**Portfolio Project: Predicting Financial Failure of Companies in the United States**

Brigitte Wetz

Colorado State University Global

MIS581: Capstone – Business Intelligence and Data Analytics

Dr. Jamia Mills

June 30, 2024

**Abstract**

The 2008 financial crisis and 2020 global COVID-19 pandemic resulted in thousands of companies failing and filing for bankruptcy. For companies and investors alike, determining warning signs of potential for failure can be beneficial. Managers and executives can make changes at a company to prevent an eventual bankruptcy filing while investors can decide to withdraw money from companies they have invested in to prevent a large personal financial loss. Financial statement data was downloaded through Python’s ‘yfinance’ library for 1,602 companies which have not filed for bankruptcy in the past four years, and 15 companies which have filed for bankruptcy in the past four years. Gross Domestic Product (GDP) and Consumer Price Index (CPI) were also incorporated into the dataset as well as the calculation of ten financial ratios: Current Ratio, Acid-test Ratio, Debt Ratio, Debt-to-Equity Ratio, Interest Coverage Ratio, Return on Assets Ratio, Return on Equity Ratio, Earnings Per Share Ratio, Return on Total Assets Ratio, and Working Capital Ratio. Based on statistical t-tests, five independent variables (Current Ratio, Acid-test Ratio, Debt Ratio, Return on Assets Ratio, and CPI) were used in four machine learning models: Logistic Regression, Support Vector Machine, XGBoost, and Artificial Neural Networks (ANN). Logistic Regression had an accuracy of 99.65% while the remaining three models each performed with an accuracy of 99.71%. The highest ROC AUC curve value, 0.86, belongs to the XGBoost model, while the lowest value, 0.64, belongs to the Logistic Regression. Only the XGBoost model was able to predict a bankrupt company based on the confusion matrix for each.

**Portfolio Project: Predicting Financial Failure of Companies in the United States**

Companies in the United States have endured two black swan events over the past seventeen years, starting with the global financial crisis in 2008, and, more recently, the COVID-19 global pandemic in 2020 (Bhattacharjee & De, 2022). While financial institutions bore the weight of the 2008 financial crisis, many non-financial companies such as Chrysler and General Motors required government bailout loans to remain in business, and others such as Charter Communications and Lyondell Chemical filed for bankruptcy but then underwent restructuring to re-list on the stock market (Schaefer, 2011).

During COVID-19, when people were forced to stay home, online shopping boomed while some brick-and-mortar stores had a harder time transitioning their consumer base to shop online. With a lack of need for work outfits, clothing retailers JCPenny, J.Crew, and Tailored Brands (owner of Men’s Warehouse and Jos. A. Bank) filed for bankruptcy, while a travel lull for business and pleasure caused Hertz, a rental car company, to also file for bankruptcy (Divine, 2023). Inflation and rising interest rates have put more pressure on businesses as loans have been restructured with higher interest payments, the cost of goods and services have increased, and consumers are spending less resulting in less company revenue.

Investors who are unaware of any financial distress in a company stand to lose most or all their investments if quarterly or annual reports are not analyzed on a regular basis. Financial institutions regularly calculate financial ratios using a company’s financial statements to determine credit worthiness and if a loan warrants a higher interest rate when a company wishes to refinance. If a company’s financial ratios become undesirable compared to an industry average, whether too high or too low, it may be harder for companies to obtain new funding (McGregor, 2022). For companies themselves, internal analyses of their financial health may lead to the realization that changes must be made before the company reaches the point of filing for bankruptcy.

The specific problem of this research that will seek to be answered is: how much of an effect do interest rate and inflation rate have on financial ratios of a company in predicting financial distress of a company in the United States through machine learning in addition to financial ratios?

**Objectives**

The objective of this research is to test multiple machine learning models and develop an accurate machine learning model for predicting financial distress of a company in the United States by incorporating interest rate, inflation rate, and financial ratios such as liquidity, leverage, efficiency, profitability, and market value.

**Organizational Benefits**

If there is a correlation between either or both rates with changes in financial ratios, companies may have a better idea of which aspects of the business would be most beneficial from preliminary actions to prevent financial ratios from worsening leading to bankruptcy. Healthy financial ratios will also allow for better credit ratings and therefore better interest rates for loans if a company wanted to expand its operations or purchase a competitor. Once a model is developed, companies can create a dashboard that is updated on a regular basis (daily, weekly, monthly) with alerts set up for the most important indicators from the model. If one of these indicators or variables changes in a way that predicts future financial distress, a root cause analysis can begin to lessen the effects or possibly reverse the outcome for the company.

For an investor, the knowledge of this correlation is also beneficial to prevent a large loss on his or her investment. From an investor perspective, even without daily financial data from a company, a dashboard can still be useful by incorporating quarterly reports into the predictive model, as well as other information an investor may find important such as stock price, press releases, and social media sentiment.

**Overview of Study**

The use of financial ratios to determine a company’s financial health as well as a predictor in determining financial distress leading to bankruptcy has been well documented for several decades. This study builds on previous work by incorporating macroeconomic data, Gross Domestic Product and Consumer Price Index, with these financial ratios into machine learning algorithms. Data from balance sheets, cash flow statements, and income statements were used to calculate financial ratios which resulted in 5,781 data points. While logistic regression has been used to predict company failure since the 1970’s, more modern machine learning algorithms such as Artificial Neural Networks (ANN), which are more of a black-box calculation, are able to find correlations and connections in data for greater predictive accuracy, much like the neural network of the human brain.

**Research Questions and Hypotheses**

**Research Question 1**

The first research question to explore is: Can bankruptcy be correlated to any financial statement variables, ratios, or economic variables (Gross Domestic Product growth, Consumer Price Index as an indication of inflation)? Hypotheses are given as:

H0: Company bankruptcy cannot be correlated with any variables.

H1: At least one variable has a moderate to strong correlation to bankruptcy.

This question can be answered by conducting a statistical analysis of each variable against the binary target of whether a company has filed for bankruptcy (flagged with the number 1) or not (flagged with the number 0). Two correlation techniques will be used: Point-Biserial Correlation and two-sample t-test (normal and Welch) to determine p-values. Point-Biserial Correlation measures how much two variables are associated with one another (Statistics Solutions, n.d.) and can be used on data with unequal variances (Bonett, 2020) while two-sample t-tests compare the statistical significance between the means of two independent groups or populations (JMP, n.d.). P-values measure the probability of the current value of the coefficient being the same if the two variables were unrelated. Common practice is to state that if a p-value is below 0.05, then the two variables have a statistically significant relationship with one another.

**Research Question 2**

The second research question to explore is: Can bankruptcy be predicted by machine learning models? Hypotheses are given as:

H0: Machine learning models cannot predict a company filing for bankruptcy.

H1: At least one machine learning model can predict a company filing for bankruptcy.

Four machine learning models will be used for supervised learning and dividing the data into training and testing sets: Logistic Regression, Support Vector Machine (SVM), XGBoost, and Artificial Neural Networks (ANN). Some machine learning models, such as logistic regression, perform best when data is normally distributed. Therefore, histograms for each variable will be analyzed for skewness and kurtosis. Those variables which are not normally distributed may be scaled in Python. Any variables with large variances will also require scaling. XGBoost, ANN, and SVM do not require normally distributed data for models to perform well.

A correlation matrix will be constructed to view how strong of a correlation exists between variables. When two variables are highly correlated, including both in a model will cause issues in the p-values and confidence intervals, leading to a model which is not an accurate depiction of the predictive nature of the model (Singh et al., 2023). Precision, recall, F1-score, and ROC curves will be used to evaluate the predictive performance of each model. Shapley Additive exPlanations (SHAP) is another technique for evaluating variables and their effect on the target outcome (Mickle & Deb, 2022).

**Literature Review**

Edward Altman was the pioneer of predicting a company’s financial distress. He developed the Z-score model in 1968 which is still used today (Altman, 2000). His model employs five financial ratios: working capital/total assets, retained earnings/total assets, EBIT/total assets, market value equity/book value of total debt, and sales/total assets in conjunction with unique coefficients for each (Altman, 1968). Multiplying each financial ratio by its coefficient and summing the values results in a final value known as the z-score. A company is healthy if this value is 3.00 or above, in distress when the value is below 1.88, and in danger of becoming distressed when between 1.88 and 2.99 (Simanca et al., 2022). This initial research was done using data from the manufacturing industry, and other researchers have since applied this technique to other industries. There is also evidence that financial stress prediction models are best developed for each country, given differences in accounting standards and company sizes (Abdelkader & Wahba, 2024).

Hassan & Nassar (2023) used four financial ratios to measure how well a company performed in terms of its return on assets, another financial ratio. They also incorporated macroeconomic data on inflation, unemployment rate, GDP, and government bond rates. Return on assets is one of the most used ratios to determine the financial health of a company, as a higher return is better to lenders. It was determined that GDP and debt-to-equity had the biggest impact on a company’s financial health.

In a study by Kristanti & Dhaniswara (2023), five financial ratios were used to predict financial distress for companies in the industrial sector on the Indonesian Stock Exchange. Instead of predicting with continuous data, dummy variables for each financial ratio were created based on various cutoffs for each. Two machine learning models, artificial neural networks and logistic regression, were created and evaluated for accuracy. Logistic regression outperformed the artificial neural networks model quite considerably, with an accuracy of 98% in predicting financial distress for a company versus 82.5% accuracy.

Another method for a logistic regression model is to include cluster analysis (Li, 2016). By incorporating cluster analysis, multicollinearity issues can be resolved in a more effective way. Each cluster then becomes a dummy variable fed into the logistic regression model. In Li’s study, the cluster-based logistic regression model had an accuracy of 98.4% while logistic regression on its own had an accuracy of 97.8%.

While Li’s research focused on the travel industry, its principles can be applied to this study by clustering financial data according to sector or strength (or weakness) of financial ratios. This study will also include other financial ratios as well as those from other studies in a more comprehensive attempt to model financial distress (for a total of ten). A combination of machine learning models from other studies will also be developed with one general machine learning model for all data and individual models for each of the sectors represented on the United Stock Exchange which contain enough data.

**Research Design**

**Methodology**

A quantitative approach was used for this study for the comparisons of independent variables to a dependent variable: the binary target, i.e. whether a company declares bankruptcy or not. Financial ratios allow for comparisons between small-cap and large-cap corporations to eliminate large differences in financial statement values. For example, Microsoft has reported revenue in the trillions of dollars, and comparing this value to a smaller company which may only have reported revenue in the hundreds of millions can create a large skew in an analysis.

Three different types of ratios are used in this study: liquidity, leverage, and profitability. Liquidity ratios are used to determine if a company can pay its short- and long-term obligations in a timely manner. Leverage ratios determine how a company finances its operations, whether from loans or shareholders. Profitability ratios measure how well a company can generate income relative to various aspects of the business such as revenue or operating costs.

The consumer price index is a monthly calculation of the price difference in goods and services commonly purchased by consumers in the United States and is a common method to measure inflation (Baxa & Widdersheim, 2023). During COVID-19, the index fell to 1.2 in 2020 but reached a peak of 6.6 in 2022. Inflation affects all sectors as the cost of manufacturing increases and revenues may decrease due to consumers pulling back on discretionary spending. GDP values are indicators of how well a country’s economy is performing and can aid in identifying recessions (Stobierski, 2021). It represents the yearly value of goods and services produced in a country. Values can be compared to the previous year to determine the growth (or contraction) rate. In 2020, the GDP contracted -2.8% in the United States due to COVID-19 but rebounded to a growth rate of 5.9% in 2021. Once inflation began increasing, the rate fell to 1.9% in 2022, 2.5% in 2023, and 1.3% thus far in 2024. Comparing older classification techniques (logistic regression) with newer techniques (XGBoost, ANN) will help companies and investors when deciding which model to use in an independent analysis..

**Methods**

Currently active companies were chosen who trade on the New York Stock Exchange or NASDAQ. A list of companies who have filed for bankruptcy for the past four years and the filing date was obtained from <https://stockanalysis.com/actions/bankruptcies/2024/>. Financial statement data (balance sheets, income statements, and cashflow statements) for these two groups of companies were downloaded from Python’s ‘yfinance’ library using Jupyter Notebook. Eleven sectors are represented by these stocks: Basic Materials, Communication, Consumer Cyclical, Consumer Defensive, Energy, Finance, Healthcare, Industrials, Real Estate, Technology, and Utilities. Table 1 gives the definition of each sector according to yahoo!finance (yahoo!finance, 2024).

**Table 1**

*Sector Definitions*

|  |  |
| --- | --- |
| **Sector** | **Definition** |
| Basic Materials | Manufacturers of chemicals, building materials, and paper products or are active in the exploration and processing of commodities such as oil or gold. |
| Communication | Providers of services which allow users to communicate using fixed-line networks or wireless networks, advertising and marketing services, entertainment, or interactive media. |
| Consumer Cyclical | Companies in the retail, auto, restaurant, lodging, or entertainment industries. |
| Consumer Defensive | Manufacturers of food, beverages, household and personal products, tobacco; providers of education or training. |
| Energy | Oil and gas producers or refiners, oil and gas equipment manufacturers, pipeline operators. |
| Finance | Financial service providers such as banks, asset management companies, credit cards, investment brokerages, insurance companies. |
| Healthcare | Includes hospitals, pharmaceutical and biotechnology companies, medical equipment suppliers. |
| Industrials | Manufacturers of machinery and hand-held tools. Also includes aerospace, defense, and transportation companies. |
| Real Estate | Companies which develop, acquire, and manage real estate properties. |
|  |  |
| Technology | Manufacturers of computers, data storage and networking products, semiconductors, operating systems, and applications. |
| Utilities | Providers of utilities to consumers such as electric, gas, and water. |

Two separate data frames were created when downloading financial data, one for bankrupt companies and one for non-bankrupt companies. These two data frames were then combined and only the columns needed for calculating the financial ratios in Table 2 were saved. Once the ratios were created, financial statement data and rows with any missing values were dropped from the data frame. A linear regression was performed for each company’s financial ratio over the time of the data and slope data extracted for analysis. The rationale for this step was to analyze if bankrupt companies have a higher slope value (either negative or positive depending on financial ratio) than non-bankrupt companies and if a trend in slope can be a predictor in a company’s financial health.

**Table 2**

*Financial Ratios*

|  |  |  |
| --- | --- | --- |
| **Ratio Type** | **Ratio** | **Formula** |
| Liquidity | Current | Current Assets / Current Liabilities |
| Liquidity | Acid-Test | (Current Assets – Inventory) / Current Liabilities |
| Liquidity | Working Capital | (Current Assets – Current Liabilities) / Total Assets |
| Leverage | Debt | Total Liabilities / Total Assets |
| Leverage | Debt-to-Equity | Total Liabilities / Stockholders Equity |
| Leverage | Interest Coverage | EBIT / Interest Expense |
| Profitability | Return on Assets | Net Income / Total Assets |
| Profitability | Return on Equity | Net Income / Stockholders Equity |
| Profitability | Earnings Per Share | Given in financial data (Net Earnings / Total Shares Outstanding) |
| Profitability | Return on Total Assets | EBIT / Total Assets |
|  |  |  |

Consumer price indices (CPI) and gross domestic product (GDP) growth were downloaded as Excel files from the United States Bureau of Labor Statistics and The World Bank respectively. The average CPI value for each year was calculated and imported into Python as well as the yearly GDP growth file and joined with the financial ratio data frame on reporting year. The final data frame consisted of 5,766 datapoints for non-bankrupt companies and 15 datapoints for bankrupt companies. Figure 1 shows a breakdown of financial sectors for each binary outcome.

**Figure 1**

*Number of Datapoints for Each Sector and Binary Outcome*

|  |  |
| --- | --- |
| A graph of a bankruptcy  Description automatically generated with medium confidence | A graph of a number of different colored bars  Description automatically generated with medium confidence |
| *Note.* The figure on the left is the number of non-bankrupt companies in each sector; the figure on the right is the number of bankrupt companies in each sector. | |

Feature selection involved computing a correlation matrix for the independent variables in addition to ANOVA, feature importance calculations, and two-sample t-tests between independent and dependent variables. The data set was split into training and testing sets with a testing size of 30%. Variables with a statistically significant difference (p-value less than 0.05) between means of bankrupt and non-bankrupt companies, based on t-tests, were selected for each initial model building as long as their correlation coefficients were low to prevent multicollinearity issues. Other combinations of variables were included in models and the highest accuracy model chosen as the best. Highest accuracy is defined as the model with the highest count of true positives and true negatives divided by the total number of predictions in the testing set.

After feature selection, Logistic Regression, SVG, XGBoost, and ANN models were created using Python’s ‘scikit-learn’ library. Receiver Operating Characteristic (ROC) curves, which plot the true positive rate against the false positive rate, were also constructed and evaluated using the Area Under the Curve (AUC) values. In general, when an AUC value is above 0.6, then the model has a higher probability of distinguishing between bankrupt and non-bankrupt companies. Values for this analysis were compared between the four models to determine which model is the best predictor/classifier.

**Limitations**

Python’s ‘yfinance’ data is limited to the past four years, with data needed for calculating financial ratios missing from several companies. This led to a small population sample for companies which filed for bankruptcy. While quarterly data is available and would have added a level of granularity to the analysis, yearly 10-K reports were used for a representation of a company’s financial data for the year it declared bankruptcy. Several companies, a majority of which are in the financial sector, were missing data needed for the calculation of financial ratios which affected an initial goal to model each individual sector to compare accuracies between sectors.

**Ethical Considerations**

Data used for this project is publicly available and prone to hackers altering the accuracy. Therefore, data was downloaded from each source, saved to an Excel file, and then imported into Jupyter Notebook to preserve accuracy or data disappearing online. The program for scraping financial data using Python’s yfinance library was obtained from a public notebook on Kaggle and modified (Kevin, 2024). Spot checks were performed on the financial statement data downloaded through Python’s yfinance library for accuracy against public filings on the U.S. Security and Exchange Commission’s website.

No conflicts of interest apply to this project. There is concern that companies use the results of this study to note which variables could be indicators that a company will file for bankruptcy and worry about investors selling stock prematurely, thereby creating a false crash in the company’s value. Conversely, managers may attempt to manipulate financial data reported to the government to artificially make variables or financial variables more appealing than they are to make the company appear to be in better financial health.

**Findings**

Descriptive statistics were performed on the final data set (Table 3) and the limitation on number on bankrupt company datapoints greatly affects the standard deviation and variance between the binary outcomes. The GDP growth for years in which companies declared bankruptcy averaged 2.35% versus 2.01% for companies which did not declare bankruptcy while the Consumer Price Index averaged 4.07 for companies which did not declare bankruptcy and 5.61 for those which did. Table 4 shows the descriptive statistics for the slope calculations.

**Table 3**

*Descriptive Statistics of Data Set by Binary Outcome*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Status** | **N** | **Minimum** | **Maximum** | **Mean** | **Std. Dev.** | **Variance** |
| GDP\_Growth | 0 | 5,766 | -2.77 | 5.95 | 2.01 | 3.09 | 9.54 |
| 1 | 15 | 1.94 | 5.95 | 2.35 | 1.03 | 0.98 |
| CPI\_Yearly\_Avg | 0 | 5,766 | 1.70 | 6.15 | 4.07 | 1.63 | 2.65 |
| 1 | 15 | 3.57 | 6.15 | 5.61 | 0.84 | 0.66 |
| CR | 0 | 5,766 | 0.00 | 54.02 | 2.39 | 2.42 | 5.85 |
| 1 | 15 | 0.11 | 3.01 | 1.26 | 0.88 | 0.72 |
| ATR | 0 | 5,766 | -2.15 | 54.02 | 1.79 | 2.25 | 5.06 |
| 1 | 15 | 0.08 | 2.61 | 0.76 | 0.78 | 0.57 |
| DR | 0 | 5,766 | 0.01 | 251.33 | 0.73 | 4.03 | 16.27 |
| 1 | 15 | 0.38 | 2.84 | 1.26 | 0.68 | 0.43 |
| DTER | 0 | 5,766 | -16,549.56 | 2,290.82 | -3.58 | 245.84 | 60,426.24 |
| 1 | 15 | -38.00 | 17.65 | -2.52 | 12.53 | 146.49 |
| ICR | 0 | 5,766 | -176,158.00 | 31,447.78 | -26.35 | 2,639.66 | 6,966,594.29 |
| 1 | 15 | -84.44 | 24.56 | -11.94 | 27.44 | 702.67 |
| ROAR | 0 | 5,766 | -64.46 | 1.48 | -0.03 | 1.19 | 1.42 |
| 1 | 15 | -4.58 | 0.16 | -0.69 | 1.17 | 1.27 |
| ROER | 0 | 5,766 | -982.00 | 1,288.56 | 0.03 | 24.29 | 590.05 |
| 1 | 15 | -9.26 | 17.30 | 0.71 | 5.52 | 28.42 |
| EPS | 0 | 5,766 | -8,928.60 | 525.20 | -1.55 | 146.24 | 21,382.88 |
| 1 | 15 | -21.00 | 3.23 | -5.21 | 7.41 | 51.20 |
| ROTAR | 0 | 5,766 | -44.85 | 1.91 | 0.00 | 0.92 | 0.84 |
| 1 | 15 | -3.48 | 0.23 | -0.57 | 0.92 | 0.79 |
| WCR | 0 | 5,766 | -235.06 | 0.95 | 0.10 | 3.74 | 13.97 |
| 1 | 15 | -2.23 | 0.31 | -0.24 | 0.71 | 0.47 |
|  |  |  |  |  |  |  |  |

*Note.* CR = Current Ratio, ATR = Acid-test Ratio, DR = Debt Ratio, DTER = Debt-to-Equity Ratio, ICR = Interest Coverage Ratio, ROAR = Return on Assets Ratio, ROER = Return on Equity Ratio, EPS = Earnings Per Share, ROTAR = Return on Total Assets Ratio, WCR = Working Capital Ratio

**Table 4**

*Descriptive Statistics of Slope Calculations*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Status** | **N** | **Minimum** | **Maximum** | **Mean** | **Std. Dev.** | **Variance** |
| GDP\_Growth | 0 | 1,487 | -4.01 | 8.71 | 1.07 | 1.56 | 2.43 |
| 1 | 62 | -5.06 | 8.71 | 1.65 | 3.02 | 8.99 |
| CPI\_Yearly\_Avg | 0 | 1,487 | -1.38 | 2.57 | 1.20 | 0.61 | 0.37 |
| 1 | 62 | -0.50 | 2.57 | 1.56 | 0.71 | 0.49 |
| CR | 0 | 1,487 | -22.49 | 18.32 | -0.06 | 1.15 | 1.33 |
| 1 | 62 | -9.97 | 6.26 | -0.20 | 1.60 | 2.53 |
| ATR | 0 | 1,487 | -23.03 | 17.95 | -0.09 | 1.14 | 1.29 |
| 1 | 62 | -9.98 | 6.37 | -0.20 | 1.59 | 2.49 |
| DR | 0 | 1,487 | -23.34 | 0.56 | -0.04 | 0.63 | 0.40 |
| 1 | 62 | -2.52 | 72.41 | 1.11 | 9.23 | 83.77 |
| DTER | 0 | 1,487 | -1,416.89 | 8,277.13 | 4.60 | 219.69 | 48,229.56 |
| 1 | 62 | -91.26 | 8.67 | -2.59 | 13.05 | 167.52 |
| ICR | 0 | 1,487 | -175,831.71 | 26,889.83 | -114.59 | 4,665.39 | 21,751,247.49 |
| 1 | 62 | -2,300.53 | 4,715.67 | 65.85 | 753.77 | 559,010.99 |
| ROAR | 0 | 1,487 | -1.13 | 1.06 | 0.01 | 0.09 | 0.01 |
| 1 | 62 | -10.89 | 1.96 | -0.31 | 1.53 | 2.31 |
| ROER | 0 | 1,487 | -644.65 | 37.10 | -0.76 | 19.17 | 367.21 |
| 1 | 62 | -8.94 | 37.86 | 0.81 | 5.69 | 31.89 |
| EPS | 0 | 1,487 | -549.49 | 373.09 | 0.79 | 18.57 | 344.78 |
| 1 | 62 | -75.56 | 5,353.64 | 161.32 | 873.93 | 751,428.25 |
| ROTAR | 0 | 1,487 | -1.24 | 1.05 | 0.01 | 0.10 | 0.01 |
| 1 | 62 | -6.54 | 1.95 | -0.24 | 1.03 | 1.05 |
| WCR | 0 | 1,487 | -0.56 | 2.18 | 0.00 | 0.10 | 0.01 |
| 1 | 62 | -69.20 | 0.44 | -1.20 | 8.78 | 75.89 |
|  |  |  |  |  |  |  |  |

Skew and kurtosis were calculated to determine if each variable is normally distributed. Most variables show non-normal distributions with large amounts of kurtosis (Table 5). This data will require scaling for use in machine learning models.

**Table 5**

*Skew and Kurtosis for Each Variable*

|  |  |  |
| --- | --- | --- |
| **Variable** | **Skew** | **Kurtosis** |
| GDP\_Growth | -0.32 | -0.95 |
| CPI\_Yearly\_Avg | -0.18 | -1.22 |
| CR | 6.03 | 70.36 |
| ATR | 6.93 | 88.54 |
| DR | 50.93 | 2,869.98 |
| DTER | -55.84 | 3,631.04 |
| ICR | -53.43 | 3,491.12 |
| ROAR | -41.23 | 1,911.33 |
| ROER | 13.03 | 1,888.39 |
| EPS | -50.62 | 2,828.82 |
| ROTAR | -36.54 | 1,503.36 |
| WCR | -52.38 | 3,000.80 |
|  |  |  |

To determine if the difference in means between the two groups of companies is significant, t-tests and ANOVA were performed. If the ratio of the variance for the two groups was greater than 5, then Welch’s t-test was used, which assumes unequal variances between two populations. For both the t-test and ANOVA, results are considered statistically significant when the p-values are less than 0.05. Based on this criteria, the variables which are considered statistically significant between the two groups of companies are: CPI, Current Ratio, Acid-test Ratio, Return on Assets Ratio, and Return on Total Assets Ratio (Table 6). When comparing the slopes of each company’s variable over time for the two groups of companies, GDP and CPI are the only variables which are statistically significant for t-tests. However, comparing the variances with ANOVA, the Debt Ratio, Return on Assets Ratio, EPS, Return on Total Assets Ratio, and Working Capital Ratio are also considered to be of statistical significance.

**Table 6**

*Results of T-tests and ANOVA*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **T-test P-value** | **F-Score** | **F-Score P-value** | **Slope T-test P-value** | **Slope F-Score** | **Slope F-Score P-value** |
| GDP\_Growth | 0.2127\* | 0.1915 | 0.6617 | 0.0063 | 7.4755 | 0.0063 |
| CPI\_Yearly\_Avg | 0.0002 | 13.5218 | 0.0002 | 6.8657x10-6 | 20.3698 | 6.8657x10-6 |
| CR | 0.0002\* | 3.2412 | 0.0719 | 0.3488 | 0.8784 | 0.3488 |
| ATR | 0.0002\* | 3.1401 | 0.0764 | 0.4659 | 0.5318 | 0.4660 |
| DR | 0.0108\* | 0.2549 | 0.6137 | 0.3296\* | 21.1275 | 4.6464x10-6 |
| DTER | 0.8179\* | 0.0003 | 0.9867 | 0.2258\* | 0.0663 | 0.7967 |
| ICR | 0.6846\* | 0.0004 | 0.9831 | 0.2429\* | 0.0926 | 0.7610 |
| ROAR | 0.0322 | 4.5897 | 0.0322 | 0.1031\* | 60.9522 | 1.0700x10-14 |
| ROER | 0.6484\* | 0.0117 | 0.9138 | 0.0769\* | 0.4113 | 0.5214 |
| EPS | 0.1828\* | 0.0094 | 0.9228 | 0.1532\* | 50.3739 | 1.9287x10-12 |
| ROTAR | 0.0161 | 5.7972 | 0.0161 | 0.0581\* | 74.2057 | 1.7066x10-17 |
| WCR | 0.0969\* | 0.1194 | 0.7296 | 0.2869\* | 28.0153 | 1.3764x10-7 |
|  |  |  |  |  |  |  |

*Note*. Values with an (\*) indicate Welch’s t-test

Table 7 shows the correlation coefficient and p-value for each variable to the binary target. Like Pearson’s correlation coefficient, a value of -1 indicates a perfect negative relationship between two variables, a value of +1 indicates a perfect positive relationship between two variables, and a value of 0 indicates two variables are not correlated with one another. These results show there are three variables that have a statistically significant correlation with bankruptcy: Return on Assets Ratio, Return on Total Assets Ratio, and CPI. Given that two separate tests have found statistical significance between the independent variables and the dependent variable, the null hypothesis for research question one can be rejected.

**Table 7**

*Point-Biserial Correlation Results*

|  |  |  |
| --- | --- | --- |
| **Variable** | **Correlation** | **P-value** |
| GDP\_Growth | 0.0058 | 0.6617 |
| CPI\_Yearly\_Avg | 0.0483 | 0.0002 |
| CR | -0.0237 | 0.0719 |
| ATR | -0.0233 | 0.0764 |
| DR | 0.0066 | 0.6137 |
| DTER | 0.0002 | 0.9867 |
| ICR | 0.0003 | 0.9831 |
| ROAR | -0.0282 | 0.0322 |
| ROER | 0.0014 | 0.9138 |
| EPS | -0.0013 | 0.9228 |
| ROTAR | -0.0317 | 0.0161 |
| WCR | -0.0046 | 0.7297 |
|  |  |  |

**Logistic Regression Model**

Logistic regression models perform best when data is normally distributed and does not display multicollinearity. A correlation matrix and heatmap are visual aids in determining which variables are correlated with one another (Figure 2). Another method is the Variance Inflation Factor (VIF) which compares interactions between each variable. The results of VIF analysis are in line with the correlation matrix. Return on Assets and Return on Total Assets showed the greatest correlation, followed by Working Capital, Debt Ratio, Current Ratio, and Acid-test Ratio.

**Figure 2**

*Variance Inflation Factor and Heatmap of Correlation Matrix*

A screenshot of a computer

Description automatically generatedVIF

22.50

18.97

53.22

1.26

1.01

170.72

1.26

1.02

106.44

56.73

1.72

4.02

*Note*. VIF values on the left correspond to the same variables, in order, of the heatmap.

A logistic regression model was created using five variables: Current Ratio, Acid-test Ratio, Debt Ratio, Return on Assets Ratio, and CPI\_Yearly\_Avg. Python’s ‘scikit-learn’ library’s ‘RobustScaler’ package was used for scaling data due to its ability to handle outliers and skewness. The ‘lbfgs’ solver was used as well as a regularization value of 1 to avoid overfitting the model, with a maximum number of iterations of 1,000. The model resulted in an accuracy of 99.65%, and ROC AUC value of 0.64 (Figure 3). Precision and recall are unreliable due to a lack of data for bankrupt companies as the model failed to predict any bankrupt companies. However, based on the accuracy and AUC values, the null hypothesis for research question two can be rejected.

**Figure 3**

*Confusion Matrix and ROC AUC Curve for Logistic Regression Model*

A graph of positive and negative variables

Description automatically generated

**Support Vector Machine Model**

The same process for building a SVM model was followed from the logistic regression models in terms of variables used and train-test split. An ‘rbf’ kernel was used in the model with a gamma value of 1 to prohibit overfitting. Accuracy was slightly higher than the logistic regression model at 99.71% and the ROC AUC curve increased to 0.78. Figure 4 shows the confusion matrix and ROC AUC curve for the model. The model was not able to correctly predict any companies filing for bankruptcy. Precision and recall are unreliable due to a lack of data for bankrupt companies as the model failed to predict any bankrupt companies.

**Figure 4**

*Confusion Matrix and ROC AUC Curve for Support Vector Machine Model*

A graph of positive and negative values

Description automatically generated with medium confidence

**XGBoost Model**

The same process for building an XGBoost model was followed from the logistic regression and SVM models. However, as XGBoost is a decision tree-based model, data was not scaled as in the previous models. A binary logistic objective with an evaluation metric of AUC was used with a learning rate of 0.1 and 1,000 for the number of estimators. While the accuracy was the same as the SVM model at 99.71%, the ROC AUC curve increased to 0.86. Figure 5 shows the confusion matrix, ROC AUC curve, and SHAP values for XGBoost. SHAP analysis shows that higher values of the Return on Assets Ratio indicate companies are less likely to file for bankruptcy, while higher CPI values are an indication that a company is more likely to file for bankruptcy. XGBoost was able to correctly predict a company filing for bankruptcy based on the confusion matrix. The precision calculated for the model was 0.5 and the recall was 0.2.

**Figure 5**

*Confusion Matrix, ROC AUC Curve, and SHAP Values for XGBoost Model*

A graph with a line going up

Description automatically generated

**Artificial Neural Network Model**

The same process for building an ANN model was followed from the logistic regression and SVM models. Robust scaling was used to transform the data due to large amounts of variance and kurtosis. Parameters for the model include using an activation of logistic and solver of ‘lbfgs’. The alpha parameter was set to 1 to mitigate overfitting of the data and a max iteration number of 1,000 was used. While the accuracy was the same as the SVM and XGBoost models, 99.71%, the ROC AUC curve was 0.79. Figure 6 shows the confusion matrix and ROC AUC curve for the model. Precision and recall are unreliable due to a lack of data for bankrupt companies as the model failed to predict any bankrupt companies.

**Figure 6**

*Confusion Matrix and ROC AUC Curve ANN Model*

**A graph of positive and negative variables

Description automatically generated**

**Conclusion**

Both research questions’ null hypotheses were rejected based on statistical testing and machine learning model performance. Of the twelve initial variables, five were deemed statistically significant in predicting company bankruptcy: Current Ratio, Acid-test Ratio, Debt Ratio, Return on Assets Ratio, and CPI. Of the four machine learning models built, the accuracy values were most consistent, while the highest AUC belonged to the XGBoost model. This model was also the only one to correctly predict a company filing for bankruptcy. It is possible the models suffer from overfitting of the data given the small sample of bankrupt companies and therefore require more tweaking.

**Table 8**

*Summary of Model Results*

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | **AUC** |
| Logistic Regression | 0.9965 | 0.64 |
| SVM | 0.9971 | 0.78 |
| XGBoost | 0.9971 | 0.86 |
| ANN | 0.9971 | 0.79 |
|  |  |  |

**Recommendations**

The biggest hurdle in this project was the sample size for bankrupt companies. Therefore, it is recommended to try another Python library, ‘EdgarTools,’ which may contain more data than the ‘yfinance’ library. It may even be best to visit the SEC website and download yearly or quarterly reports directly and store the data in a database which Python can then connect to for analysis. This would hopefully add to the number of data points and alleviate any overfitting seen in this study. Combination models have worked well in other studies and is another recommendation for further work to be done in this study. There are other variables which could be incorporated as well, such as social media sentiment on companies, company size, and how much foreign translation exposure a company has.

**References**

Abdelkader, N. A. M., & Wahba, H. H. (2024). A proposed multidimensional model for predicting financial distress: an empirical study on Egyptian listed firms. *Future Business Journal*, *10*(1), 1–16. <https://doi.org/10.1186/s43093-024-00328-2>

Altman, Edward. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance*, *23*(4), 589-609. [Altman1968.pdf (calctopia.com)](https://www.calctopia.com/papers/Altman1968.pdf)

Altman, Edward. (2000). Predicting Financial Distress Of Companies: Revisiting The Z-Score And Zeta. In A. Bell, C. Brooks, & M. Prokopczuk (Eds.), *Handbook of Research Methods and Applications in Empirical Finance*. Edward Elgar Publishing. 10.4337/9780857936097.00027.

Baxa, A., & Widdersheim, M. M. (2023). Inflation and U.S. Public Libraries: Three Approaches for Measuring Inflation in Historical Data. Journal of Library Administration, 63(3), 339–357. <https://doi.org/10.1080/01930826.2023.2177925>

Bhattacharjee, N., & De, A. (2022). Black Swan Event and The Stock Market Volatility Response to Shocks in Developed, Emerging, Frontier and the BRIC Markets: Lessons from the COVID-19 Pandemic. Brazilian Business Review (Portuguese Edition), 19(5), 492–507. <https://doi.org/10.15728/bbr.2022.19.5.2.en>

Bonett, D. G. (2020). Point‐biserial correlation: Interval estimation, hypothesis testing, meta‐analysis, and sample size determination. British Journal of Mathematical & Statistical Psychology, 73, 113–144. <https://doi.org/10.1111/bmsp.12189>

Divine, J. (2023, May 12). 7 Companies That Went Bankrupt Due to COVID. *U.S. News*. <https://money.usnews.com/investing/stock-market-news/slideshows/covid-bankrupt-companies>

Hassan, M., & Nassar, R. (2023). Effects of Financialand Macroeconomic Factors on the Financial Performance of a Company. *Journal of International Business Disciplines*, *18*(2), 19–29.

JMP. (n.d.). The t-Test. *JMP*. <https://www.jmp.com/en_us/statistics-knowledge-portal/t-test.html>

Kevin, J. (2024). *How to Use Yfinance for Generating this Data* [Jupyter Notebook]. Kaggle. <https://www.kaggle.com/code/junkevin/how-to-use-yfinance-for-generating-this-data/notebook>

Kristanti, F. T., & Dhaniswara, V. (2023). The Accuracy of Artificial Neural Networks and Logit Models in Predicting the Companies’ Financial Distress. *Journal of Technology Management & Innovation,* *18*(3), 42–49. <https://doi.org/10.4067/s0718-27242023000300042>

Li, J., Weng, J., Shao, C., & Guo, H. (2016). Cluster-Based Logistic Regression Model for Holiday Travel Mode Choice. *Procedia Engineering*, 137(2016), 729-737. <https://www.sciencedirect.com/science/article/pii/S1877705816003374?via%3Dihub>

McGregor, R. (2022, September 30). The 4 main reasons why you should care about financial ratios. *CPA*. <https://www.bccpa.ca/news-events/cpabc-newsroom/2022/september/the-4-main-reasons-why-you-should-care-about-financial-ratios/>

Mickle, C. F., & Deb, D. (2022). Early prediction of patient discharge disposition in acute neurological care using machine learning. BMC Health Services Research, 22(1), 1–15. <https://doi.org/10.1186/s12913-022-08615-w>

Sajja, R. (2023, November 22). Understanding and Analyzing “Correlation” in Python: A Practical Guide. *Medium*. <https://medium.com/@ratankumarsajja/understanding-and-analyzing-correlation-in-python-a-practical-guide-6ac485a066e7>

Schaefer, S. (2011, August 10). The Great Recession’s Biggest Bankruptcies: Where Are They Now? *Forbes*. <https://www.forbes.com/sites/steveschaefer/2011/08/10/the-great-recessions-biggest-bankruptcies-where-are-they-now/?sh=3e405f0b4b7e>

Simanca H., F. A., Barbosa Guerrero, L. M., & Palacios Rozo, J. J. (2022). Altman Z-Score Econometric Insolvency Prediction Model For Risk Coverage In Companies. *Webology*, *19*(6), 565–575.

Singh, P., Singh, S., & Paprzycki, M. (2023). Detection and elimination of multicollinearity in regression analysis. International Journal of Knowledge Based Intelligent Engineering Systems, 27(1), 105–111. <https://doi.org/10.3233/KES-221622>

Stobierski, T. (2021, June 8). What is GDP & Why is it Important? *Harvard Business School*. <https://online.hbs.edu/blog/post/why-is-gdp-important>

U.S. Bureau of Labor Statistics. (2024). *Consumer Price Index for All Urban Consumers (CPI-U)* [Data Set]. U.S. Bureau of Labor Statistics. <https://data.bls.gov/timeseries/CUUR0000SA0L1E?output_view=pct_12mths>

World Bank. (2024). *GDP growth (annual %) – United States* [Data Set]. World Bank. <https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?end=2022&locations=US&start=2012>

Yahoo!finance. (2024). Sectors. *Yahoo!finance*. <https://finance.yahoo.com/sectors>.