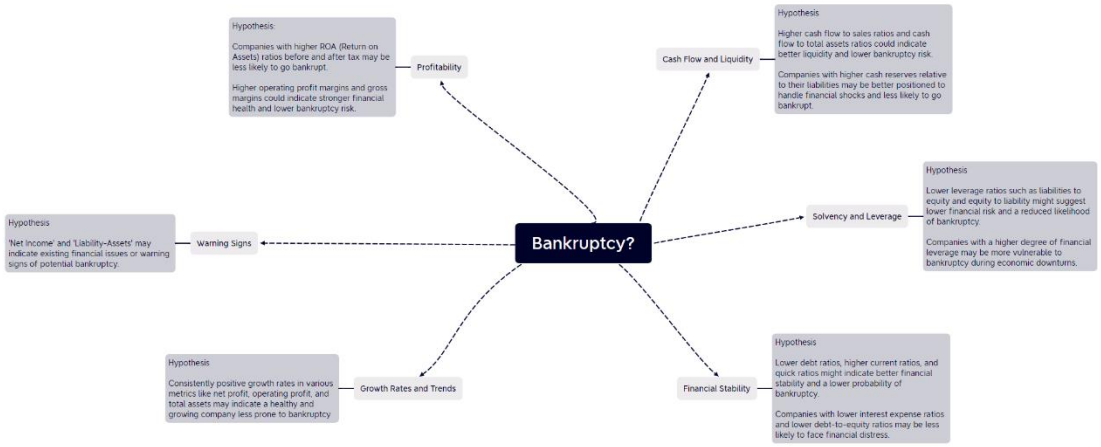
Introduction

In this Data Analytics report my focus is centred around predicting bankruptcy of companies in Taiwan. The reason I decided to do research around this topic is because I have a passion for investing and understanding company performance. When deciding whether to invest in certain companies there is a lot of due diligence required, this includes (just to name a few) looking at direction of the company, industry the company is in, current leadership and finally looking at company financials. This sparked a question that I never though about previously which is centred around understanding whether a company is going to go bankrupt or not. This to me seemed extremely important. A company from an outside perspective could be doing well but if the financials and the money in the background is not supporting it then it is in trouble. I then began reading up articles and understanding how important determining bankruptcy is for many other industries, not just investing. At this point I was sold and decided to go all in!

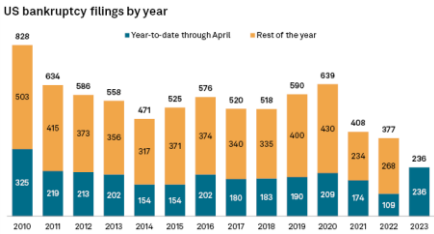
Hypothesis

The good thing about looking at company financial data is that there can be many hypotheses broken down into many different categories. Which is exactly what I did. Below is a mind map of the different hypothesis ideas I came up with and the different avenues I could explore.



Coming up with hypothesis was enjoyable, since there are 95 features in my dataset this allowed me to come up with hypothesis from all angles. For clarity I will take one from above and give a brief explanation. One of the first features that caught my eye when having a brief look at the data were 'Net Income' and 'Liability-Assets'. These are so crucial in understanding what direction the company is heading. Having a look at these columns could allow us to see if the company is having/had any financial issues which could lead to bankruptcy. If a company is experiencing continuous negative net income or continuous fluctuations and no stability this could likely lead to bankruptcy or was there a certain range to which when a company entered there was a high probability of bankruptcy.

Relevant Reading Material



Over the last couple years, the number of companies going bankrupt around the world has showed an upward trend. A brief breakdown for the US market is captured in Jack Kelly Forbes article where it outlines some of the potential reasons for bankruptcy being rising inflation, supply chain disruptions and interest rates. What I found more interesting was the market intelligence data from the S&P Global that was linked. This data showed that number of bankruptcies in the US alone for Year to date through April was the highest since the financial crash of 2008-09. Although the project I am doing does not focus on external economic factors, there is scope for further research and incorporating these external factors.

An article that I read was written by Talha Mahboob Alam, on Corporate Bankruptcy Prediction: An Approach Towards Better Corporate World. This was a fascinating article to read because it faced similar challenges that I was expecting to run into, for example, imbalanced data. Paraphrased, the study highlights important factors that can increase the risk of a company going bankrupt. These include consistently low profits even when they're putting more money into resources, big drops in how much they pay their employees, and not investing enough in things like training and hiring. Also, if they're too focused on getting more resources without making sure their operations run smoothly and their staff are well-trained, it can make their financial situation even worse and might lead to them having to shut down.

Reading ‘Failure processes and causes of company bankruptcy: a typology’ by Hubert Ooghe and Sofie De Prijcker was very intersting and showed a different approach to mine. The article starts of by outlining that many of the predictions models focus heavily on the financial side of the companies. Whereas what the authors are trying to incorporate is a balance of both looking at financials and considering characteristics of the company, they are trying to bridge that gap. The authors focus on 4 factors, these are around: company characteristics, management, environment, and characteristics of the entrepreneurs. Again, like the previous example I found it fascinating that through their findings, the characteristics of management/entrepreneur was deemed to be a major cause of bankruptcy. From the article I found out that they only focused on 12 Belgian companies of different industries, sizes, and ages. Although this is a good subset of companies, I felt like 12 companies was not enough and that the sample should have been bigger, to full test the hypothesis.

Work by Wenhao Zhang called Machine Learning Approaches to Predicting Company Bankruptcy. This article was similarly related to the project that I was looking to undertake. Like me, Wenhao Zhang explains that predicting value of companies is something that can ‘harm the entire system’ and ‘that the value of one company will always get better’ which is not always the case. What I found fascinating and something I did not think about is related to the employees of a company. Wenhao Zhang talks about predicting company bankruptcy rate and deciding whether people should quit the company early, rather than waiting till the company goes bankrupt and protect themselves. Which is something I never thought about but is a very valid point. The author also focuses on the simple machine learning algorithms, for example, K Nearest Neighbour and Random Forest. He further goes on to explain that the reason he does this is because he had read other research papers that use Neural Networks and they give skewed results and believes that sometimes the best approach is the simple algorithms, they perform better.

For this next article I took a different approach. I wanted to look at those bankruptcy prediction models that were already produced and being used. I did this through Irina Berzkalne and Elvira Zelgalve article called ‘Bankruptcy prediction models: a comparative study of the Baltic listed companies. The 7 models they focused on were Altman, Altman Z’, Altman Z’’, Springate, Fulmer,

Zmijewski and Šorins/Voronova. Before doing research, I was unaware that there was industry standard models in place. But after doing research and reading this article it became even more clear to me the significance of finding out company bankruptcy and how much of a major topic it is.

Obtaining the Dataset

When looking for suitable dataset I used Kaggle to locate appropriate data that would suit my requirements. When obtaining the data, I wanted to make sure that it was as genuine as possible, many datasets nowadays are manmade that don't have real number, which don't show the true outcome/projections of companies. I managed to come across Taiwan Economic Journal (TEJ), which when doing further research about the TEJ is known to be the most trusted financial site in Asia. For a bit of context about the data, the data was collected from the TEJ for the years 1999 to 2009. The decision on whether a company was bankrupt or not was based on the business regulations of the Taiwan Stock Exchange. In terms of the data itself, it had 6819 instances and 95 features.

Two points I would like to address about the data. Firstly, the reason that I picked companies from Taiwan was because I wanted to look at different markets, I am constantly looking at the companies within the UK and US. Having a deeper dive into different markets would allow me to get a better understanding of other markets and make comparisons between the different markets too. Secondly, I liked that this dataset had around 95 features. Since within companies there are so many financials that can have an impact on company bankruptcy, I wanted to see what features the bigger drives were and just for my own interest how companies on the Taiwan Stock Exchange were performing.

Sample of the Dataset

Below is a sample of the dataset that was imported through Kaggle. Since there are 96 columns in my data showing the full data would be very difficult. However, within my notebook you can see the full dataset and more!

	Bankrupt?	ROA(C) before interest and depreciation before interest	ROA(A) before interest and % after tax	ROA(B) before interest and depreciation after tax	Operating Gross Margin	Realized Sales Gross Margin	Operating Profit Rate	Pre-tax net Interest Rate	After-tax net Interest Rate	Non-industry income and expenditure/revenue	Continuous interest rate (after tax)	Operating Expense Rate	Research and development expense rate	Cash flow rate	Interest-bearing debt interest rate	Tax rate (A)	Net Value Per Share (B)	Net Value Per Share (A)	Net Value Per Share (C)	Persistent EPS in the Last Four Seasons	Cash Flow Per Share	Revenue Per Share (Yuan ¥)	Operating Profit Per Share (Yuan ¥)
0	1	0.370594	0.424389	0.405750	0.601457	0.601457	0.998969	0.796887	0.808809	0.302646	0.780985	1.256969e-04	0.000000e+00	0.458143	7.250725e-04	0.000000	0.147950	0.147950	0.147950	0.169141	0.311664	0.017560	0.095921
1	1	0.464291	0.538214	0.516730	0.610235	0.610235	0.998946	0.797380	0.809301	0.303556	0.781506	2.897851e-04	0.000000e+00	0.461867	6.470647e-04	0.000000	0.182251	0.182251	0.182251	0.208944	0.318137	0.021144	0.093722
2	1	0.426071	0.499019	0.472295	0.601450	0.601364	0.998857	0.796403	0.808388	0.302035	0.780284	2.361297e-04	2.550000e+07	0.458521	7.900790e-04	0.000000	0.177911	0.177911	0.193713	0.180581	0.307102	0.005944	0.092338
3	1	0.399844	0.451265	0.457733	0.583541	0.583541	0.998700	0.796967	0.808966	0.303350	0.781241	1.078888e-04	0.000000e+00	0.465705	4.490449e-04	0.000000	0.154187	0.154187	0.154187	0.193722	0.321674	0.014368	0.077762
4	1	0.465022	0.538432	0.522298	0.598783	0.598783	0.998973	0.797366	0.809304	0.303475	0.781550	7.890000e+09	0.000000e+00	0.462746	6.860686e-04	0.000000	0.167502	0.167502	0.167502	0.212537	0.319162	0.029690	0.096898
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6814	0	0.493687	0.539468	0.543230	0.604455	0.604462	0.998992	0.797409	0.809331	0.303510	0.781588	1.510213e-04	4.500000e+09	0.463734	1.790179e-04	0.113372	0.175045	0.175045	0.175045	0.216602	0.320966	0.020766	0.098200
6815	0	0.475162	0.538269	0.524172	0.598308	0.598308	0.998992	0.797414	0.809327	0.303520	0.781586	5.220000e+09	1.440000e+09	0.461978	2.370237e-04	0.371596	0.181324	0.181324	0.181324	0.216697	0.318278	0.023050	0.098608
6816	0	0.472725	0.533744	0.520638	0.610444	0.610213	0.998984	0.797401	0.809317	0.303512	0.781546	2.509312e-04	1.039086e-04	0.472189	0.000000e+00	0.490839	0.269521	0.269521	0.269521	0.210929	0.324857	0.044255	0.100073
6817	0	0.506264	0.559911	0.554045	0.607850	0.607850	0.999074	0.797500	0.809399	0.303498	0.781663	1.236154e-04	2.510000e+09	0.476123	2.110211e-04	0.181294	0.213392	0.213392	0.213392	0.228326	0.346573	0.031535	0.111799
6818	0	0.493053	0.570105	0.549548	0.627409	0.627409	0.998080	0.801987	0.813800	0.313415	0.786079	1.431695e-03	0.000000e+00	0.427721	5.900000e+08	0.000000	0.220766	0.220766	0.220766	0.227758	0.305793	0.000665	0.092501

6819 rows × 96 columns

Data Preparation/Data Exploration

At first glance the data looked like it had already been processed and was in a good format. But to make sure that this was the case I did some final checks on the data.

One of the first checks I carried out was checking what datatypes all the columns where in my data. This is an important step because when inputting data into my machine learning models it must be in numerical datatypes. The check came back and showed me that columns in my data were either 'int64' or 'float64' which was good news.

The next step for me was to check for whether there was any null values/missing data. I ran .info() and .describe() on my data to get a better understanding of my data. This led me to some interesting conclusions about my data. The first spot was that there were no null values and that all my columns had 6819 values, which was good. I did a finally check by running some code to double check this

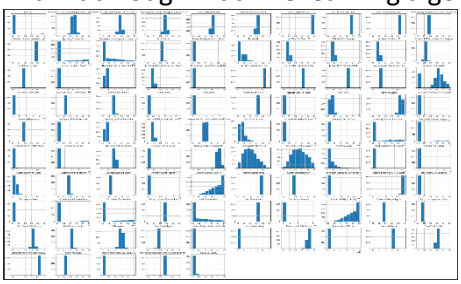
and both showed no null/missing values. Also, I had a look to see if there were any duplicate values, which there were not.

The second spot was that I noticed that my min and max of all features in my data were 0 and 1. This meant that there was a min max scaling normalisation applied to the data already. Although this had a min max scaling thinking about future implementations, like PCA, I know that I will have to apply a standard scaling where the features have mean of 0 and variance of 1.

I wanted to look at variable that I would be predicting and look at the samples that I had and what the percentage split was between the samples. I found out that 6599/6819 samples were for not bankrupt, and 220/6819 samples were for bankruptcy. This was a big concern of mine as this would skew the machine learning models into predicting 96% of the time that the company would not go bankrupt. There were 3 keyways I found that could have solved these problems. These were downsampling, upsampling and SMOTE. I decided that I would go with the SMOTE technique which essentially uses the current data sample to generate similar data. The reason I did not use the other two techniques was because with downsampling, I would be removing over 6000 samples from the sample where the companies did not go bankrupt, which would make the dataset have 440 samples, which is small. With the upsampling, that focuses on duplicating data that is already in the sample of companies that have gone bankrupt, and I felt like doing this would have no variety in the samples, which could skew the machine learning model outputs. Which is why I settled for SMOTE.

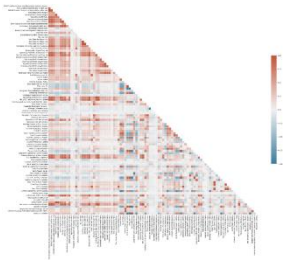
Feature Selection

As I have mentioned above there are around 95 features in my data. This is too much. I want to try and reduce the number of features in my data. The first thing I did was check to see if there were any columns in my data that had one unique value. From my analysis I found out that 'Net Income Flag' had 1 unique value. This meant that regardless of whether a company was bankrupt or not the value remained unchanged. This added no differentiation, which meant that it would have no impact when I ran it through machine learning algorithms, so I dropped the column. This still left another 93 features which I had to reduce. Going through these one by one did not seem feasible and when charting these variables through histograms and heatmaps there was some interesting finds, but the real bulk of reducing the number of features came through PCA.



For further clarification around the distribution of my features I decided to produce a histogram. Just having a look at the diagram without knowing what features it is clear to see that a lot of the features are skewed towards one side or the other. There are only a handful features that have variation.

Having a look at a heatmap that has 94 features is not the easiest. But here I just wanted to get my point across that there were many features which were highly correlated and that could be dropped.



I wanted to make sure that there was some form of reason of reducing the features before I applied PCA, which is why I plotted the histogram and the heatmap. I then ran PCA on my data and created a chart that showed me that 99% of my data was being explained from around 9 features in my data. This was great as it meant that the number of features dropped from 94 to 9.

Logistic Regression

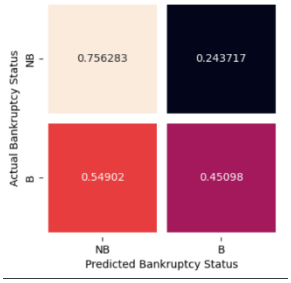
Logistic Regression is a supervised learning algorithm that focuses on binary classification. It is a type of regression analysis where the dependent variable is categorical. The Logistic Regressions model is used to predict the probability that a given input sample belongs to a certain category. The focus of Logistic Regression is to split the data through producing a line, it does this by relying on the sigmoid function.

The reason why I chose to use Logistic Regression as part of machine learning algorithms is because my data is about categorising whether a given company is bankrupt or not, which means that it is categorical. Using the features that I have I can create a split in the data that separates those companies that are bankrupt from those that aren't.

Results

I ran two different logistic regression models. The first model was the standard model with the SMOTE technique applied. I did run the logistic regression on the data as it came but because of the imbalanced data the training accuracy result came in at 96%, which was expected because I had around 96% of the data attributed to companies that had not gone bankrupt. Figure 1 is the confusion matrix for the data with the smote technique applied. The accuracy of this model was 60.7%. I also used PCA and hyperparameters to increase the accuracy of the model, which is the second model I ran.

Fig 1 – Standard (SMOTE Technique applied)



From running PCA I worked out that 9 components explained over 99% (Fig 3) of my data and I went with an inverse regularisation strength of 1.2. This increased my accuracy to 61.8%. Although these tweaks have increased the accuracy, I was surprised, since it was a smaller increase then I expected.

Fig 2 – Standard (SMOTE Technique applied) + PCA.

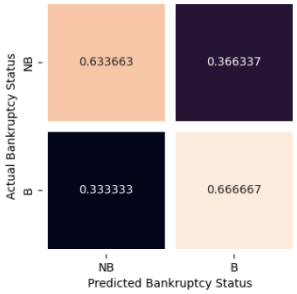


Fig 4 – Hyperparameter Regularisation Graph

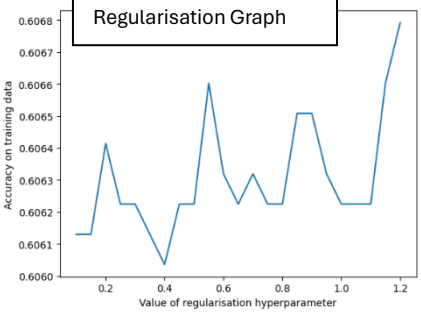
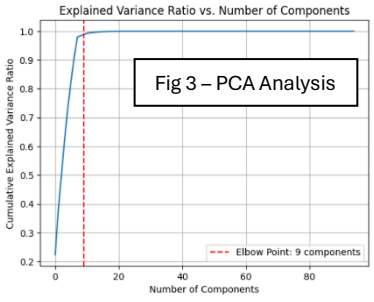


Fig 3 – PCA Analysis



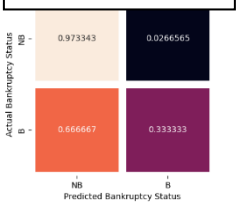
Binary Decision Trees

When researching which algorithms to pick for my data I came across Binary Decision Trees. Apon researching further some of the key benefits I found out that binary decision trees are good for datasets that have unbalanced data and help to identify critical factors in the data. This was music to my ears because as stated above the imbalance of data in my dataset was extreme and I also had over 90 features. This felt like the best algorithm for the job. I also thought that this would give me good opportunity to test whether having an imbalanced data or using the smote technique on my data then running it through the binary decision tree algorithm had any difference.

Results

The first iteration I ran was on the dataset as it came in, with the massive imbalance. Fig 5 shows the result. Comparing this model to the Logistic Regression it did considerably better coming in with an accuracy of 95.5%. However, when looking at the results I noticed that the accuracy on the training data was 1.0. This means that it always predicts the correct label. This is a cause of concern since it can have negative implications on the test data.

Fig 5 – First Iteration



The second iteration I ran was on the data that I had, but this time it was balanced using the SMOTE technique before ran through the decision tree algorithm. Fig 7 shows the confusion matrix. Previous iteration accuracy was 95.5%, this time I got an accuracy of 95.9%. This is a very slight change and shows that the decision tree algorithm balanced out the imbalance in my data very well. Again, like the previous iteration I spotted that the accuracy of the training data was 1.0. I did a final iteration where I used PCA + SMOTE and this can be seen Fig 9 and Fig 10. The accuracy of the model dropped to 89.9%

Fig 7 – Second Iteration

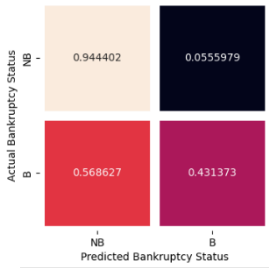


Fig 8 – Second Iteration Decision Tree generate

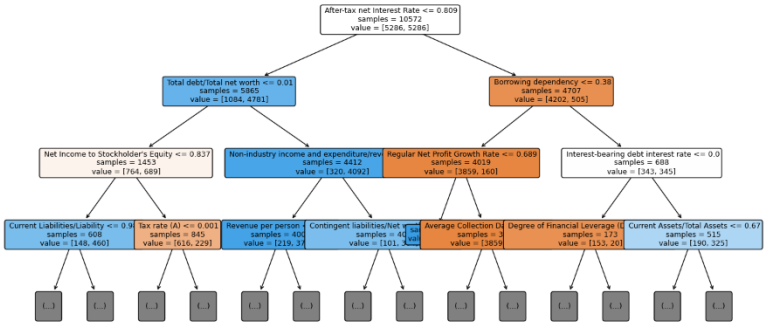


Fig 9 – Third Iteration

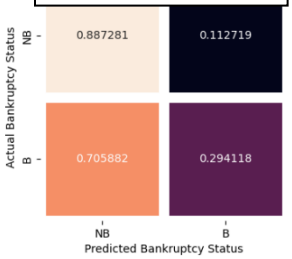
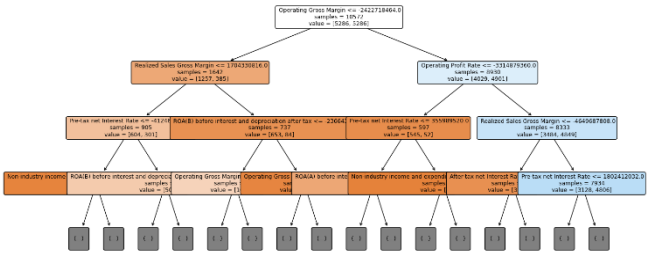


Fig 10 – Third Iteration Decision Tree generated.



Challenges

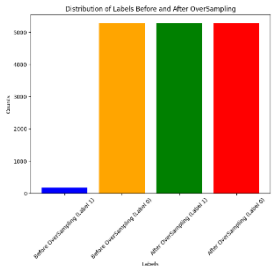


Fig 11

The characteristics of the dataset present challenges in model development. Research indicates that most companies do not go bankrupt, leading to a substantial bias within the datasets. This was clear to see when I found out that around 96% of my data consisted of those companies that were not bankrupt.

One of the main reasons I picked this dataset was because of the two challenges I saw upon investigating the data at a first glance. The two challenges were imbalance of data and the number of features. Fig 11 shows the clear imbalance of companies that did not go bankrupt vs companies that did go bankrupt before and after I applied SMOTE.

The second challenge was that I had around 95 features and I had to try and narrow this down. I felt like this would be an interesting/difficult task because when evaluating companies there are so many internal and external factors which can lead to company going bankrupt. I thought it would be interesting to go on this journey to see what factors had the biggest influence.

Key Findings

I found some very interesting findings, some that aligned with my initial hypothesis and some that surprised me. Having a look at the Decision Tree before any form of PCA showed that the 'After-tax net Interest Rate', 'Total debt/Total net worth' and 'Borrowing dependency' were three of the biggest factors in determining company bankruptcy. Thinking about this logically, this makes sense; usually those companies that have high amount to debts relative to other factors go on to become bankrupt. Linked with this is the factors of 'Borrowing dependency', again, logically this would make sense. Companies that borrow a lot of money will incur an interest rate which would have to be paid back on top of the principal which could cause financial difficulty to the company leading them to become bankrupt.

One of my initial hypotheses was centred around believing that 'Net Income' was a big driver in understanding whether a company goes bankrupt or not. Simply put, I was wrong. It played no role in determining whether a company went bankrupt, which I found extremely fascinating.

Possible Business Applications

After running the models on the processed data and seeing the output of the models I don't think that the model is anywhere near ready for professional level use, however, it provides a good base model. The reason I say this is because the models outlined some of the key features that went into determining the bankruptcy, but looking at financial of companies is only a small section in determining whether a company may go bankrupt or not, there are several other factors. The models produced an accuracy of around ~60%, which doesn't give me enough confidence for professional use.

Conclusions

To conclude, this report details creating models and predicting whether a company will be bankrupt or not depending on financial features in Taiwan. Although this report focused more on the financial factors, throughout this project with the extended reading I found out that financial factors is just one section out of many factors that can go into determining whether a company goes bankrupt or not. For example, to improve the model I could add economic conditions in Taiwan during the modelling period. From my reading I found out that economic conditions play a massive role in company survival.

One final point out would outline is that the models that I ran seemed to get similar accuracy on the test data, which was good, as it meant there was no underfitting/overfitting. With test accuracy of 74.5% and 63.5% on the logistic regressions and accuracy of 92.8% and 87.1% on decision tree.

Final point to note is that I started this report with some initial hypothesis. Now coming to the end of the report some of those hypotheses have become clearer, but not being fully answered, I believe that is because like I stated above there are some variables that I did not include.

References

Hubert Ooghe, S. D. (2008). Failure processes and causes of company bankruptcy: a typology. *Emerald Insight*.

Kelly, J. (2023). Corporate Bankruptcies Are Rising At A Concerning Rate—What To Do If Your Company Has Filed For Bankruptcy. *Forbes*.

Sabater, S. L. (2023). April adds 54 more US corporate bankruptcies; 2023 filings highest since 2010. *S&P Global*.

Zhang, W. (2017). Machine Learning Approaches to Predicting Company Bankruptcy. *Scientific Research*.

Zhang, W. (2017). Machine Learning Approaches to Predicting Company Bankruptcy. *Scientific Research*.