0.162
TimeRole	-0.001	-0.001	0.001	-1.369	0.171
x2_DE	0.001	0.001	0.000	1.344	0.179
x1_Food Produce...	0.001	0.001	0.001	1.320	0.187
x1_Leisure & Hot...	0.001	0.001	0.001	1.307	0.191
x0_2012	-0.001	-0.001	0.001	-1.307	0.191
pe_op_dil	0.003	0.003	0.002	1.303	0.193
intcov_ratio	0.001	0.001	0.001	1.262	0.207
x1_Consumer Se...	0.001	0.001	0.001	1.259	0.208
x0_2020	0.001	0.001	0.000	1.256	0.209
x2_MD	0.001	0.001	0.001	1.251	0.211
AVG_SALPCT	0.002	0.002	0.002	1.237	0.216
x1_Automobiles ...	0.001	0.001	0.001	1.206	0.228
TotCurrNoLstdBrd	-0.001	-0.001	0.001	-1.190	0.234
TimeBrd	0.002	0.002	0.002	1.188	0.235
x2_OK	0.001	0.001	0.001	1.179	0.238
x2_DC	0.001	0.001	0.001	1.178	0.239
x2_UT	0.001	0.001	0.001	1.171	0.242
x2_NJ	0.001	0.001	0.001	1.164	0.245
x0_2014	0.001	0.001	0.001	1.124	0.261
pe_op_basic	-0.002	-0.002	0.002	-1.096	0.273
x1_Health	0.001	0.001	0.001	1.092	0.275
STD_SALPCT	-0.002	-0.002	0.002	-1.089	0.276
x1_Media & Entert...	0.001	0.001	0.001	1.069	0.285
x1_Pharmaceutic...	0.001	0.001	0.001	1.056	0.291

x2_FL	0.001	0.001	0.001	1.049	0.294
x1_Food & Drug ...	0.001	0.001	0.001	1.045	0.296
GProf	-0.001	-0.001	0.001	-1.044	0.296
TimeRetirement	0.001	0.001	0.001	1.032	0.302
x1_Mining	0.001	0.001	0.001	1.020	0.308
opmbd	0.051	0.051	0.050	1.009	0.313
x2_MO	0.001	0.001	0.001	0.996	0.319
roe	-0.000	-0.000	0.000	-0.981	0.327
cash_conversion	-0.000	-0.000	0.000	-0.976	0.329
x2_CT	0.001	0.001	0.001	0.973	0.330
x0_2015	0.001	0.001	0.001	0.972	0.331
STDEVTotNoLstd...	0.001	0.001	0.001	0.970	0.332
ocf_lct	0.001	0.001	0.001	0.958	0.338
pretret_noa	0.000	0.000	0.000	0.954	0.340
fcf_ocf	-0.000	-0.000	0.000	-0.942	0.346
opmad	-0.045	-0.045	0.052	-0.872	0.383
x2_OR	-0.001	-0.001	0.001	-0.850	0.396
dpr	0.000	0.000	0.000	0.847	0.397
equity_invcap	-0.003	-0.003	0.003	-0.842	0.400
rect_act	0.001	0.001	0.001	0.831	0.406
npm	0.039	0.039	0.049	0.800	0.424
inv_turn	-0.000	-0.000	0.000	-0.791	0.429
x2_MA	0.001	0.001	0.001	0.775	0.438
invt_act	0.001	0.001	0.001	0.767	0.443
STDEVAge	-0.001	-0.001	0.001	-0.757	0.449
rect_turn	-0.000	-0.000	0.000	-0.743	0.458
x2_AZ	0.001	0.001	0.001	0.740	0.459
STD_AGE	0.000	0.000	0.000	0.739	0.460
TotCurrNoUnLstd...	-0.000	-0.000	0.001	-0.731	0.465
x1_Steel & Other ...	0.000	0.000	0.001	0.723	0.469

adv_sale	0.000	0.000	0.000	0.715	0.475
x2_WI	0.001	0.001	0.001	0.713	0.476
aftret_eq	-0.004	-0.004	0.005	-0.712	0.476
aftret_equity	0.003	0.003	0.005	0.674	0.500
x2_RI	0.000	0.000	0.001	0.671	0.502
x2_VT	0.000	0.000	0.000	0.658	0.510
x1_Telecommuni...	0.000	0.000	0.001	0.642	0.521
x2_CO	0.001	0.001	0.001	0.637	0.524
debt_capital	0.000	0.000	0.000	0.617	0.537
pe_inc	-0.000	-0.000	0.001	-0.615	0.539
lt_ppent	-0.000	-0.000	0.000	-0.599	0.549
x2_SD	0.000	0.000	0.000	0.593	0.553
x2_GA	0.001	0.001	0.001	0.591	0.555
ptpm	-0.028	-0.028	0.047	-0.590	0.555
de_ratio	0.002	0.002	0.003	0.582	0.561

x1_Electricity	0.000	0.000	0.001	0.581	0.561
efftax	0.000	0.000	0.000	0.568	0.570
x2_NM	-0.000	-0.000	0.000	-0.565	0.572
sale_equity	-0.001	-0.001	0.003	-0.550	0.583
debt_at	0.001	0.001	0.001	0.539	0.590
intcov	-0.001	-0.001	0.001	-0.531	0.595
x2_IN	-0.000	-0.000	0.001	-0.518	0.605
x2_IL	0.001	0.001	0.001	0.516	0.606
pay_turn	0.000	0.000	0.000	0.516	0.606
debt_assets	-0.001	-0.001	0.001	-0.508	0.612
pretret_earnat	0.000	0.000	0.001	0.507	0.612
x2_ND	0.000	0.000	0.000	0.498	0.618
roce	0.000	0.000	0.001	0.486	0.627
x2_MI	-0.000	-0.000	0.001	-0.477	0.634
short_debt	-0.000	-0.000	0.000	-0.465	0.642

int_totdebt	0.000	0.000	0.000	0.461	0.644
x1_Publishing	0.000	0.000	0.001	0.453	0.650
roa	0.000	0.000	0.001	0.449	0.653
x2_MN	0.000	0.000	0.001	0.444	0.657
x2_AR	0.000	0.000	0.001	0.441	0.660
x2_WV	-0.000	-0.000	0.000	-0.440	0.660
x2_NH	0.000	0.000	0.000	0.438	0.661
x2_MT	0.000	0.000	0.000	0.432	0.665
x2_IA	0.000	0.000	0.001	0.410	0.682
x1_Renewable E...	-0.000	-0.000	0.000	-0.402	0.688
x2_ID	-0.000	-0.000	0.000	-0.396	0.692
x2_OH	0.000	0.000	0.001	0.381	0.703
curr_debt	-0.000	-0.000	0.001	-0.378	0.705
x1_Utilities - Other	0.000	0.000	0.001	0.376	0.707
TotNoUnLstdBrd	-0.000	-0.000	0.001	-0.361	0.718
x2_SC	0.000	0.000	0.001	0.314	0.754
x1_Chemicals	0.000	0.000	0.001	0.300	0.764
x2_TN	-0.000	-0.000	0.001	-0.285	0.776
x2_BC	-0.000	-0.000	0.000	-0.281	0.779
ptb	0.000	0.000	0.000	0.280	0.779
dltt_be	-0.000	-0.000	0.000	-0.277	0.782
CAPEI	-0.000	-0.000	0.000	-0.276	0.782
x2_PR	0.000	0.000	0.000	0.275	0.783
at_turn	-0.000	-0.000	0.001	-0.271	0.786
x2_QC	0.000	0.000	0.000	0.269	0.788
aftret_invcapx	0.000	0.000	0.001	0.268	0.789
x1_Beverages	0.000	0.000	0.001	0.254	0.799
Succession	0.000	0.000	0.001	0.247	0.805
x1_Oil & Gas	0.000	0.000	0.001	0.247	0.805
cash_ratio	0.001	0.001	0.002	0.241	0.810

x2_NE	-0.000	-0.000	0.001	-0.238	0.812
TotNoLstdBrd	0.000	0.000	0.001	0.226	0.821
x1_Electronic & El...	-0.000	-0.000	0.001	-0.221	0.825
x2_MS	0.000	0.000	0.001	0.220	0.826
pe_exi	-0.000	-0.000	0.001	-0.212	0.832
STDEVNoQuals	0.000	0.000	0.000	0.208	0.835
bm	-0.000	-0.000	0.001	-0.205	0.837
x2_NV	0.000	0.000	0.001	0.194	0.846
x2_WA	0.000	0.000	0.001	0.172	0.864
pcf	0.000	0.000	0.000	0.168	0.867
cash_lt	0.000	0.000	0.001	0.162	0.872
x1_Tobacco	-0.000	-0.000	0.000	-0.135	0.893
x2_KS	0.000	0.000	0.001	0.134	0.893
int_debt	0.000	0.000	0.000	0.130	0.897
x1_Leisure Goods	0.000	0.000	0.000	0.129	0.897

x1_Diversified Ind...	-0.000	-0.000	0.001	-0.121	0.903
x2_CA	-0.000	-0.000	0.002	-0.111	0.912
sale_nwc	0.000	0.000	0.000	0.109	0.913
NationalityMix	0.000	0.000	0.000	0.099	0.921
x0_2013	-0.000	-0.000	0.001	-0.094	0.925
quick_ratio	0.000	0.000	0.004	0.088	0.930
profit_lct	-0.000	-0.000	0.001	-0.084	0.933
AVG_TOTAL_CO...	-0.000	-0.000	0.000	-0.079	0.937
debt_invcap	-0.000	-0.000	0.003	-0.070	0.944
x1_Real Estate	-0.000	-0.000	0.001	-0.067	0.947
NoQuals	-0.000	-0.000	0.001	-0.066	0.947
x2_WY	-0.000	-0.000	0.000	-0.059	0.953
sale_invcap	-0.000	-0.000	0.001	-0.047	0.962
accrual	-0.000	-0.000	0.001	-0.039	0.969
gpm	-0.000	-0.000	0.005	-0.024	0.981

x2_ON	0.000	0.000	0.000	0.017	0.987
evm	-0.000	-0.000	0.000	-0.013	0.989
x1_Information Te...	-0.000	-0.000	0.001	-0.007	0.994
totdebt_invcap	0.000	0.000	0.001	0.006	0.995
curr_ratio	-0.000	-0.000	0.003	-0.001	0.999
base_prediction0	0	0	?	?	?

Appendix D: Baseline Model - Linear Regression using only Last Year's Board Percentage

Generalized Linear Model

```
Model Metrics Type: RegressionGLM
 Description: N/A
 model id: rm-h2o-model-generalized_linear_model-216
 frame id: rm-h2o-frame-generalized_linear_model-216
 MSE: 0.00511848
 RMSE: 0.07154355
 R^2: 0.71044075
 mean residual deviance: 0.00511848
 mean absolute error: 0.043657135
 root mean squared log error: 0.061402056
 null DOF: 16724.0
 residual DOF: 16723.0
 null deviance: 295.6444
 residual deviance: 85.606575
 AIC: -40753.17
GLM Model (summary):
   Family     Link Regularization Number of Predictors Total Number of Active Predictors Number of Iterations
 gaussian identity           None                     1                               1                    1
Scoring History:
            timestamp   duration iterations negative_log_likelihood objective
 2021-04-12 03:10:19  0.000 sec          0               295.64438   0.01768

H2O version: 3.30.0.1-rm9.8.1
```

Appendix E: Descriptions of Different Types of Attempted Modes

Model 1: Ordinary Least Squares (OLS) Regression

The first model we explored was an Ordinary Least Squares (OLS) Regression. We first tested this model as a baseline to compare its performance to more complex models built hereafter. We used nearly all features we had available, except for *FEMALE_COUNT* and *EXEC_COUNT*, as these variables were used to create the target variable and would most likely introduce multicollinearity in our models.

The first main advantage of using OLS regression is that it simplifies the relationship between multiple predictor variables and the target variable. Another advantage of using OLS regression comes from its computational efficiency. Given the simplistic nature of the model, OLS does not require complicated calculations and can run promptly predictions even when the large amount of data is entered into the model. Lastly, ease of interpretability of the output is the reason we have decided to start our analysis with OLS regression. The ability of OLS regression to determine the relative influence of one or more predictor variables to the predicted value using the coefficient can comprehensively express which changes in the predictor variable cause which changes in the target variable.

On the other hand, the limitation of using OLS regression is that it is too simplistic. Even when the model uses a complex and vast amount of features, simplistic linear representation of the model output can distort the underlying relationship between predictors and target variable. Additionally, OLS regression is severely influenced by the presence of outliers. The orientation and direction of the best fit line can be altered by even one instance of outliers, resulting in a model that does not capture the information. It is also worthwhile to mention that OLS regression assumes that the independent variables are not correlated with the target variable. Thus, the

output of the model is not accurate unless the user did a thorough job to remove multicollinearity of the variables beforehand. Lastly, OLS regression cannot determine the feature importance given the nature of OLS regression lying on its 'independence' assumption.

We have used two OLS regressions: one for statistical analysis and one for machine learning. OLS regression in which predictors directly fit the target variable returned an R-squared value of 0.722, indicating that more than 70 percent of the variance in the target variable is explained by the independent variables we entered into the model. F-statistic compares our model with zero predictor variables (the intercept only model) and decides whether added coefficients improved the model. Given the value, we can treat this model to have rejected the null hypothesis and thus be statistically significant. However, when OLS regression is computed after splitting the dataset into training and testing set to validate the model performance, R-squared value went down to -0.499, indicating that chosen model does not follow the trend of the data, so fits worse than a horizontal line. We have decided to explore other regression models given such a discrepancy. The next two models we considered are Ridge and Lasso regression. These models reduce model complexity and prevent overfitting, which may have been inferred by the R-squared value (0.840) from simple linear regression.

Model 2: Ridge Regression

The second model we examined, using the same set of features, was a ridge regression model. In ridge regression, the cost function is modified by adding a penalty term corresponding to the square of the magnitude of the coefficients. So the advantage of the ridge regression is that it shrinks the coefficients, reduces the model complexity and thus multicollinearity. This model was a continuation of our first, as now we are using a L2 regularization parameter that penalizes large weights in the model. The single parameter that is entered as the argument of the Ridge

regression is *"alpha"*, which controls the magnitude of regularization strength. The higher this value is, the stronger the regularization is and the lower variance the model has. We have set our *"alpha"* value as 1, and the following table is the performance measure of the ridge regression. The limitations of Ridge regression is that it trades the variance for bias by shrinking the coefficients towards zero. Additionally, ridge regression cannot perform feature selection. Thus, we also have explored Lasso regression, which is discussed below.

Model 3: Lasso Regression

The main difference Lasso Regression from the Ridge regression is that it uses different regularization (L1). In L1 regularization, absolute value of the coefficient is considered instead of taking the square of the coefficients. This leads to zero coefficients; thus, Lasso regression not only helps in reducing overfitting but it can help us in feature selection.

Model 4: Elastic Net Regression

Elastic net regression can be thought of as the combination of Ridge and Lasso regression. The argument of elastic regression accepts two parameters. In the parameter tuning, alpha parameter stands for the sum of lambda1 and lambda2. Thus, Elastic Net can be thought of as an extension of linear regression that adds regularization penalties to the loss function.

Model 5: Decision Tree Regression

The decision tree algorithm is characterized by recursive partitioning — Starting from the root, the data is split on the feature that results in the largest Information Gain. In an iterative process, we then repeat this splitting procedure at each child node until the leaves are pure, or in another word, until all samples at each node belong to the same class. Decision trees regression normally uses mean squared error (MSE) as a metric to determine the point at which split needs to be made. Then, a prediction is made by running the data points through the entire tree until it

reaches a leaf node. The final prediction is the average of the value of the dependent variable in that leaf node. Disadvantages of using decision tree and other tree based models is that they are generally limited to extrapolate to unseen data, especially future time periods since it is just taking the average value out of the observed points. In a problem where any tree based algorithm is rendered useless, neural net or linear regression models are deemed to be more preferred models given that we want to use a model that can fit the data and consequently extrapolate with a higher level of accuracy.

(from:https://gdcoder.com/decision-tree-regressor-explained-in-depth/)

Model 6: Gradient Boosted Tree Regression (Ensemble)

Gradient boosted tree works by first calculating the average of the target label. This leaf will be used as a baseline to approach the correct solution in the proceeding steps. Next, the residuals are calculated for every sample, and a decision tree is constructed with a goal of predicting the residuals. In the next step, target label is predicted using all of the trees within the ensemble after which the new residuals are computed. This process is repeated several times until the number of interactions matches the number specified by the hyperparameter. Finally, all of the trees in the ensemble are used to compute a final prediction.

(from:https://towardsdatascience.com/machine-learning-part-18-boosting-algorithms-gradient-boosting-in-python-ef5ae6965be4)

Model 7: AdaBoost Regression (Ensemble)

An AdaBoost regressor is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction. As a general principle Adaboost builds an ensemble by sequentially adding members which have been trained

on those instances of data which are proving most difficult to correctly predict. Each new predictor is given a training set where the difficult examples are increasingly represented, this is achieved either through weighting or resampling.

(From:https://datascience.stackexchange.com/questions/6949/how-will-ada-boost-be-used-for-solving-regression-problems)

Model 8: Bagging Regression (Ensemble)

Bagging method often considers homogeneous weak learners, learns them independently from each other in parallel and combines them following some kind of deterministic averaging process. The idea of bagging is then simple: we want to fit several independent models and "average" their predictions in order to obtain a model with a lower variance. However, we can't, in practice, fit fully independent models because it would require too much data. So, we rely on the good "approximate properties" of bootstrap samples (representativity and independence) to fit models that are almost independent.

(From:https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205)

Model 9: Extra Trees Regression (Ensemble)

Extra Trees is an ensemble machine learning algorithm that combines the predictions from many decision trees. Thus, it can be regarded as the extension of the decision tree algorithm, and it is also related to the widely used random forest algorithm. The main advantage of extra tree regression is that although it uses a simpler algorithm it can often achieve equivalent, if not, better performance compared to the random forest algorithm.

(From:https://machinelearningmastery.com/extra-trees-ensemble-with-python/#:~:text=Extra%20Trees%20is%20an%20ensemble,predictions%20from%20many%20decision%20trees.&text=Extra%20Trees%20ensemble%20is%20an,and%20regression%20with%20scikit%2Dlearn.)

Model 10: Random Forest Regression (Ensemble)

Random forests Learning trees are very popular base models for ensemble methods. Strong learners composed of multiple trees can be called "forests". The random forest approach is considered as a bagging method where deep trees, fitted on bootstrap samples, are combined to produce an output with lower variance. Random forest algorithm combines the concepts of bagging and random feature subspace selection to create more robust models.

(From:https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205)

Model 11: Stacking Regression (Ensemble)

Often considered as one of the heterogeneous weak learners, stacking ensembles learn and combine the training set by computing a prediction based on the different weak models predictions. The main difference of the stacking from bagging and boosting is characterized by stacking's heterogeneity, meaning different learning algorithms are combined using a meta-model; whereas, both bagging and boosting combine homogeneous weak learners using more deterministic algorithms. Second, stacking learns to combine the base models using a meta-model whereas bagging and boosting combine weak learners following deterministic algorithms.

Model 12: Histogram Gradient Boosting Regression (Ensemble)

As discussed, gradient boosting is an ensemble of decision trees algorithms known for one of the most popular techniques for structured classification and predictive modeling

problems. A major disadvantage of tree-based methods is that it is slow to train the model, which can be problematic when training large datasets. However, the training process via histogram-based gradient boosting ensembles can be dramatically sped up by discretizing and transforming the continuous features into a few hundred unique categorical values.

Appendix E: Individual Contributions

- **Emmy:** Data preparation (merging and cleaning data sources), modeling, writeup, presentation
- **Qiyu:** Data engineering, modeling, writeup, presentation
- **Sean:** Data sourcing, modeling (running models and hyperparameter tuning in Python), writeup
- **Stephen:** Preliminary modeling, writeup, presentation