Project name: Bankruptcy Prediction using Python

Prepared by Team 7: Harsha Singh, Ranjana Jalali, Sourya Peddina,
Xiasun (Yuri) Huang, Ganesh Babu Bamdhamravuri

Date: APRIL 10, 2022

Purpose of This Report

Bankruptcy is disastrous for companies and will be the end of the game if it happens. Fortunately, bankruptcy prediction theory is not new for people, and many financial professionals are working on modeling to provide better guidance for companies facing such problems. In the late 1960s, Edward Altman researched the extent to which analysis of different financial ratios could be used to predict business failure and bankruptcy (Altman, p565, ACCA). In his Z score model, he identified five key indicators of the likely failure or non-failure of business:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5$$

, where X1 measures liquidity, X2 measures profitability, X3 measures activity/efficiency, X4 measures leverage, and X5 measures solvency.

This project is inspired by the Z score model of Altman. We do not assign the same score as Altman's do but collect data in the same indicators (categories in our case), with which we train predictive models to advise businesses on how to tackle the problem by taking advantage of financial data.

To summarize our procedures, first, we determine the most significant categories based on their exploratory information contribution to our target variable. Second, we train several predictive models to evaluate the classification performance in terms of precision and recall. Third, we discuss our best model on how businesses can use it and its limitations of using financial data.

Our most appropriate model for businesses is the KNN model, which can be used as an intuitive tool to detect bankruptcy. We recommend the high precision feature of the model and suggest business leaders should be cautious and set contingent margins based on their risk appetite. By the end of the report, our team also discusses how business leaders can consider other factors to understand the root cause of business failures. We point out that we use bankruptcy and business failures interchangeably throughout the report.

An Introduction to Our Dataset

The dataset was taken from Kaggle that was collected from the Taiwan Economic Journal. The dataset contains 96 columns and 6819 rows with the details of company bankruptcy based on the business regulations of the Taiwan Stock Exchange. Most of the data have numerical attributes that help evaluate the possibility of bankruptcy.

We summarize the 96 variables into Altman's categories to have an overall view of the variable diversity. Figure 2.1 shows that we have not only comprehensive Altman categories in our dataset but also an additional category that evaluates shareholder values.

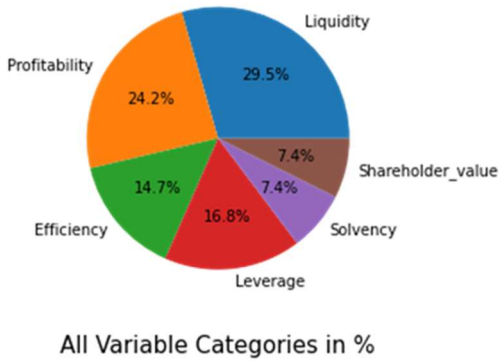


Figure 2.1

During our data preprocessing stage, we did not identify any missing values. To prepare for the modeling, we standardized all the independent numeric variables using python StandardScaler. We also set our training data proportion to 40%.

To reduce the number of variables, we look into the correlation matrix of all variables and remove variables from the highly correlated pairs (abs $\geq$ 0.3). Appendix I visualizes all correlations among all variables.

To select the top variables for our model, we do feature selection based on the explanatory power between the variable and the target variable. Figure 2.2 shows there are five variables at the first tier of exploratory power. These are the key variables of our model, and we summarize their categories into a pie chart to see if the categories are diversified in the light of Altman's Z score model. Figure 2.3 tells that our key variables have four Altman's categories. We are confident that these variables will provide comprehensive information for the prediction.

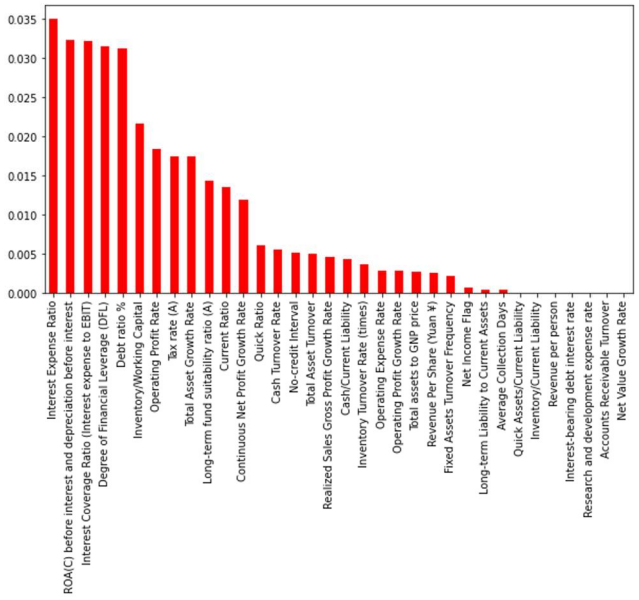


Figure 2.2 Mutual Values

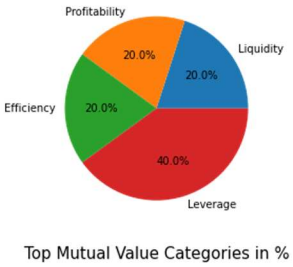


Figure 2.3

Our Modeling Process

We have 5 models on our basket (KNN, Perceptron, Logistic Regression, Naive Bayesian, and Decision Tree), from which we will select the best one to advise businesses. In the testing data set, we evaluate the performance of all the models in two dimensions, precision, and recall. Figure 3.1 and Figure 3.2 show the rankings of the models in terms of the two performance indicators. We can see that KNN ranks the first in terms of precision. Perceptron performs the best to contribute to recall.

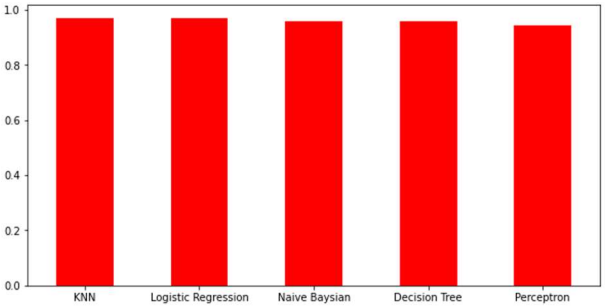


Figure 3.1 Precision Scores

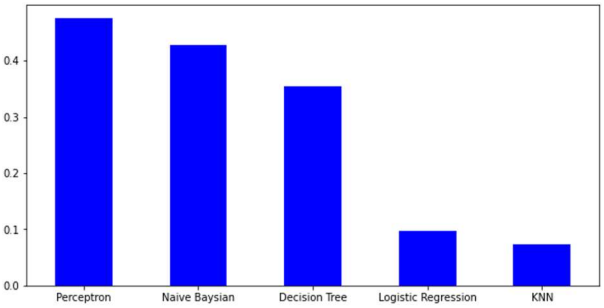


Figure 3.2 Recall Scores

Our Model Interpretation

In terms of precision, KNN turns out to be the best classification model (Figure 3.1), in which data points are grouped into 14 neighbors based on their Euclidean distance between each other. Using such a non-parametric supervised learning method, we achieve a very impressive overall precision score, of 97% (Figure 3.1). When it comes to recalling, Perceptron is at the top of the ranking, with a recall score of 0.48. We check the confusion matrix of the two models to further evaluate the performance in the classification.

1. Confusion Matrix & Classification Report

The confusion matrix in Figure 4.1 and Figure 4.2 evaluates the classification performance in two dimensions. When evaluating the model's performance in dividing companies with low risks of bankruptcy, we can see that KNN has 2,642 successful instances in the data pool of 2,718 records in total. This means that the model has impressive accuracy in telling if a company is healthy. When it comes to classifying bankruptcy, KNN has successfully 6 attempts in 10 records of such events. In a much riskier environment, the model fails to escalate the performance in dividing data into distinctive groups, but the accuracy rate (60%) in this case is satisfying for the team as it is better than that in the case should we randomly tell (50%).

The classification report of KNN helps summarize the discussion above by listing the precision scores for each case, whether classifying 0 (97%) or 1 (60%). Statistically speaking, decision-making based on merely precision is so biased that it ignores other needs of the model user who may be cautious about other key performance indicators. In our case, companies may care more about whether it's risky of going into bankruptcy than whether it's safe, as the consequences of bankruptcy will be "an obituary" on tomorrow's latest headlines.

When looking into the recall of this model, we can see that KNN only has a score of 7% in proportion to the total occurrences of bankruptcy. This can be disastrous for companies, as bankruptcy tends to be misclassified and the resulting consequences can be "a bloody hell".

```
[[2642    4]
 [  76    6]]
```

	precision	recall	f1-score	support
0	0.97	1.00	0.99	2646
1	0.60	0.07	0.13	82
accuracy			0.97	2728
macro avg	0.79	0.54	0.56	2728
weighted avg	0.96	0.97	0.96	2728

Figure 4.1

It's arguable that Perceptron should be the best one to go, given the highest recall score. Figure 4.2 shows that despite the lower overall accuracy (94%) than KNN, Perceptron penalizes heavily on bankruptcy misclassification and successfully identifies 48% of the total bankruptcy records. However, the much lower precision rate (27%) will mean a lot of false alarms. In a real business environment, the low precision rate will make Perceptron unreliable for people. Once the predicted results come out, businesses may not be cautious since Perceptron "lies" a lot in this case.

```
[[2538  108]
 [  43   39]]
```

	precision	recall	f1-score	support
0	0.98	0.96	0.97	2646
1	0.27	0.48	0.34	82
accuracy			0.94	2728
macro avg	0.62	0.72	0.66	2728
weighted avg	0.96	0.94	0.95	2728

Figure 4.2

2. Coefficients of variables

The KNN model has another disadvantage it does not give the coefficients of key variables. Therefore, it's not possible to see the level of sensitivity of our target variable to each of the key variables. To see how the key variables will explain the target one, we migrate the list of coefficients from the logistic regression model we used for the model comparison.

Table 4.1 shows that ROA and Debt Ratio have significant coefficients with an absolute value of around 1. Using Altman's categories in his Z score model, we can also see that the classification of bankruptcy becomes very sensitive to the company's efficiency and leverage levels.

Variable	Coefficient	Category
ROA (Return on Assets)	-0.943414	Efficiency
Interest Expense Ratio	0.008479	Profitability
Debt Ratio %	0.987256	Leverage
DFL (Degree of Financial Leverage)	0.013022	Leverage
Interest Coverage Ratio	0.021385	Liquidity

Table 4.1

Another way to think about the coefficient of key variables will be a visualization of the relationship between each key variable and the target variable. For this, our team employs Mosaic plots to better understand how key variables will have an impact.

Figure 4.3 – Figure 4.7 provides an intuitive graph to explore the pattern of each pair of key variable – target. We can see that ROA and Debt Ratio have a very seeable relationship with the target variable. The lower the ROA, the higher the chance of the bankruptcy of the company. The higher the debt ratio the company has, the higher chance of the bankruptcy for sure. For the other key variables, most of the companies have the same level of financial parameters. In very extreme cases, we do not have sufficient data to explore the relationship between the key variable and the target.

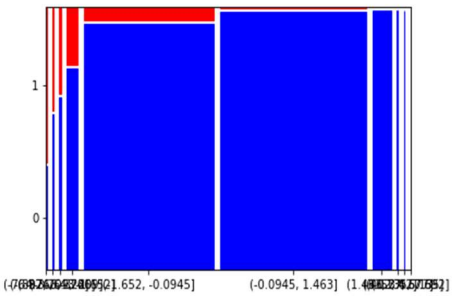


Figure 4.3

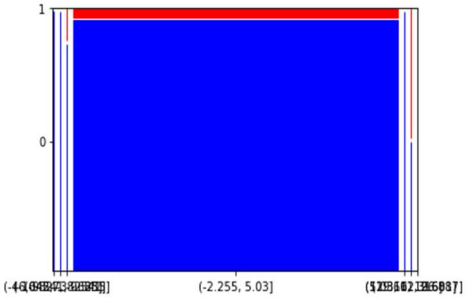


Figure 4.4

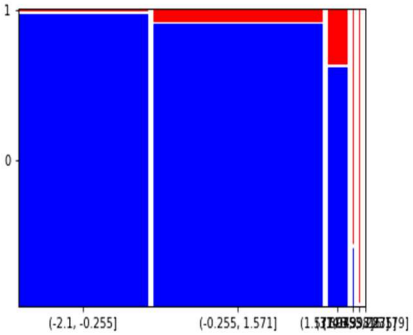


Figure 4.5

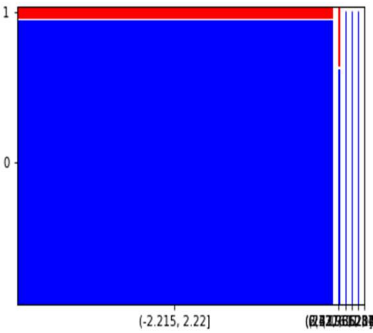


Figure 4.6

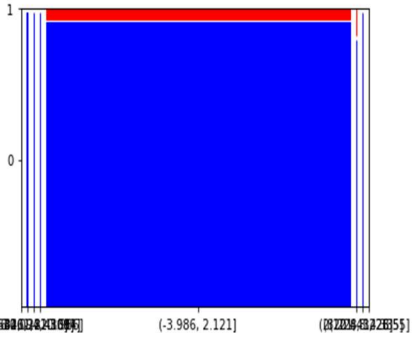


Figure 4.7

The visualizations are in line with what we can tell from the key variable coefficients (Table 4.1) and prove that efficiency and leverage are the two significant variables that have an obvious relationship with our target variable.

3. A “Formula” in an Image

Despite the weakness in writing down the prediction formula for the use, KNN excels at marking distinctive regions on the dataset. The regions can be used as an intuitive classification tool for businesses, which can use KNN as a “map” to identify risk areas. The compass, in this case, will be our key variables, or mainly ROA and Debt Ratio. Using python, we can draw the landscape based on the financial parameters given, while determining an appropriate size of the map. For more details on the procedure of “map drawing”, please refer to the enclosed Colab notebook (Appendix II).

Figure 4.8 is the product of the KNN model which can be used as an intuitive tool for companies “navigating” every day in the business world. Metaphorically, the “pink ocean” area is marked as the safety area, whilst the “green mainland and its surrounding isles” are marked as the risky area. On this map, companies can determine the coordinates of their financial condition, and change the “sailing route” before the “ship” is stranded on the shore. We call this map “the Greenland Model”.

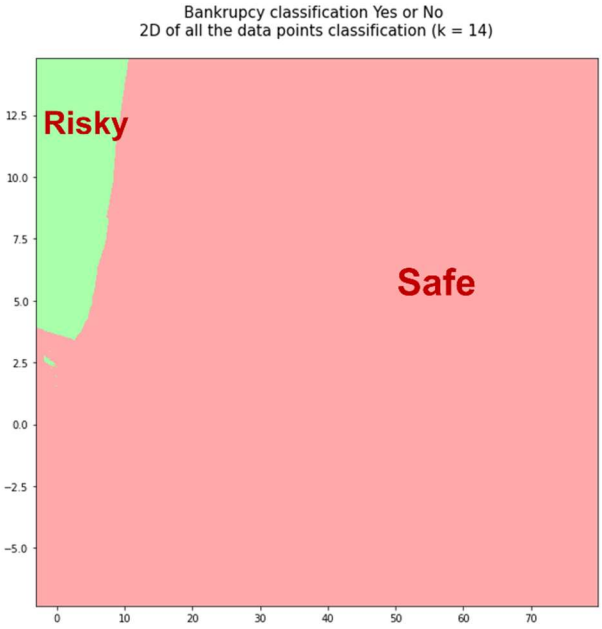


Figure 4.8

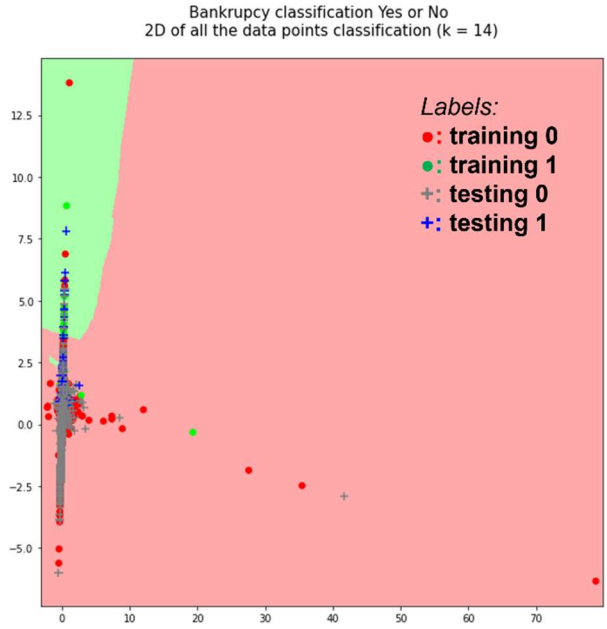


Figure 4.9

To see how our data points will spread across the map, we translate the five-dimensional dataset into a two-dimensional piece, using the Python package, PCA. Figure 4.9 shows that many bankruptcy records from the training and testing dataset fall on the "mainland" and its surrounding small islands. However, some data points indicating bankruptcy also escape the detecting system, clustering with the data points of the other classification or floating somewhere in the deep "pink ocean." Also, blue crosses and green dots are not the only "visitors" on the "mainland," meaning some of the data points predicting 0s are misclassified as part of the "mainlanders." Misclassification happens a lot.

Apart from visualizing the classification performance of the KNN model, the "Greenland Model" provides a two-dimensional bird's view of the spread of all data points. Data points are spread vertically or follow some linear line into the deep "pink ocean" area at the lower right corner. Our guess is that there are two powerful forces (X-axis: ROA and Y-axis: Debt Ratio) that have an obvious relationship with our target. Most of the data points follow the two directions guided by the two powers. This makes sense because companies with low efficiency and high debt rate have a much high chance of bankruptcy. It's not difficult to tell that the low-efficiency-and-high-debt-rate area is the "mainland" on the map.

4. The Failure of Quantified Financial Model?

The KNN stands out with an accuracy rate of 97%, but this overall precision score can't tell the model can do better than the case should we arbitrarily say all companies are not going into bankruptcy (all the predicted are 0s). The proportion of 0s in the data set is so high (97%) that even if we take a naïve random selection strategy, we still have satisfying accuracy! From this perspective (compared with a

naïve random selection), our KNN model fails to identify the distinctive features of the companies going into bankruptcy in terms of financial parameters. When we look at the key variable visualization, we can see that most companies share the same level of the financial parameter for some variables (Figure 4.4, 4.6, 4.7). For the variables that have an obvious relationship (Figure 4.3 & Figure 4.5), bankrupted companies' features are only distinctive in very extreme cases. Yes, identifying ROA and Debt Ratio as the overwhelming powers is a great step, but the records for such features (risky ROA and risky Debt Ratio) are still limited. When it comes to the Perceptron model, the large number of false alarms will make the model much less trustworthy and the usability questioned.

In his article, Risks and Crises, Danielsson questions the prominent use of financial models as they give wrong signals and wrong risk underestimation very frequently (Danielsson, 2011). Apart from the financial data itself (like lack of predicting features in our case), Danielsson also criticizes the vast use of historical data purported for future prediction. Even though we test the model before it can be used, the testing dataset itself is passive, and the data is not likely to have a say in a future event.

In his A-score model for company failure prediction, Argenti uses judgment to assign scores to each problem area (Argenti, p566, ACCA). Instead of looking into only financial data, he argues companies should focus on the symptoms, which will only become transparent in the later stages of failure. He believes the main cause of business failure lies in the management's ability to lead a business, and such ability is evaluated in the following categories.

- Defects: The company shows defects in its management style: for example, autocratic CEO, a passive board, lack of budgetary control.
- Mistakes: The company starts to make mistakes: making wrong decisions, like overtrading.
- Symptoms: The company becomes “ill”, for example, declining morale in its personnel and declining quality.

Acknowledging the limitations of our financial model and the importance of leadership, our team believes that the interpretation of our model matters in real business practice. KNN, in our opinion, is the best model that can be used in a compromised way to address the concerns aforementioned.

Recommendations and Suggestions

Despite the low recall score, the model provides an intuitive "map" with which business leaders can see a clear direction of the "navigation" and avoid crashes to the shore of "the mainland". To address the concerns of low recall scores, business leaders may incorporate their level of risk appetite into the map. For example, a margin of "shallow waters" can be added (mainly to the south) to the border of the green "mainland" (so that more blue crosses can be monitored!). In this way, KNN will alarm the map users that things can go wrong if the financial data point of the company falls into this "shallow water." A business leader can then take a U-turn to make sure the "ship" is always in the safety zone. However, the add-ons of the "shallow waters" have limitations very similar to the Perceptron model, making a lot of false alarms. The business leader's tolerance of these false alarms depends on their risk appetite. If the risk appetite of the leader is low, then the margin should be wide (because more blue crosses can be detected).

The 60% precision score for detecting bankruptcy on "the mainland" can definitely be a powerful tool for leaders. The green territory should be used as a benchmark so that leaders can understand why they should be concerned about their company's financial conditions and how concerning the situation is.

Business leaders should not limit their responsibilities to detecting how the financial data can go wrong. Understanding the root causes of the problem and taking corrective actions are also important for the company's survival. Argenti's three categories provide a good scorecard that will help evaluate the company's underlying management issues.

Conclusion

Our team has processed a dataset with around 7K records and 100 columns using Python and selected our models' top five features. This project follows statistical procedures before the project team can find valuable indicators for bankruptcy prediction. We evaluate our models in terms of precision and recall and determine that KNN has the best precision, and the Perceptron model has the highest recall core. This report emphasizes the model interpretation and the model application in the real business world. When evaluating the impacts of high precision and low recall, we suggest that businesses should use KNN as an intuitive tool for day-to-day financial condition control. Our team also recommends setting risk appetite to compensate for the shortcomings of our best model, which can be used as a benchmark for risk evaluation. In addition to monitoring and controlling financial data, our team also suggests an investigation of the root causes of the poor financial conditions and proactive actions to improve by considering qualitative factors using Argenti's scorecard.

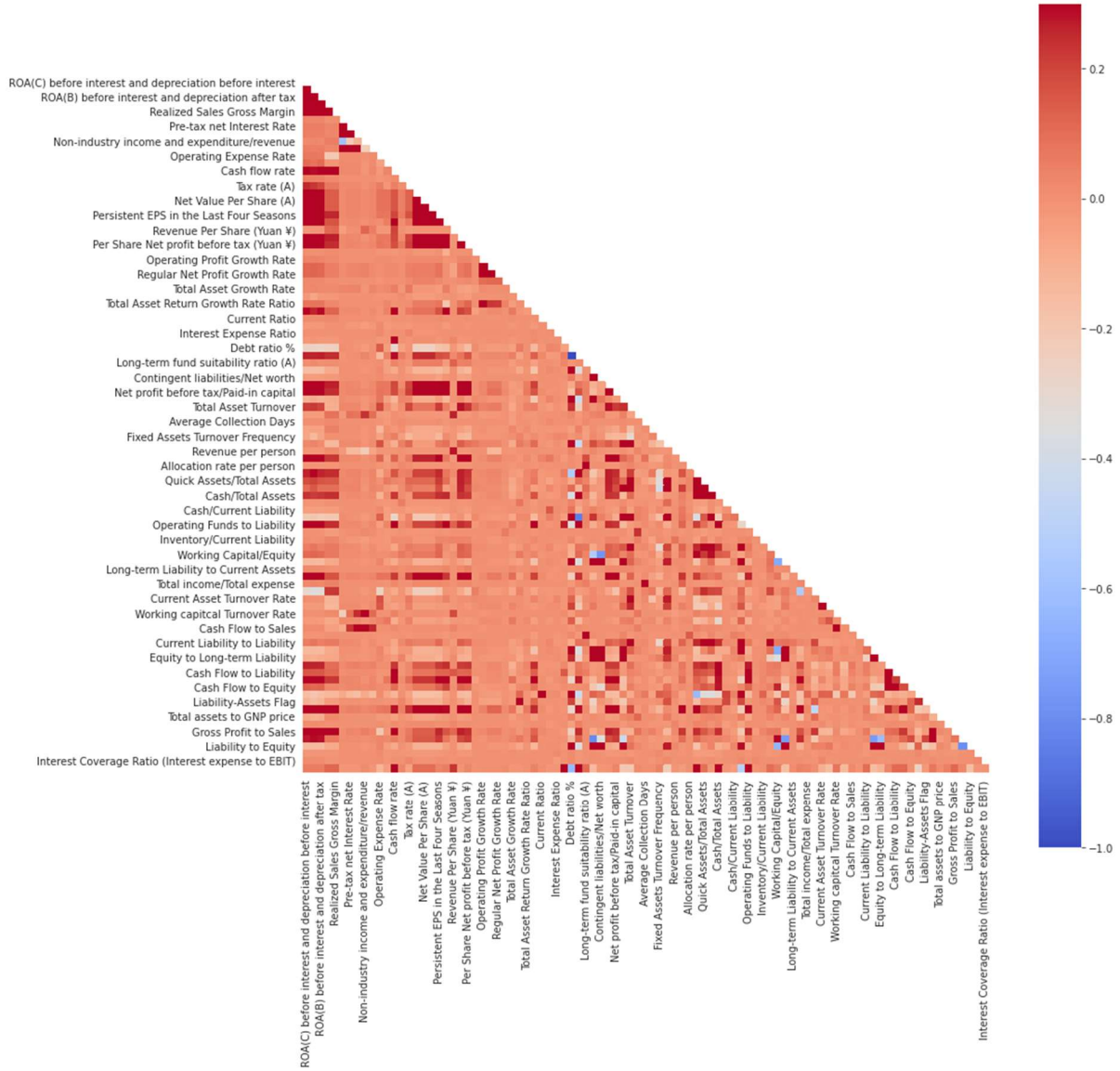
References

The Association of the Chartered Certified Accountants. Advanced Performance Management. BPP Media, 8th Edition, 2015, p565, p566.

Jon Danielsson, 2011. Risk and Crises. VOXEU. https://voxeu.org/article/risk-and-crises-how-models-failed-and-are-failing

Appendix

I. All variable correlation matrix



II. Procedures and Scripts

Please refer to our Colab:

https://colab.research.google.com/drive/1ZPJYC97foyZ0atunUf-jouc1e8851ut0?usp=sharing