Milestone 3 – Data Project

Cybersecurity Analytics

November 19, 2024

Deliverable Requirements:

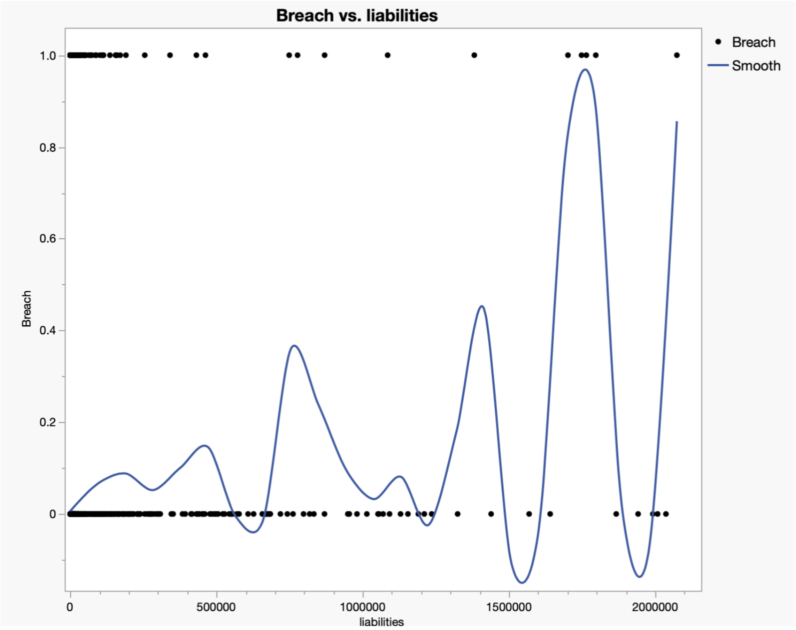
A description of and results from at least two advanced analysis methods (e.g. linear regression, logistic regression), which variable sets were used, and an assessment of the performance of these analyses. The group should build up the analysis slowly by first estimating simple models with few predictors and sequentially adding more variables. Identification of key variables that predict breach and how they match (or don't match) your initial expectation will be key to receiving full credit. You can also assess goodness of fit of these models and how much of the variation in data breaches they explain. Also, each group will be required to give a 15-minute presentation documenting their entire process through all three milestones as well as the result of their prediction effort.

Variables of Importance:

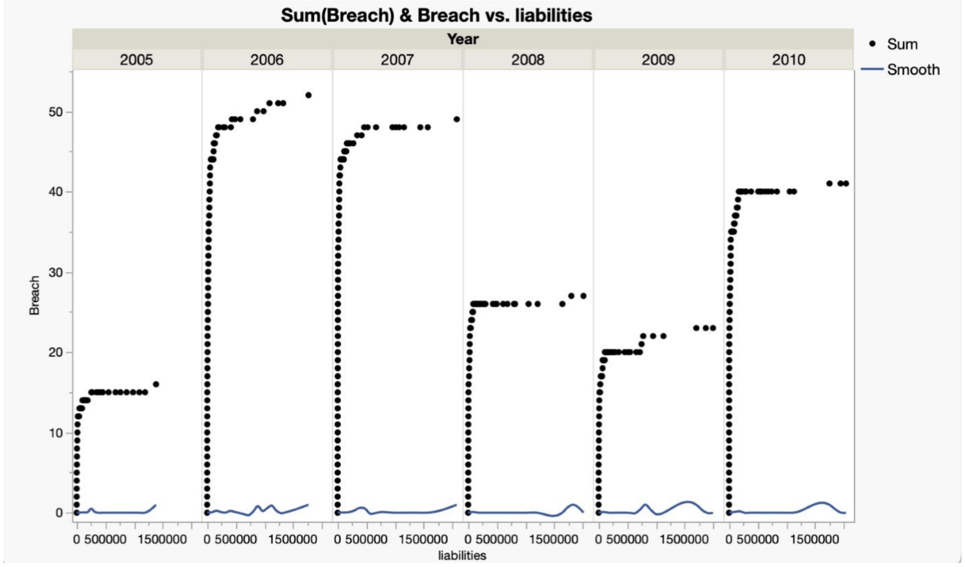
1. StockHolderEquity
2. Liabilities

Liabilities Analysis

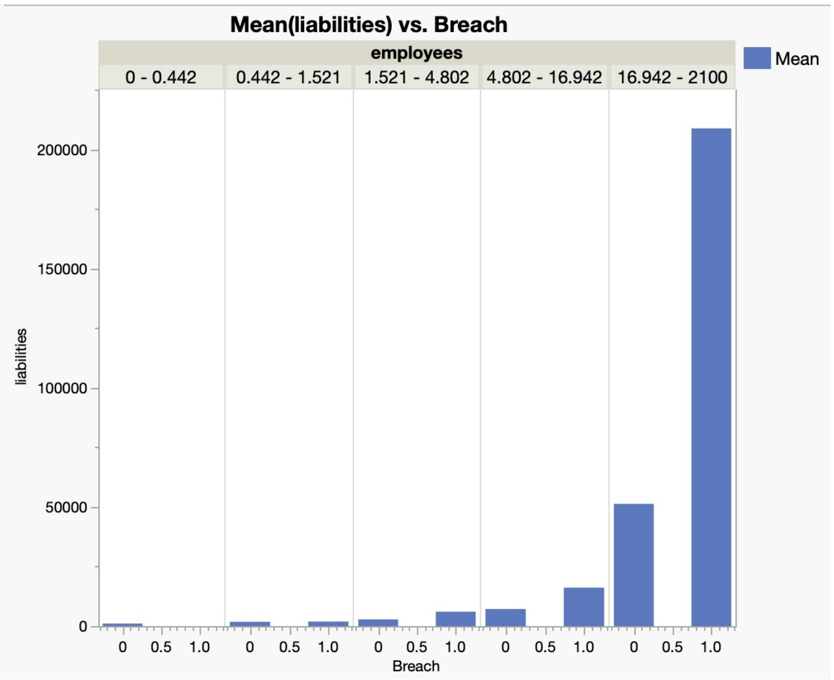
From Milestone 2:



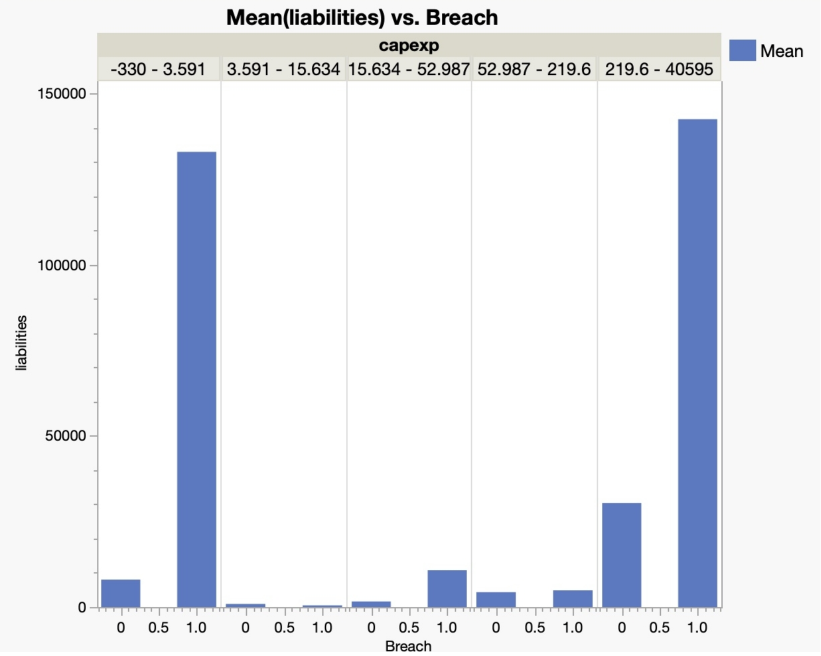
The two variables we are focused on in this graph are breach and liabilities. The variable "liabilities" is the accounting measure of the total obligations that a company owes. The variable breach is a binary variable that describes whether a company has been breached or not, 1 = breached and 0 = not breached. Our correlation matrix shows that there is a positive correlation between the two variables with a value of 0.1272. From the graph above you cannot really tell too much so we decided to separate the number of breaches by year for the liability amounts and got a graph that is much more descriptive.



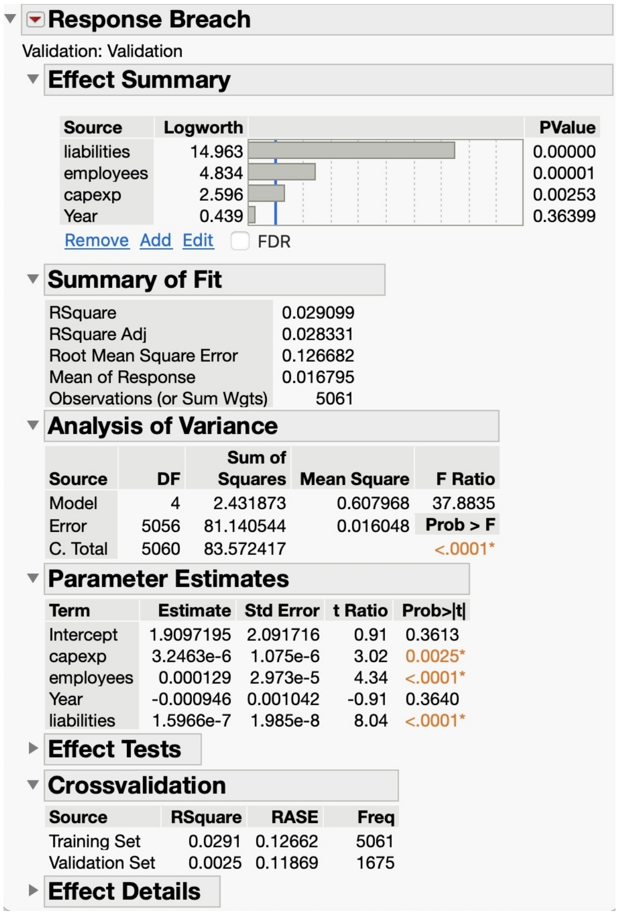
This graph shows us a lot more about the correlation between the two variables. You can see that the number of breaches that occur within a year increases as the amount of liabilities increases. In 2006 the number of breaches increases to a maximum of 53 for the company with the highest liability. In 2005, 2008, 2009, and 2010 you can see that the logarithmic curve for the number of breaches levels our much more than those in 2006 and 2007 and this could be because of the invention of cloud computing in 2006.



Now continuing this analysis, we first started by expanding on the simpler analyses from milestone 2. We started by changing how we split up the graph for breach v liabilities by now splitting the liabilities values by the number of employees each company has and then looking at the mean number of breaches the companies have. As you can see from this graph above, as the number of employees in each company increases, so does the number of breaches that the company has. Now knowing that there is a positive correlation between employees, liabilities, and breach we can expand on this analysis.



Now, we are continuing the correlation between the liabilities v breach, but this time split up the analysis by the values of capital expenditures. Here you can see more of an interesting split in the data. The distribution here is bimodal, meaning there are 2 distinct peaks in the data. When the capital expenditures are low and the liabilities are high, the number of breaches also increases. When the capital expenditures are high and so are the liabilities, the number of breaches is also increased.



Response Breach

Validation: Validation

Effect Summary

Source	Logworth	PValue
liabilities	14.963	0.00000
employees	4.834	0.00001
capexp	2.596	0.00253
Year	0.439	0.36399

Summary of Fit

RSquare	0.029099
RSquare Adj	0.028331
Root Mean Square Error	0.126682
Mean of Response	0.016795
Observations (or Sum Wgts)	5061

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	2.431873	0.607968	37.8835
Error	5056	81.140544	0.016048	Prob > F
C. Total	5060	83.572417		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1.9097195	2.091716	0.91	0.3613
capexp	3.2463e-6	1.075e-6	3.02	0.0025*
employees	0.000129	2.973e-5	4.34	<.0001*
Year	-0.000946	0.001042	-0.91	0.3640
liabilities	1.5966e-7	1.985e-8	8.04	<.0001*

Effect Tests

Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.0291	0.12662	5061
Validation Set	0.0025	0.11869	1675

Effect Details

Here we have continued with a more in-depth analysis of liabilities v breach now including the variables we looked at above: capital expenditures, employees, and year. We first establish a significance level of α = 0.05. Below shows the null and alternative hypotheses for the data.

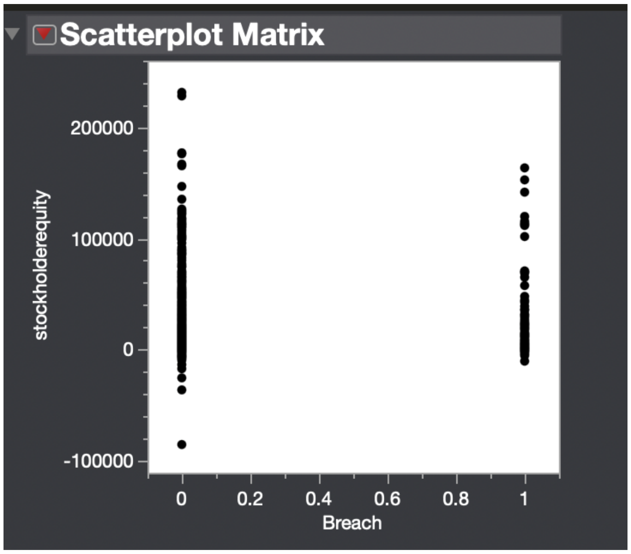
H_0: The predictor variables have NO significant impact on the probability of a company getting breached.

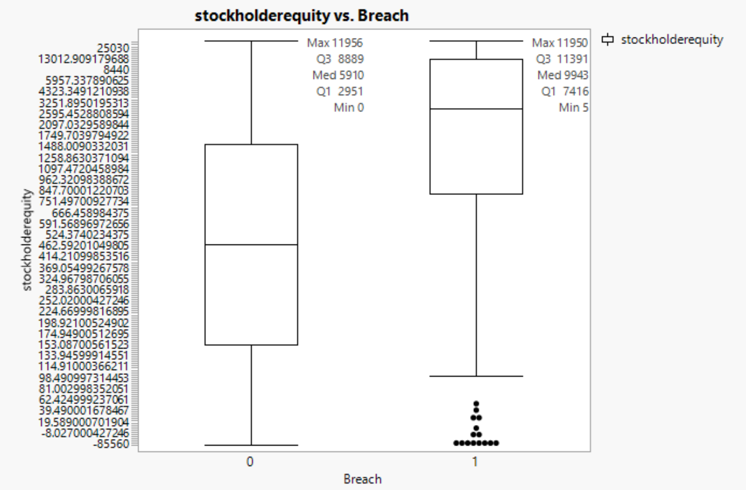
H_A: The predictor variables have a significant impact on the probability of a company getting breached.

For the parameters we can see that most of them have a statistically significant impact on the probability of a company getting breached. The p-value for capital expenditures = 0.0025, the p-value for employees = <0.001, and the p-value for liabilities = <0.001 which are all statistically significant in this linear regression model. The p-value for a year = 0.3640 on the other hand is not statistically significant on its own but put with the other variables we get a different conclusion with a p-value of <0.001.

Conclusion: At a significance level of α = 0.05 and with a p-value of <0.001 we reject the null hypothesis as we have significant evidence to suggest that the predictor variables have a significant impact on the probability of a company getting breached.

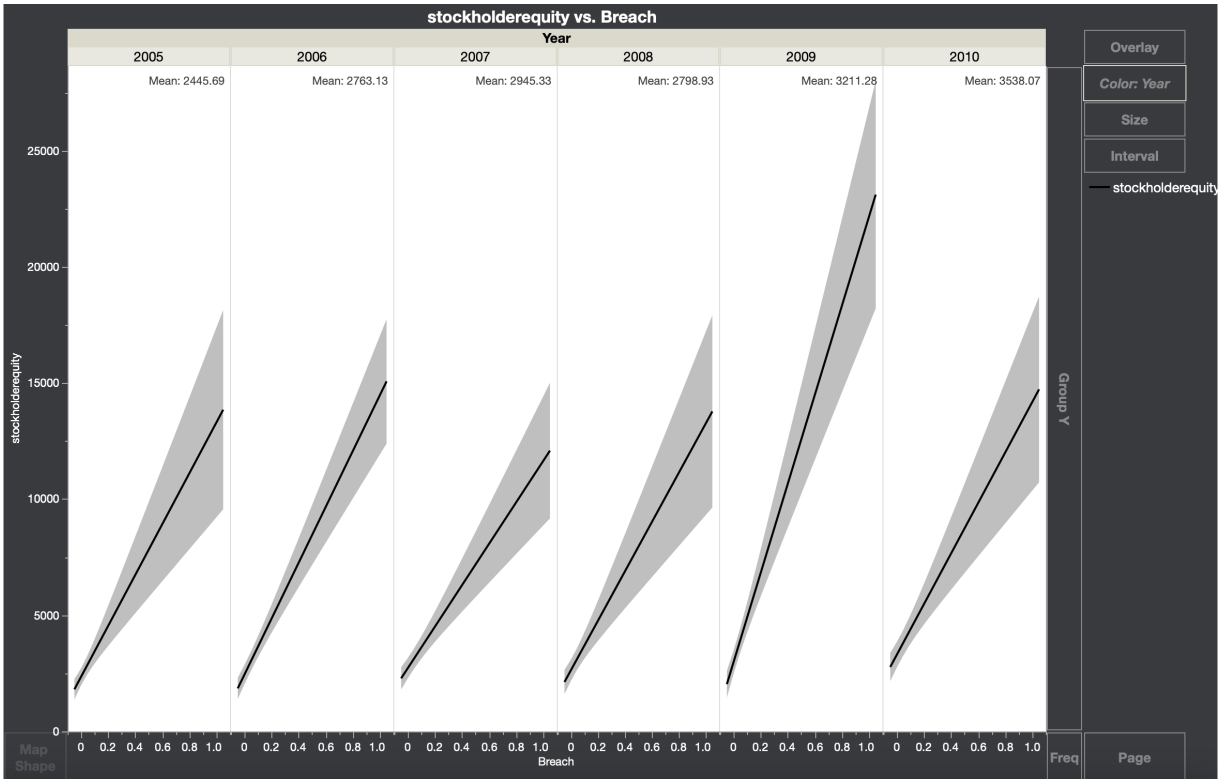
StockHolderEquity Analysis



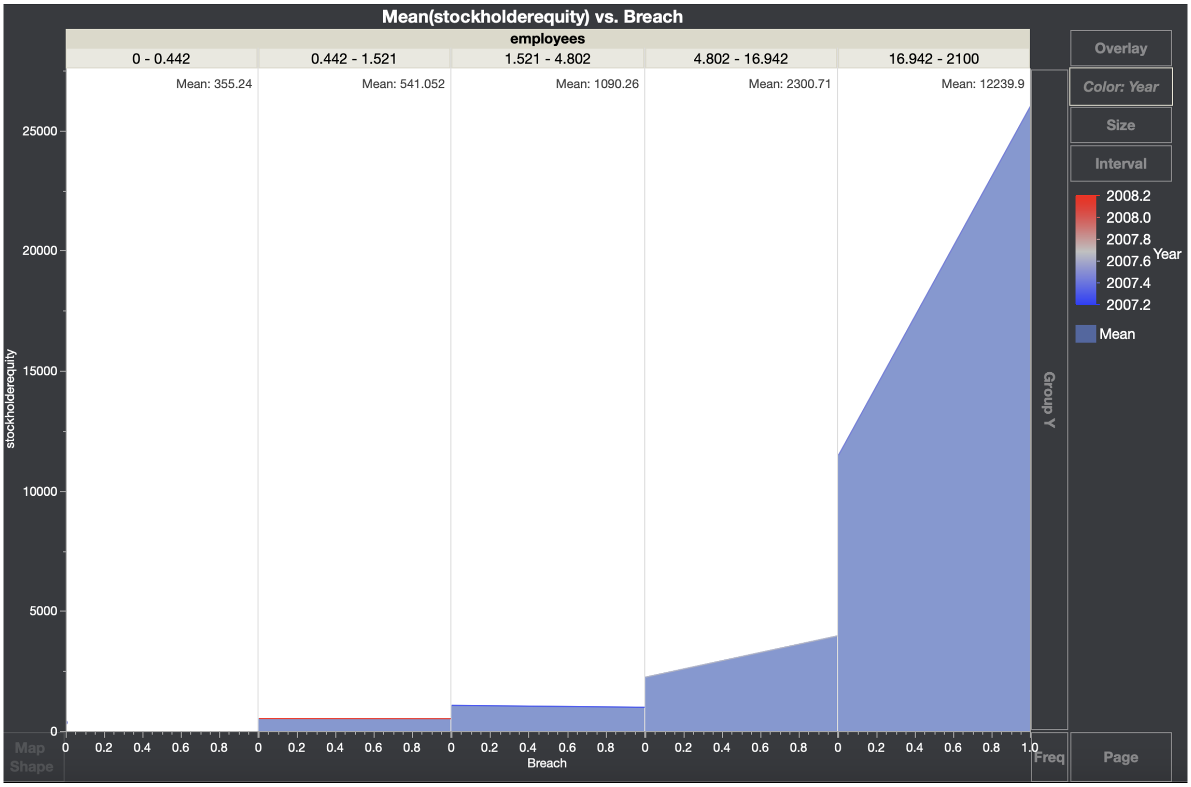


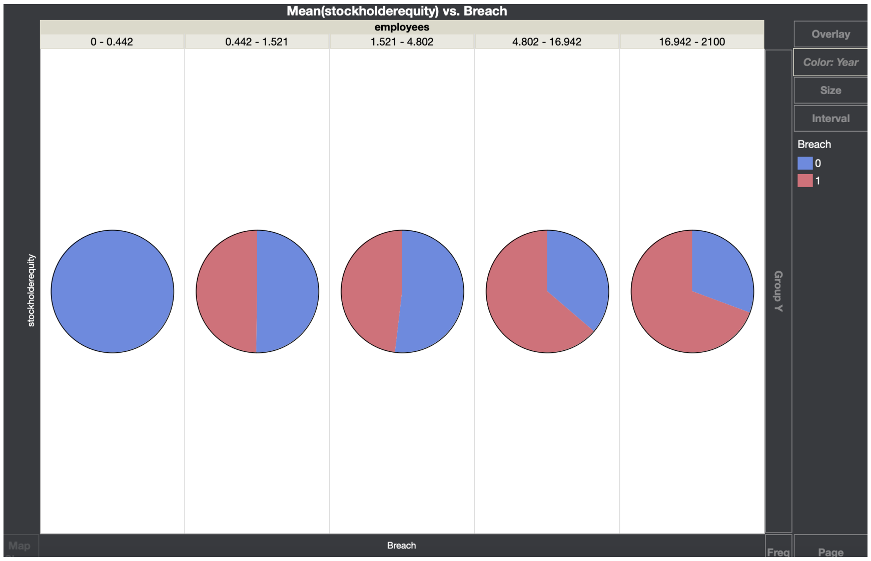
Positive Correlation: 0.1435

Based on the Correlation Matrix, we found a moderate, positive relationship of 0.1435 between the stockholder equity and breach variables. This implies that there's a tendency for breach occurrences to rise in parallel with stockholder equity. Using a Scatterplot Matrix, a linear relationship is displayed between the variables. In addition, through the box plot we can gather that non-breached companies tend to have a lower stockholder equity average of 5910, while breached companies have a higher stockholder equity average of 9943. It's also important to note the present outliers in breached companies. With the min being 5, the small values in Y are displayed as outliers likely due to the companies(s) resources, size or structure.



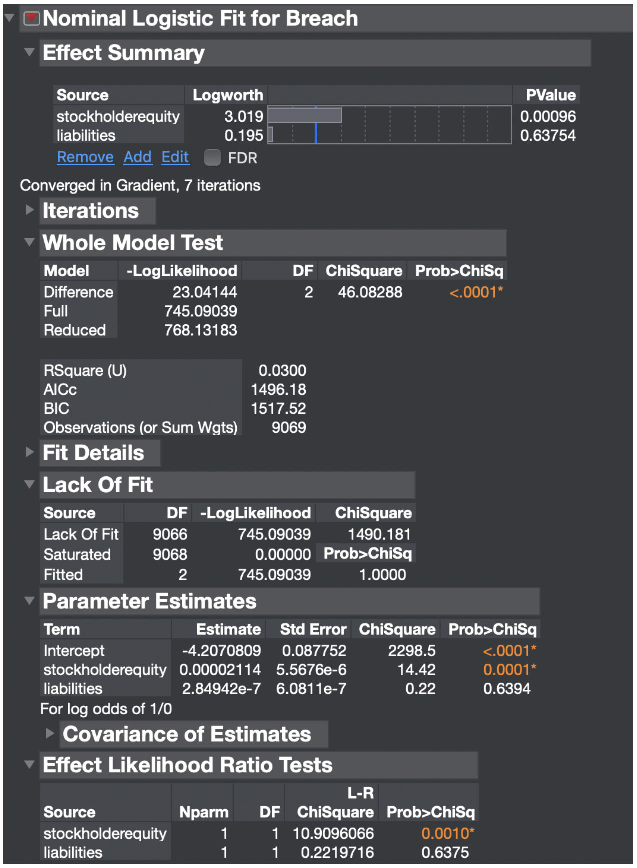
Once we incorporate the Year variable, we can confidently suggest that the mean of stockholder equity steadily increases over time, from 2005 to 2010. Starting with a mean of 2445.69 in 2015 and growing gradually until we reach the mean of 3538.07 in 2010 indicates that companies have been growing financially over time. This may also suggest that the financial performance across all companies during those 5 years expanded, which can influence how they invest in counter breach efforts or their likelihood to become a target.





Once we incorporate the Number of Employees variable into, you can see the mean almost doubles per employee range. As the range increases, the stockholder equity mean grows significantly. This can suggest that larger companies, based on employee numbers and stockholder equity, are more financially secure. Furthermore, they might invest more in resources and solutions, while also being an attractive target due to their high exposure.

Nominal Logistic Fit Model for Breach



Nominal Logistic Fit for Breach

Effect Summary

Source	Logworth	PValue
stockholderequity	3.019	0.00096
liabilities	0.195	0.63754

Remove Add Edit ☐ FDR

Converged in Gradient, 7 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	23.04144	2	46.08288	<.0001*
Full	745.09039			
Reduced	768.13183			

RSquare (U)	0.0300
AICc	1496.18
BIC	1517.52
Observations (or Sum Wgts)	9069

Fit Details

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	9066	745.09039	1490.181
Saturated	9068	0.00000	Prob>ChiSq
Fitted	2	745.09039	1.0000

Parameter Estimates

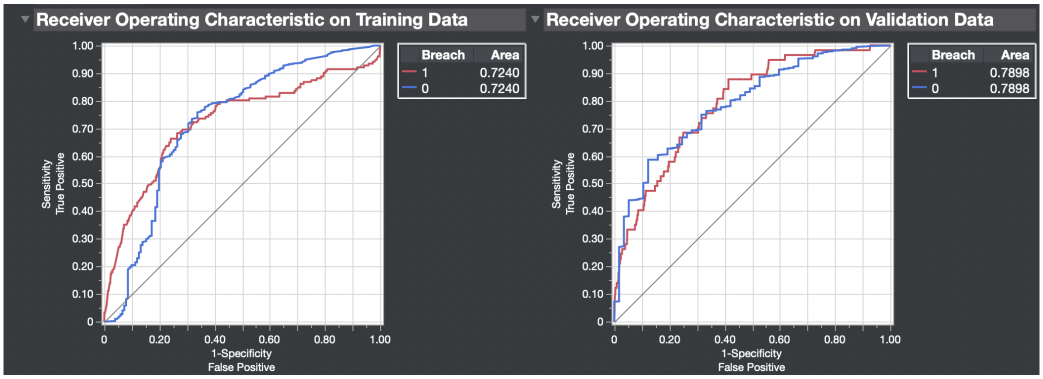
Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-4.2070809	0.087752	2298.5	<.0001*
stockholderequity	0.00002114	5.5676e-6	14.42	0.0001*
liabilities	2.84942e-7	6.0811e-7	0.22	0.6394

For log odds of 1/0

Covariance of Estimates

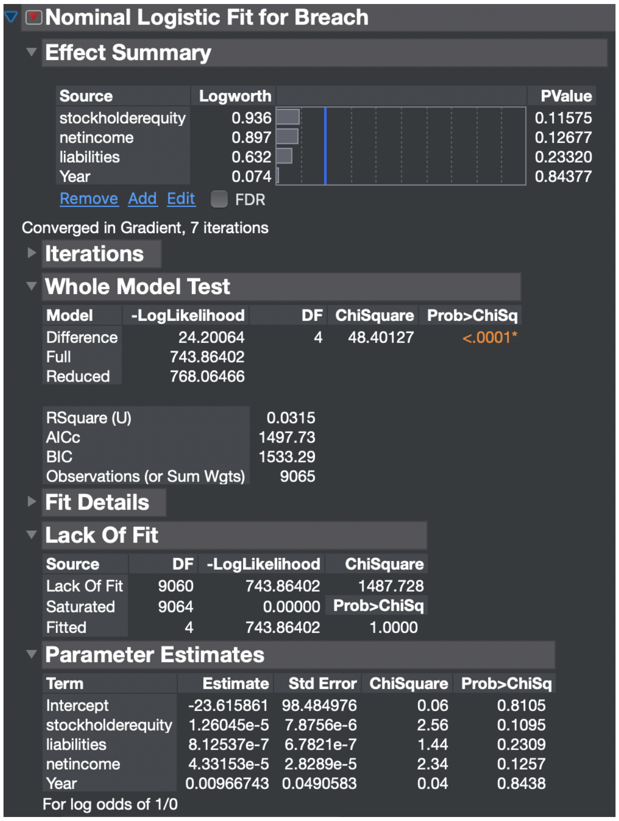
Effect Likelihood Ratio Tests

			L-R	
Source	Nparm	DF	ChiSquare	Prob>ChiSq
stockholderequity	1	1	10.9096066	0.0010*
liabilities	1	1	0.2219716	0.6375



- *Null hypothesis (H0)* – "The predictor has NO significant impact on the probability of company getting breached"
- *Alternative hypothesis (H1)* – "The predictor has a significant impact on the probability of a company being breached"

This is the Nominal Logistic Fit along with the ROC curve for breaches compared against the base variables selected – stockholderequity and liabilities. Paired into this is also the validation column produced from our data sheet. According to the p-values liabilities out of the two would be the stronger predictor. When looking at the RSquare we see these variables only making up 3% of the variation in company breaches; therefore, it is not a good predictor



Nominal Logistic Fit for Breach

Effect Summary

Source	Logworth	PValue
stockholderequity	0.936	0.11575
netincome	0.897	0.12677
liabilities	0.632	0.23320
Year	0.074	0.84377

Remove Add Edit □ FDR

Converged in Gradient, 7 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	24.20064	4	48.40127	<.0001*
Full	743.86402			
Reduced	768.06466			

RSquare (U)	0.0315
AICc	1497.73
BIC	1533.29
Observations (or Sum Wgts)	9065

Fit Details

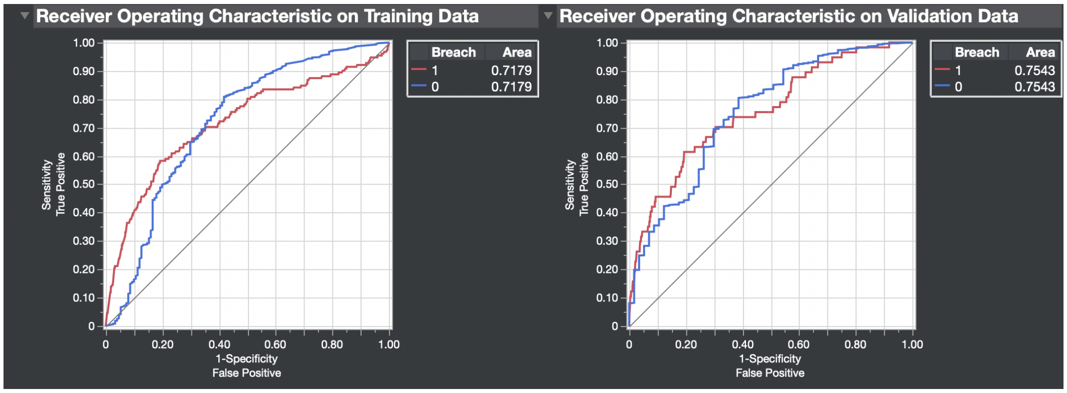
Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	9060	743.86402	1487.728
Saturated	9064	0.00000	Prob>ChiSq
Fitted	4	743.86402	1.0000

Parameter Estimates

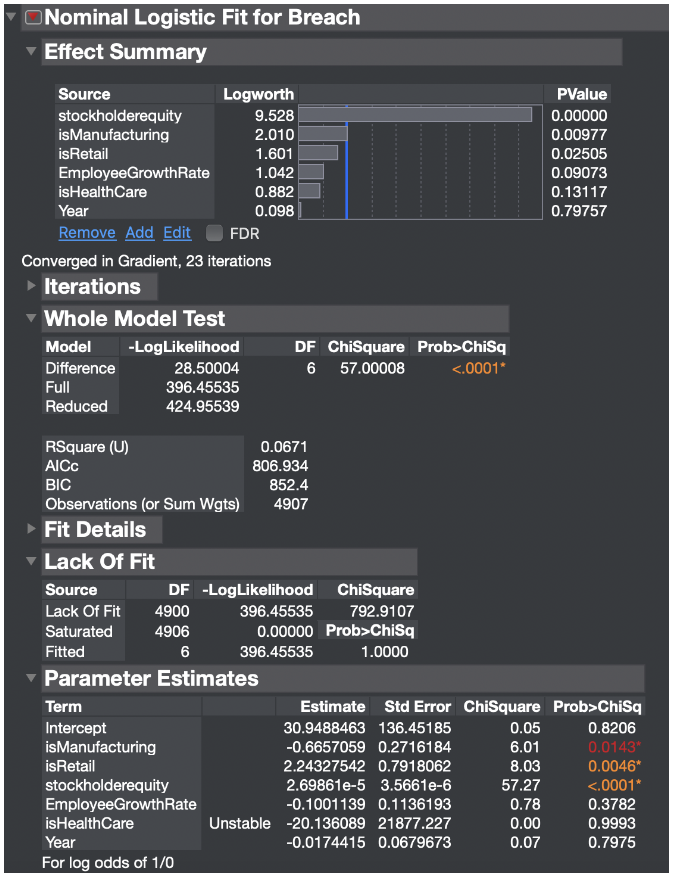
Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-23.615861	98.484976	0.06	0.8105
stockholderequity	1.26045e-5	7.8756e-6	2.56	0.1095
liabilities	8.12537e-7	6.7821e-7	1.44	0.2309
netincome	4.33153e-5	2.8289e-5	2.34	0.1257
Year	0.00966743	0.0490583	0.04	0.8438

For log odds of 1/0



- *Null hypothesis (H0)* – "The predictor has NO significant impact on the probability of company getting breached"
- *Alternative hypothesis (H1)* – "The predictor has a significant impact on the probability of a company being breached"

This is the Nominal Logistic Fit Model for breach compared against stockholder equity, net income, liabilities, and year. This is also paired with a validation column produced. We see the two higher p-values of liabilities and year being the stronger variables out of the 4 but still outputting a weak RSquare of 3.2% only a slight increase from just stockholder equity and liabilities.



Nominal Logistic Fit for Breach

Effect Summary

Source	Logworth	PValue
stockholderequity	9.528	0.00000
isManufacturing	2.010	0.00977
isRetail	1.601	0.02505
EmployeeGrowthRate	1.042	0.09073
isHealthCare	0.882	0.13117
Year	0.098	0.79757

Remove Add Edit □ FDR

Converged in Gradient, 23 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	28.50004	6	57.00008	<.0001*
Full	396.45535			
Reduced	424.95539			

RSquare (U)	0.0671
AICc	806.934
BIC	852.4
Observations (or Sum Wgts)	4907

Fit Details

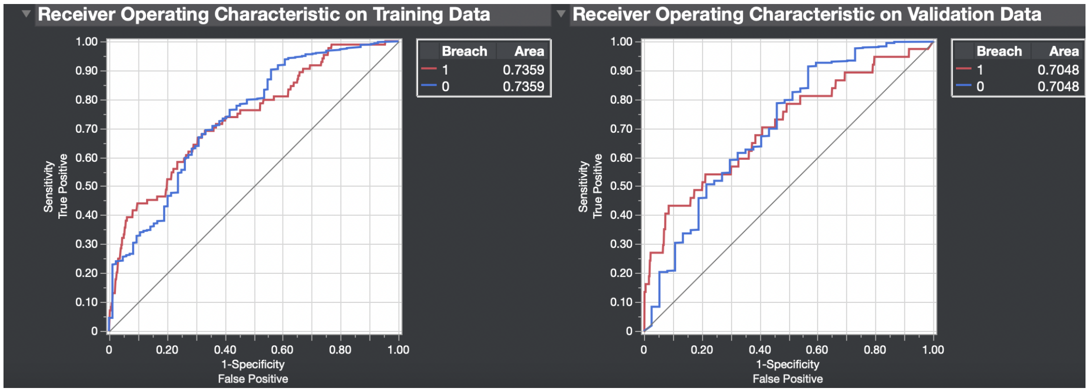
Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	4900	396.45535	792.9107
Saturated	4906	0.00000	Prob>ChiSq
Fitted	6	396.45535	1.0000

Parameter Estimates

Term		Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept		30.9488463	136.45185	0.05	0.8206
isManufacturing		-0.6657059	0.2716184	6.01	0.0143*
isRetail		2.24327542	0.7918062	8.03	0.0046*
stockholderequity		2.69861e-5	3.5661e-6	57.27	<.0001*
EmployeeGrowthRate		-0.1001139	0.1136193	0.78	0.3782
isHealthCare	Unstable	-20.136089	21877.227	0.00	0.9993
Year		-0.0174415	0.0679673	0.07	0.7975

For log odds of 1/0



Breach	Area
1	0.7359
0	0.7359

Breach	Area
1	0.7048
0	0.7048

In this model we compared stockholder equity, is Manufacturing, is Retail, Employee Growth Rate, is Healthcare, and year based on previous models that were showing a higher

Logworth than others. This led the RSquare to two-fold to .0671 compared to the previous models .0315 with only two extra variables added to this current one. You can also tell the ROC charts lines overlap a bit more compared to others, this indicates that both models may have very similar performance in terms of their ability to discriminate between positive and negative classes. The lines are not as close to the upper left as other models meaning it doesn’t perform as well.