STAT 510 Fall 2023
Final Report
Teah Thies & Lakely Nealis

Bankruptcy Factors

1. Motivation

No business owner starts a business thinking that they will have to file for bankruptcy one day. Although, doing so helps the one who owes money as well as helps the person that is owed money. For the person owing money, the stress is alleviated by debt discharge, repayment plans, and asset protection. For the person who is owed money, they receive fair distribution, fair treatment, and debtor rehabilitation so that the creditor is returned a fair amount of money within time.

A statistical article claims that “the main causes of bankruptcy remain constant over the years: Job loss and medical expenses”" [Bankruptcy]. We will be looking further into the factors that drive a company or business to bankruptcy. We aim to collect knowledge that empowers aspiring entrepreneurs, business owners, and financial professionals to implement proactive measures and minimize risk to create a steady account. On the flip side, those also can prepare if the day does come where one has to file: they can face it with resilience and know how to transform adversity into opportunity. We will look at actual data collected from the “Taiwan Economic Journal for the years 1999 to 2009. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange” [Fedesoriano]. By analyzing these factors that lead to financial distress, we gain insight into the forces that disrupt stability.

2. Data Summary

We will look at actual data collected from the “Taiwan Economic Journal for the years 1999 to 2009. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange” [Fedesoriano].

<u>Output Feature:</u>
Y - Bankrupt? [Class label (1=Yes, 0=No); Binary]

<u>Input Features:</u>

- RoaC - before interest and depreciation before interest: Return On Total Assets(C)
- RoaA - before interest and % after tax: Return On Total Assets(A)
- RoaB - before interest and depreciation after tax: Return On Total Assets(B)
- GrossMargin - Operating Gross Margin: Gross Profit/Net Sales
- RealizedGrossMargin - Realized Sales Gross Margin: Realized Gross Profit/Net Sales
- ProfitRate - Operating Profit Rate: Operating Income/Net Sales

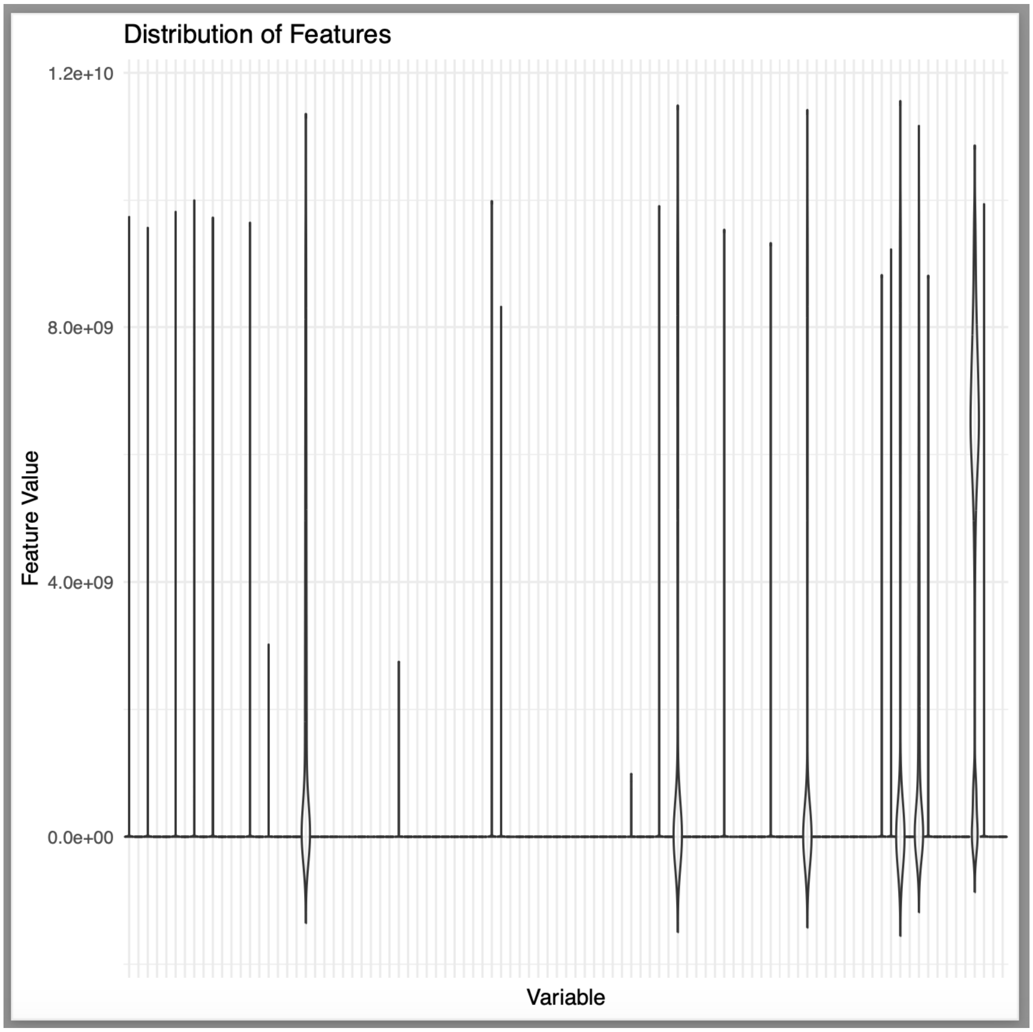
- PreTaxInterestRate - Pre-tax net Interest Rate: Pre-Tax Income/Net Sales
- AfterTaxtInterestRate - After-tax net Interest Rate: Net Income/Net Sales
- NonIndustryIncome - Non-industry income and expenditure/revenue: Net Non-operating Income Ratio
- ContInterestRate - Continuous interest rate (after tax): Net Income-Exclude Disposal Gain or Loss/Net Sales
- OpExpenseRate - Operating Expense Rate: Operating Expenses/Net Sales
- ResearchExpenseRate - Research and development expense rate: (Research and Development Expenses)/Net Sales
- CashFlowRate - Cash flow rate: Cash Flow from Operating/Current Liabilities
- InterestDebtEquity - Interest-bearing debt interest rate: Interest-bearing Debt/Equity
- TaxRate - Tax rate (A): Effective Tax Rate
- BValPerShare - Net Value Per Share (B): Book Value Per Share(B)
- AValPerShare - Net Value Per Share (A): Book Value Per Share(A)
- CValPerShare - Net Value Per Share (C): Book Value Per Share(C)
- EPS - Persistent EPS in the Last Four Seasons: EPS-Net Income
- CashFlowPerShare - Cash Flow Per Share
- RevPerShare - Revenue Per Share (Yuan ¥): Sales Per Share
- IncomePerShare - Operating Profit Per Share (Yuan ¥): Operating Income Per Share
- NetProfitPreTax - Per Share Net profit before tax (Yuan ¥): Pretax Income Per Share
- GrowthRate - Realized Sales Gross Profit Growth Rate
- OpGrowthRate - Operating Profit Growth Rate: Operating Income Growth
- PostTaxGrowthRate - After-tax Net Profit Growth Rate: Net Income Growth
- NetProfitRate - Regular Net Profit Growth Rate: Continuing Operating Income after Tax Growth
- ContNetProfit - Continuous Net Profit Growth Rate: Net Income-Excluding Disposal Gain or Loss Growth
- TotalGrowthRate - Total Asset Growth Rate: Total Asset Growth
- NetValGrowthRate - Net Value Growth Rate: Total Equity Growth
- AssetReturn - Total Asset Return Growth Rate Ratio: Return on Total Asset Growth
- CashReinvest - Cash Reinvestment %: Cash Reinvestment Ratio
- CurrentRatio - Current Ratio
- QuickRatio - Quick Ratio: Acid Test
- InterestExpenseRatio - Interest Expense Ratio: Interest Expenses/Total Revenue
- TotalNetWorth - Total debt/Total net worth: Total Liability/Equity Ratio
- DebtRatio - Debt ratio %: Liability/Total Assets
- Assets - Net worth/Assets: Equity/Total Assets
- FundSuitability - Long-term fund suitability ratio (A): (Long-term Liability+Equity)/Fixed Assets
- InterestCost - Borrowing dependency: Cost of Interest-bearing Debt
- Liability - Contingent liabilities/Net worth: Contingent Liability/Equity
- OpProfit - Operating profit/Paid-in capital: Operating Income/Capital
- NetProfitPreTax - Net profit before tax/Paid-in capital: Pretax Income/Capital
- Inventory - Inventory and accounts receivable/Net value: (Inventory+Accounts Receivables)/Equity
- AssetTurnover - Total Asset Turnover
- AccountTurnover - Accounts Receivable Turnover
- AvgCollection - Average Collection Days: Days Receivable Outstanding
- InventoryTurnover - Inventory Turnover Rate (times)
- FixedAssets - Fixed Assets Turnover Frequency
- NetWorthTurnover - Net Worth Turnover Rate (times): Equity Turnover
- RevPerPerson - Revenue per person: Sales Per Employee
- ProfitPerPerson - Operating profit per person: Operation Income Per Employee
- AllocationPerPerson - Allocation rate per person: Fixed Assets Per Employee

- WorkingCapital - Working Capital to Total Assets
- QuickAssets - Quick Assets/Total Assets
- CurrentAssets - Current Assets/Total Assets
- Cash - Cash/Total Assets
- QuickCurrentLiability - Quick Assets/Current Liability
- CashCurrentLiability - Cash/Current Liability
- CurrentLiabilityAssets - Current Liability to Assets
- LiabilityFunds - Operating Funds to Liability
- InventorytoWorking - Inventory/Working Capital
- InventorytoCurrent - Inventory/Current Liability
- CurrentLiability - Current Liabilities/Liability
- WorkingCapital - Working Capital/Equity
- CurrentLiability - Current Liabilities/Equity
- LTLiability - Long-term Liability to Current Assets
- Earnings - Retained Earnings to Total Asset
- IncometoExpense - Total income/Total expense
- ExpensetoAssets - Total expense/Assets
- AssetTurnover - Current Asset Turnover Rate: Current Assets to Sales
- QuickTurnover - Quick Asset Turnover Rate: Quick Assets to Sales
- CapitalTurnover - Working capitcal Turnover Rate: Working Capital to Sales
- CashtoSale - Cash Turnover Rate: Cash to Sales
- FlowtoSale - Cash Flow to Sales
- FixedAssets - Fixed Assets to Assets
- CurrentLiability - Current Liability to Liability
- CurrentLiability2 - Current Liability to Equity
- EquityLTL - Equity to Long-term Liability
- CashtoAssets - Cash Flow to Total Assets
- FlowtoLiability - Cash Flow to Liability
- CFO - CFO to Assets
- FlowtoEquity - Cash Flow to Equity
- LiabilitytoAssets - Current Liability to Current Assets
- Flag - Liability-Assets Flag: 1 if Total Liability exceeds Total Assets, 0 otherwise
- NetIncome - Net Income to Total Assets
- AssetstoGNP - Total assets to GNP price
- NoCreditInterval - No-credit Interval
- GrosstoSales - Gross Profit to Sales
- IncometoEquity - Net Income to Stockholder's Equity
- LiabilitytoEquity - Liability to Equity
- DFL - Degree of Financial Leverage (DFL)
- EBIT - Interest Coverage Ratio (Interest expense to EBIT)
- NetFlag - Net Income Flag: 1 if Net Income is Negative for the last two years, 0
- Otherwise
- EquitytoLiability - Equity to Liability

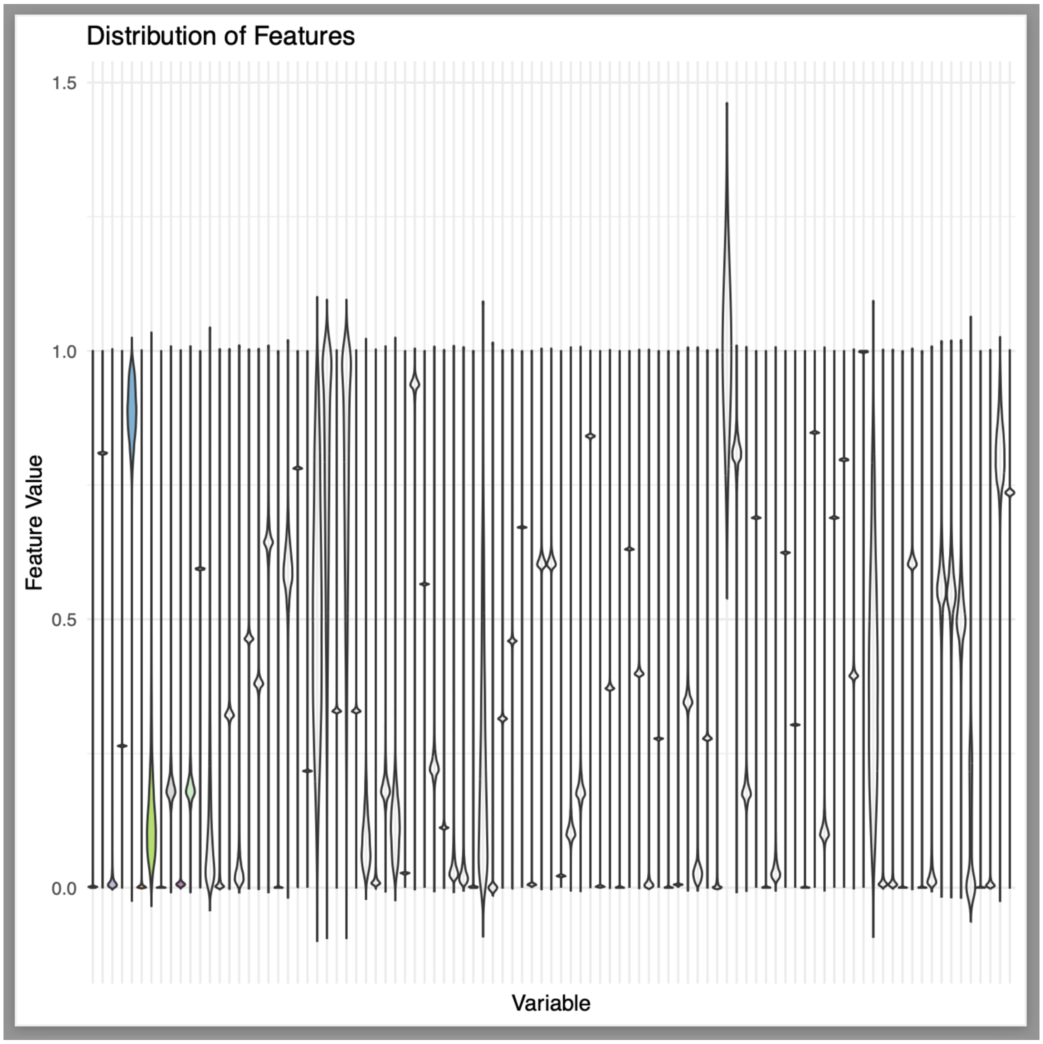
Number of observations: 2819
Missing data amount: 0 (we loaded df and removed all missing values)

3. **Descriptive Statistics**

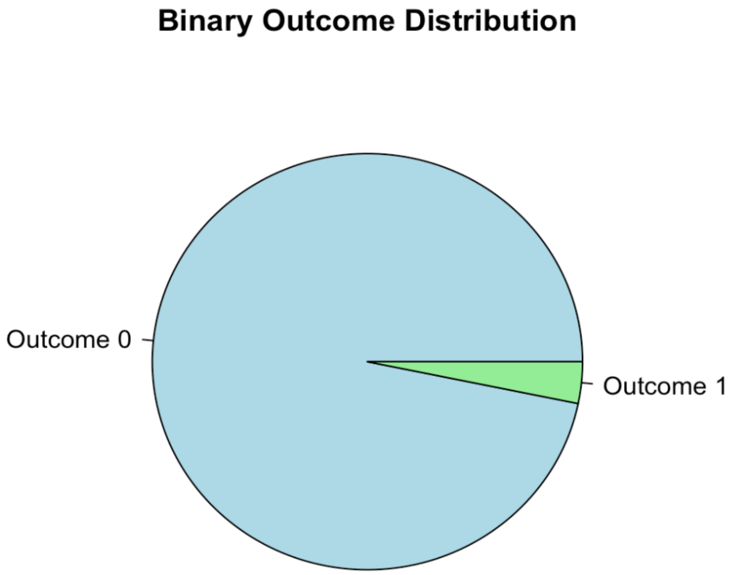
Because we have so many input features that vary from ratios to numbers that are in the millions, it is difficult to visualize this data all together. Below is a violin plot that combines aspects of a kernel density plot and a box plot. It is still difficult to gain any knowledge about our data.



Since many of our inputs are ratios, let’s subset our ratio columns and interpret this violin plot which can be seen below. The plot is still a bit crowded. Although, we are able to see that some ratios have high variability and some have very low. You can tell this by the density curves running along the lines. The more compact areas mean that the ratios have low variability and all companies tend to be in the same range for the most part in regard to that specific feature. On the other hand, the lines that have more space in the curve will have more variability across companies.

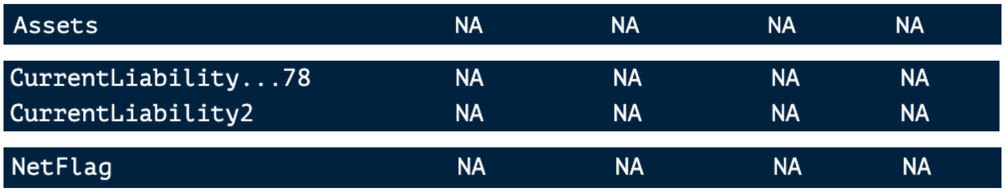


Below is a pie chart that shows the distribution of our response variable. Our output shows that 96% of companies did not go bankrupt and 4% did. This is a highly unbalanced dataset. Bankruptcy class is underrepresented and is the minority class in our data. Some potential reasons for this could be the following: natural class imbalance, sampling bias, data collection, and incomplete data.



4. **Default model**

The response variable is binary. Thus, a logistic model will be used. Our default model that contains all 95 features shows that all features are significant at alpha=0 except for the following features which show up as NA. We will remove these variables along with the coefficients that are “too close” to zero. This will reduce our model from 95 features down to 90 features which all remain significant after removal. The default model gives us no insight as to what can lead to bankruptcy. Based on the summary of the model, it is fair to rule out the default model immediately without looking any further.



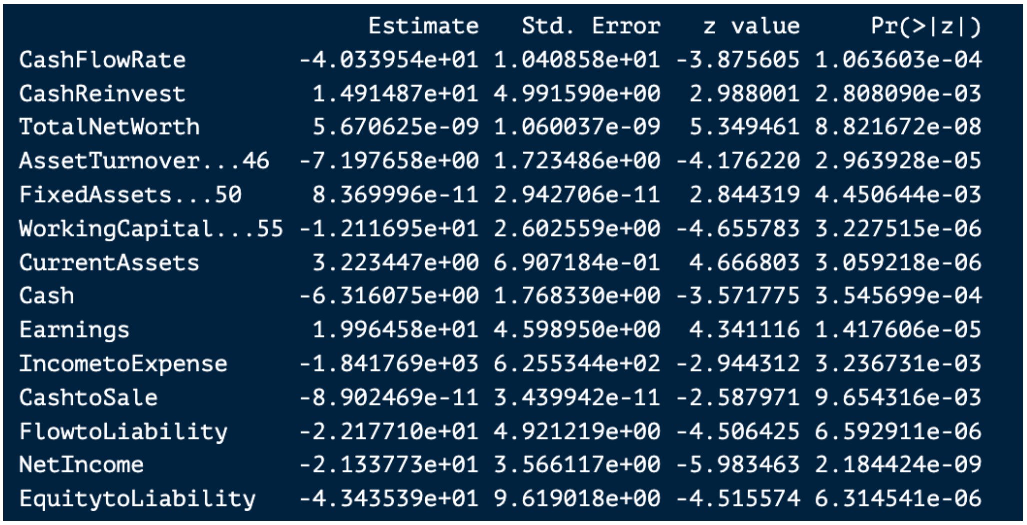
```
Assets                          NA          NA          NA        NA

CurrentLiability...78           NA          NA          NA        NA
CurrentLiability2               NA          NA          NA        NA

NetFlag                         NA          NA          NA        NA
```

5. **Modifications**

There were no characters, strings, dates, or factor level variables in our raw data. Thus, since using logistic regression on a data set with a binary response and numerical/binary input features, there is no need to change the format or type of our variables.

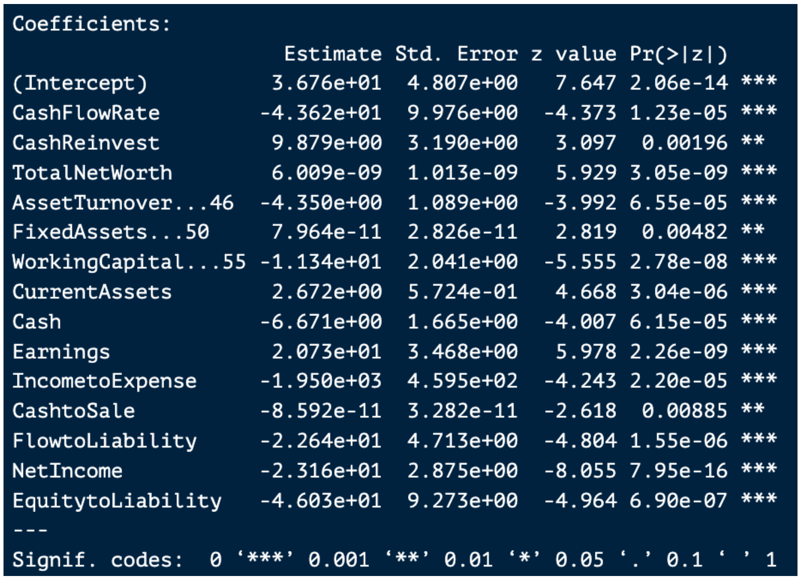
Since the distribution of our response variable is so unbalanced, we must be careful when testing outliers. If the cause of this unbalanceness is from natural class imbalance, then outliers could be crucial when interpreting our data.

Now checking for multicollinearity, we will exclude any feature with VIF > 4. The VIF method addresses how much of the variance of the coefficients is inflated due to correlation with other features. After doing so, we can reduce the model down to 53 features as opposed to 90. Starting with a data set with 95 features, it is important to decrease that number so that we aren't overfitting our model. This model looks much better based on the summary. We are beginning to get closer to our ideal model due to the fact we are narrowing in on significant factors which are listed below.



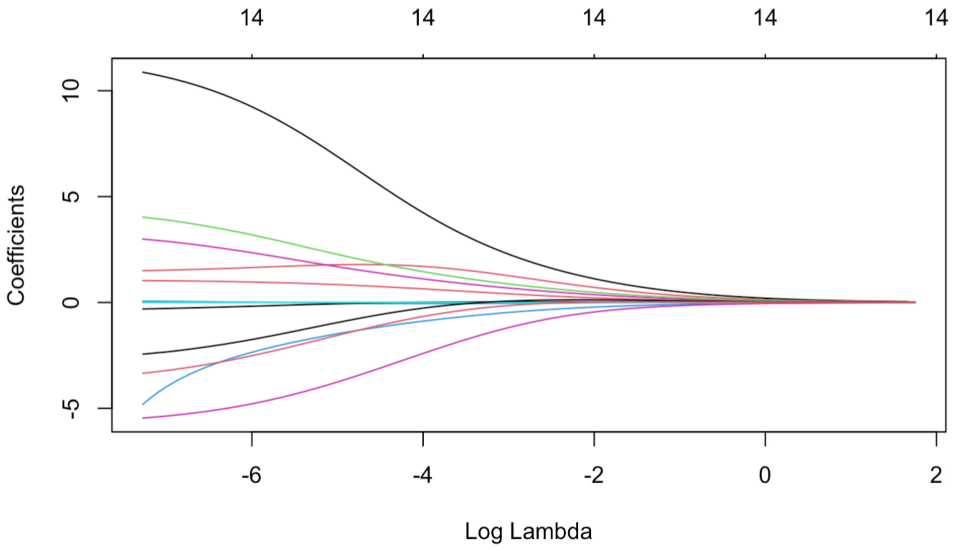
```
                         Estimate   Std. Error   z value     Pr(>|z|)
CashFlowRate         -4.033954e+01 1.040858e+01 -3.875605 1.063603e-04
CashReinvest          1.491487e+01 4.991590e+00  2.988001 2.808090e-03
TotalNetWorth         5.670625e-09 1.060037e-09  5.349461 8.821672e-08
AssetTurnover...46   -7.197658e+00 1.723486e+00 -4.176220 2.963928e-05
FixedAssets...50      8.369996e-11 2.942706e-11  2.844319 4.450644e-03
WorkingCapital...55 -1.211695e+01 2.602559e+00 -4.655783 3.227515e-06
CurrentAssets         3.223447e+00 6.907184e-01  4.666803 3.059218e-06
Cash                 -6.316075e+00 1.768330e+00 -3.571775 3.545699e-04
Earnings              1.996458e+01 4.598950e+00  4.341116 1.417606e-05
IncometoExpense      -1.841769e+03 6.255344e+02 -2.944312 3.236731e-03
CashtoSale           -8.902469e-11 3.439942e-11 -2.587971 9.654316e-03
FlowtoLiability      -2.217710e+01 4.921219e+00 -4.506425 6.592911e-06
NetIncome            -2.133773e+01 3.566117e+00 -5.983463 2.184424e-09
EquitytoLiability    -4.343539e+01 9.619018e+00 -4.515574 6.314541e-06
```

When making financial decisions for a company, it is crucial that the professional knows where the priority lies. Therefore, after finding significant factors by deleting highly correlated input features, we made a new model with only significant factors. This is a good way to narrow down what is most important and crucial. In this step, we will once again reduce our number of features down to 14. After doing so, we can see that it did not affect significance. Sometimes when making this modification, the non-significant features do have underlying insight that can help gain insight. This is not the case in our data, so it is safe to make this modification and keep it. Down below are the summary results.

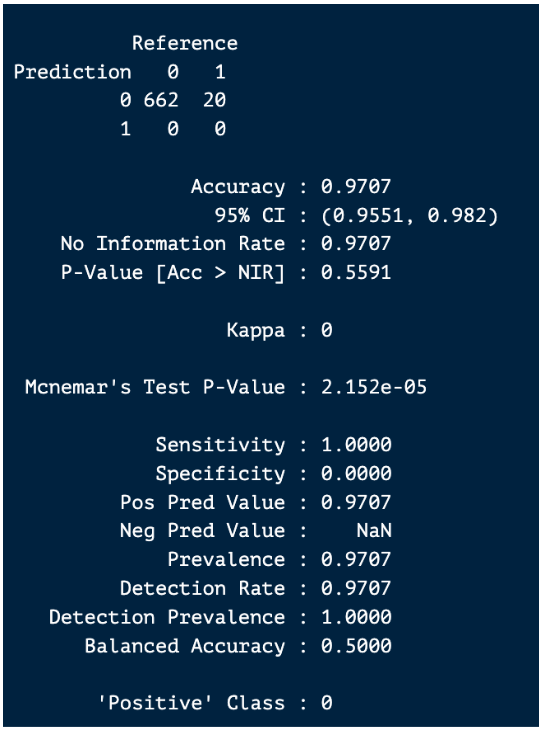


```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)         3.676e+01  4.807e+00   7.647 2.06e-14 ***
CashFlowRate       -4.362e+01  9.976e+00  -4.373 1.23e-05 ***
CashReinvest        9.879e+00  3.190e+00   3.097  0.00196 ** 
TotalNetWorth       6.009e-09  1.013e-09   5.929 3.05e-09 ***
AssetTurnover...46 -4.350e+00  1.089e+00  -3.992 6.55e-05 ***
FixedAssets...50    7.964e-11  2.826e-11   2.819  0.00482 ** 
WorkingCapital...55 -1.134e+01  2.041e+00  -5.555 2.78e-08 ***
CurrentAssets       2.672e+00  5.724e-01   4.668 3.04e-06 ***
Cash               -6.671e+00  1.665e+00  -4.007 6.15e-05 ***
Earnings            2.073e+01  3.468e+00   5.978 2.26e-09 ***
IncometoExpense    -1.950e+03  4.595e+02  -4.243 2.20e-05 ***
CashtoSale         -8.592e-11  3.282e-11  -2.618  0.00885 ** 
FlowtoLiability    -2.264e+01  4.713e+00  -4.804 1.55e-06 ***
NetIncome          -2.316e+01  2.875e+00  -8.055 7.95e-16 ***
EquitytoLiability  -4.603e+01  9.273e+00  -4.964 6.90e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Next, we will use ridge regression to find the best lambda. We will use ridge because we have already selected our features of interest and we do not want to get rid of any more. Ridge will balance the coefficients as needed while the Lasso method will try to get rid of more. We will let R choose a lambda since R selects a lamba that provides a good trade-off between model complexity and goodness of fit to the data. Cross-validation is often used to determine the best lamba that generalizes well to new data. Our best lambda using cross-validation came out to be 5.48. In summary, the plot() output provides a visual of how the coefficients change as the regularization parameter varies, helping you understand the impact of regularization on variable selection and coefficient shrinkage. The accuracy of the ridge model concludes that this is a good model.



Now that we have found significance and narrowed down our data, we will split our remaining data into training and testing data subsets. The purpose of this step is to evaluate the performance of the training data against the testing data to see how well the model will perform on new or unknown data. This will also tell us if our data is overfitted as well.



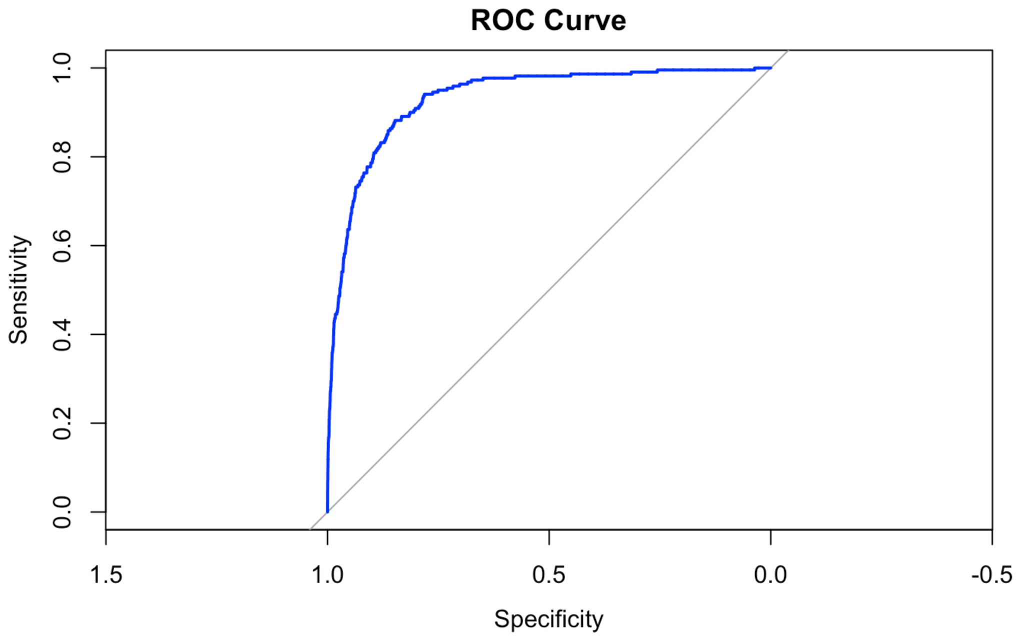
```
          Reference
Prediction   0   1
         0 662  20
         1   0   0

               Accuracy : 0.9707
                 95% CI : (0.9551, 0.982)
    No Information Rate : 0.9707
    P-Value [Acc > NIR] : 0.5591

                  Kappa : 0

 Mcnemar's Test P-Value : 2.152e-05

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.9707
         Neg Pred Value :    NaN
             Prevalence : 0.9707
         Detection Rate : 0.9707
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

6. **Model Comparison**

In linear regression, it is good to use R^2 values to compare models. In logistic regression, using this technique is not easily interpreted. It does not have a direct interpretation as the portion of variance explained. Since logistic regression is designed for binary outcomes, the R^2 value does not capture this meaning. Thus, we will use the Akaike Information Criterion.

The AIC (Akaike Information Criterion) is used for logistic model selection. AIC is based on the likelihood function and penalizes models for the number of parameters. Our goal is to find a balance between model fit and model involvement. We want to take the model with the lowest AIC value because it indicates a good-fitting model with good balance. Our default model was given an AIC value of 15682.77 while our significant feature model was given an AIC value of 1268.61. This modification to our model provided better balance.

Since we have a very unbalanced data set, it is not a good idea to use accuracy for model comparison. Since 96% of our response data classifies a business that did not go bankrupt, the accuracy of our model could be high solely because of its unbalanced nature. Thus, we will use an ROC curve for logistic regression to see which model captures the quality of performance of a binary model. The ROC curve shown below represents the trade-off between sensitivity (true positive rate) and specificity (true negative rate) for different thresholds. Our default model's AUC (area under the curve) value is 0.53, meaning that the performance simulates a random classifier. Given our summary of the default model and our eagerness to rule it out right away was proven correct by this unit of measure. A model that was deemed good using this measure was the model that had removed highly correlated predictors. The AUC value was 0.93. The AUC value for the "significant features only" model from the previously mentioned model was 0.93. Thus, they are about the same and both good to use.



7. **Result/Analysis**

Before telling which model was best for our data, we ruled out a few models for various reasons. Our default model did not provide any insight on our response variable. Thus, we did not use it. After deleting features with high multicollinearity, our new logistic model gave us results. We then ran another logistic model with only our significant features and they all still had their significance at level alpha = 0.05. Additionally, the coefficients matched up very closely to the previous model. It seemed fit to get rid of the extra noise and make interpretation more simple and concise.

Therefore, our final dataset came down to the model with 14 significant predictors (see modification section above). We used it for the ridge regression model. Based on AIC and AUC values, both models were about the same in terms of balance and wellness of fit. The ridge regression regularization is effective for our dataset. Our minimum lambda represents a point where the model generalizes well to new and unknown data. We will not use this model for interpretation due to coefficient shrinkage, but it did give us good knowledge of what our data is capable of.

Finally, after an extensive evaluation and refinement process, our final modified model is a logistic regression model incorporating 14 carefully selected features. This model has demonstrated superior performance in terms of predictive accuracy and generalization to unseen data. The feature selection process aimed to strike a balance between model complexity and interpretability, ensuring that the selected features contribute significantly to the predictive power while avoiding overfitting. The logistic regression framework provides a robust and interpretable solution for our specific problem, and the 14 chosen features capture key variables that influence the outcome. This final model has been assessed through cross-validation and other relevant metrics, affirming its suitability for the intended application and providing a reliable foundation for decision-making and insights.

Based on the suggested model, we have created a simple table below showing how our binary response variable will change if a feature increases by 1 while all of the other features remain the same. Since we are dealing with large sums of money and ratios, looking at the odds ratio value itself will not be helpful in most cases. Instead we will look at how the odds ratio will change based on an increase in a specific feature. There are 3 ways the response can change based on the odds ratio. If the odds ratio is 1, it implies that there is no change. If the odds ratio is greater than 1, it implies an increase in the odds. If the odds ratio is less than 1, it implies a decrease in the odds.

Keep in mind that an increase in odds means a company is more likely to go bankrupt. Therefore, if you are a financial professional, you would want to make decisions that minimize the features that increase in odds. Alternatively, you also want to maximize on the features that show a decrease in odds. Both of these suggestions can be easily tracked and monitored by the table below.

This table suggests that a financial planner would want to focus on maximizing the company’s cash flow, cash reinvestments, cash, flow to liability ratio, net income, and equity to liability ratio. It is interesting to see that having too many current assets could be detrimental. This can be due to it containing high risk. It is also interesting to see that high earnings is also a common factor in bankruptcy. This can be for a number of reasons, but what comes to mind first is that companies with higher earnings will take bigger risks than a company with low earnings. Another reason that comes to mind is lawsuits against a company that is making a lot of money. It is more likely for a person to file a lawsuit against a big company thinking that the business can take a hit. In reality, if it happens too much, it can lead to bankruptcy.

Input Increased by 1 Unit	**Odds Change**

CashFlowRate	Decrease
CashReinvest	Decrease
TotalNetWorth	No Change
AssetTurnover	NA
FixedAssets	NA
WorkingCapital	NA
CurrentAssets	Increase
Cash	Decrease
Earnings	Increase
IncometoExpense	NA
CashtoSale	No Change
FlowtoLiability	Decrease
NetIncome	Decrease
EquitytoLiability	Decrease

CITATIONS

"Bankruptcy Statistics [Updated for [Year]]." *Debt.org,* www.debt.org/bankruptcy/statistics/#:~:text=The%20main%causes%20of%20bankruptc.

Fedesoriano. "Company Bankruptcy Prediction." *Kaggle*, 13 Feb. 2021, www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction.

Deron Liang and Chih-Fong Tsai, deronliang '@' gmail.com; cftsai '@' mgt.ncu.edu.tw, National Central University, Taiwan
The data was obtained from UCI Machine Learning Repository:
https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction

Deron Liang and Chih-Fong Tsai, deronliang '@' gmail.com; cftsai '@' mgt.ncu.edu.tw, National Central University, Taiwan

The data was obtained from UCI Machine Learning Repository:

https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction