ISyE 6740 – Summer 2021
Final Report

Team Member Names: Po-Chou, Lin (gtID: 903621638)

Project Title: Predict corporate bankruptcy using financial ratios and other quantitative factors.

1. Abstract

Corporate Credit Risks are spiking in recent years. Companies are aggressive in funding, expanding, and developing new business thanks to the cheap interest rates in the Quantitative easing era. The Fed Reserve lowers the fed fund rate to stimulate economic activities and investments. However, banks are ignoring the company default risks due to the flooding of hot money.

Banks make profits by absorbing people's cheaper deposits and lending them out at higher rates. When the interest rate benchmark drops, the interest difference is getting narrower, which hurts banks' profitability consequently. Banks would lend out more loans to maintain profits growth rates at the cost of spiking credit risks. Among banks' businesses, corporate financing contributes greatly to the revenue. However, banks are losing control in managing the risks of corporate bankruptcy due to the rapid expansion and lower threshold for loan approvals.

If we want to use well-developed credit models to evaluate corporate credit risks, we will find most credit models are developed by financial experts before the 2009 financial crisis. Those old models are mainly developed for traditionally manufacturing models in the 20 century and only use simple ratios to linearly score the credit risks. They used to work well during that time but could be outdated in the current economic environments.

This project presents how we develop a better model for predicting corporate bankruptcy. In the first place, we explain our methods of exploring, cleaning, and preprocessing the bankruptcy data. Secondly, we apply a variety of classification models to test their performance to predict bankruptcy. Finally, we develop our own scoring model based on the Elastic Net Algorithm to achieve better performance and even require fewer variable inputs, which will save huge costs in data collections and operation efficiency. It turns out it only requires 21 input variables to achieve the 95.3% approval rate and 1.08% default rate for the bank's loan portfolios.

2. Problem Statement

The financial systems are losing the balance between returns and risks. Banks could lose 100% principles in a default case, but only earn annual 3% interest income from the balance of their issued loans. Besides, we observe the same phenomenon in the corporate bond markets. To be specific, those junk bonds only generate 5% yields, but investors are required to take the risks of 100% defaults loss. It doesn't stop investors with high-risk tolerance to pursue returns. The corporate bond markets have become the most popular investment markets due to the easing monetary policies.

Informing and solving the potential corporate credit problems have become urgent. In the past decades, banks used to rely on external credit rating agencies such as Fitch, Moody's, and Standard & Poor's. However, fewer and fewer banks trust those rating agencies because their rating solutions are qualitative, experience-based, and untransparent. Outsiders can only fully trust their brands and evaluate the risks based on their ratings. Most of the time, they are subjective and only for reference.

To relieve the increasing concerns over the crazy corporate financing markets, we would like to gain a comprehensive understanding of relationships between corporate financial ratios and the probability of bankruptcy. We leverage both linear, non-linear classifiers, and scoring methods to explore the prediction for potential corporate bankruptcy by using computational analysis and modeling techniques.

3. Data Description

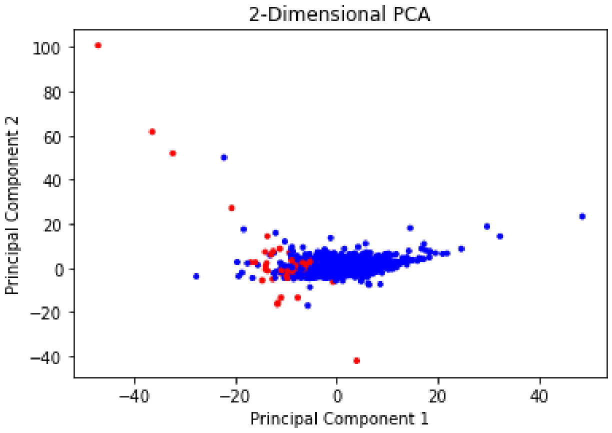
Company Bankruptcy Prediction
Bankruptcy data from the Taiwan Economic Journal for the years 1999–2009
https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction

The dataset includes 6,819 instances with 95 features and 1 label of its status of bankruptcy, collected from the Taiwan Economic Journal for the years 1999 to 2009. 93 of the features are financial ratios and the rest of the 2 are binary data, which are flag labels indicating the potential risks of a company's liability to asset ratio and net income.

Finally, the label of its status of bankruptcy, y=1, represents that the company has filed bankruptcy and y=0 means it's still operating. Of these data points, there are 220 bankrupt companies and 6,599 stable companies. That is, 3.22% of the total instances go bankrupt, and 96.78% is stable.

3.1 PCA Analysis

This is a high-dimensional dataset with 95 features, which may lead to overfitting issues in the modeling process. To solve this problem, we implement Dimensionality reduction (PCA analysis) or detect and remove collinearity to reduce complexity in the dataset. At the beginning, we scale the 95 features and apply PCA analysis to the scaled dataset. Then we project the whole dataset onto the first 2 principal components to obtain a 2-dimensional scatter plot.



As you can see from the above PCA figure, the red points represent the bankrupt companies, and the blue ones are the stable companies. This project aims to develop a classification model to classify the red points for predicting corporate bankruptcy.

3.2 Collinearity Detections

Observing 95 features, we doubt there may be the issue of collinearity. When we examine the features, we find a few of them are highly correlated. For example, there are 3 ROA indicators and 3 Net Value Per Share-related features in the dataset. Moreover, we create the correlation matrix to detect more features that are highly correlated to each other.

In the correlation matrix, we can observe many variables that are highly correlated. Therefore, we remove the features having more 0.5-0.7 correlation coefficients. The threshold of removing features is determined by multiple runs from 0.5 to 0.7. Finally, we decide to use 0.7 because it has the highest model performance than 0.6 and 0.5. It helps us to reduce the dimensions from 95 to 61 variables.

4. Z-score & Kaggle Model Performance Reviews

In the past decades, the Altman Z-Score Model is the most common corporate scoring method in the world. Many banks adopt this method to evaluate the credit risks of a company. The Z-score formula is composed of the following elements:

$$Altman\ Z - Score = 1.2A + 1.4B + 3.3C + 0.6D + 1.0E$$

- A = working capital / total assets
- B = retained earnings / total assets
- C = earnings before interest and tax / total assets
- D = market value of equity / total liabilities
- E = sales / total assets

Status	**Danger**	**Neutral**	**Safe**
Z-score Threshold	Below 1.8	1.8 to 3	Above 3

We apply the Z-score model to the dataset and examine the performance and define any instance having Z-score below 1.8 is predicted as a bankrupt company. The result is not ideal. The Z-score model fails to detect any bankruptcy case in the dataset. However, it's not a surprise for us. Because the Z-score model was developed for traditional manufacturing companies in 1967 and published in 1968.

Though studies show the Z-score method had 82% to 94% accuracy for predicting bankruptcy from 1996 to 1999, it failed to detect any bankruptcy in this dataset. The confusion matrix shows the performance of the Z-score method to the dataset. The Z-score model captured no true bankrupt company and even had 1 false negative.

Z-score Model		**Predict**	
		Stable (>=1.8)	Bankrupt (<1.8)
True	Stable	6598	1
	Bankrupt	220	0

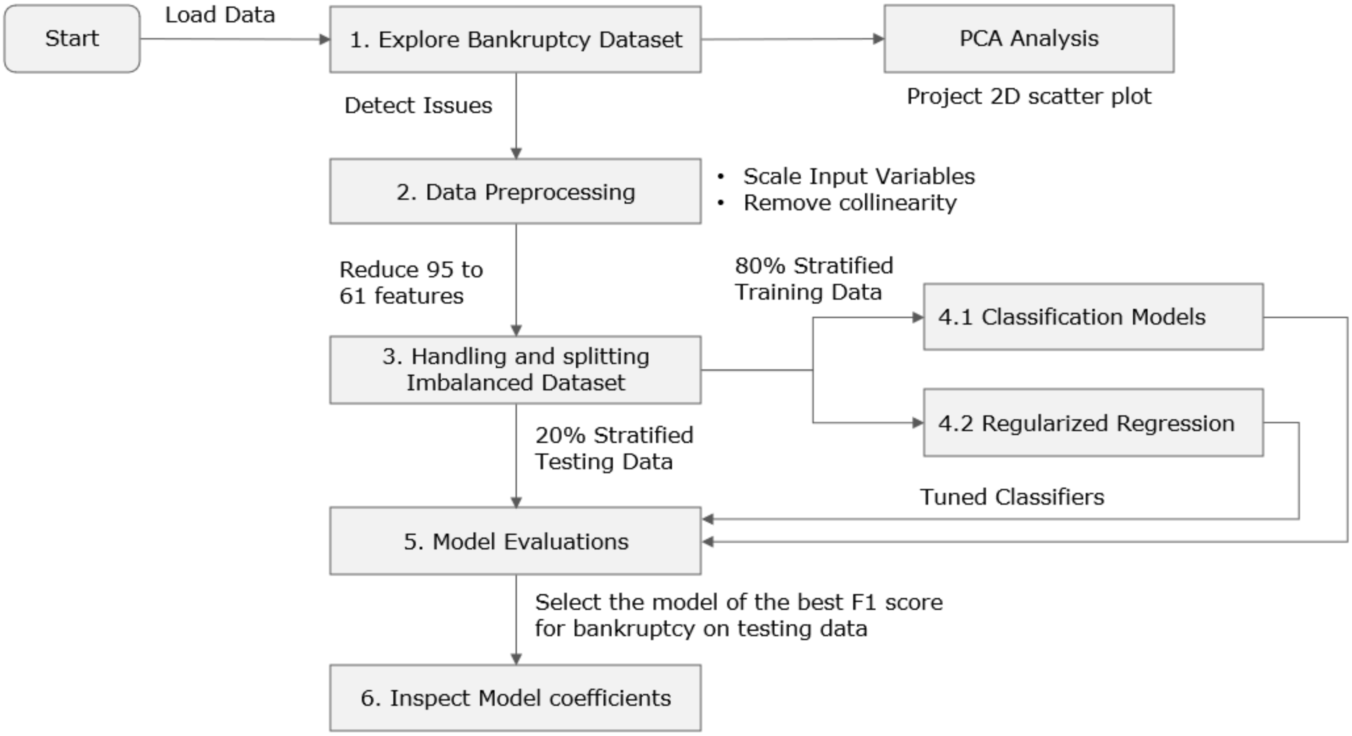
To further compare the performance of our models, we collect the model outputs registered on Kaggle. However, most of the code submissions are using oversampling methods to mistakenly enhance their accuracy. Many data analysts point out this issue in the comments. There is only 1 work submitted by Marto93 that doesn't use oversampling and achieves a decent F-1 score in predicting corporate bankruptcy. This code submission is also labeled as the silver medal on Kaggle and receives 61 votes. The following table describes its model performance:

Submitter	Classifiers	Avg. Accuracy	F-1 score- stable	F-1 score- bankruptcy
Marto93	Catboost	0.96	0.98	0.41
	Logistic Regression	0.84	0.91	0.26

Sources: Bankruptcy_Detection by Marto93

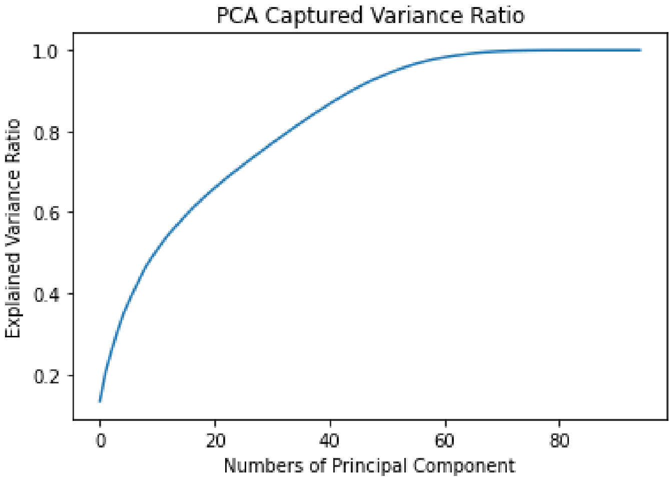
5. Methodology

In this section, we explain the process flow of developing the appropriate models to predict corporate bankruptcy. The following figure describes the process of data-pre-processing, classifiers training, testing, comparisons, and performance evaluations.


Start
Load Data
1. Explore Bankruptcy Dataset
PCA Analysis
Project 2D scatter plot
Detect Issues
2. Data Preprocessing
• Scale Input Variables
• Remove collinearity
Reduce 95 to 61 features
80% Stratified Training Data
4.1 Classification Models
3. Handling and splitting Imbalanced Dataset
4.2 Regularized Regression
20% Stratified Testing Data
Tuned Classifiers
5. Model Evaluations
Select the model of the best F1 score for bankruptcy on testing data
6. Inspect Model coefficients

5.1 Data Pre-processing

As we mention above, the corporate bankrupt dataset is high-dimensional because it has up to 95 features, which makes it complicated to use 95 features to predict testing data's class. Our first intuition is to implement PCA analysis and project the first 2 principal components onto the low-dimensional data. However, the first 2 principal components only capture 20% of the variance. It turns out that it takes more than 50 principal components to reach more than 90% explained variance for the scaled dataset. The PCA method is not very helpful in reducing dimensions. The following plot shows the cumulative variance explained by the number of principal components.

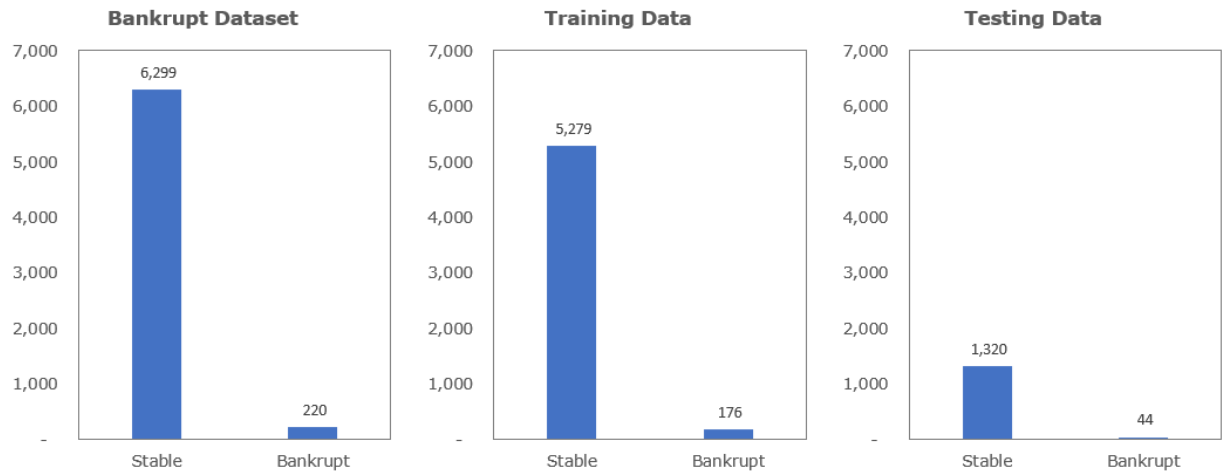


Besides, we are keen to understand what variables contribute greatly to the probability of bankruptcy. Therefore, we decide not to use PCA analysis to pre-process data. Instead, we choose to remove the Collinearity to reduce the data dimensions. Remember the 95 features are financial ratios. It’s likely that the issues of collinearity exist because financial ratios are generally highly correlated to each other. For example, if a company has very high gross margins, it would be highly possible for it to also have high net income margins.

Unsurprisingly, we detect a few variables of the dataset that have issues of collinearity. They are highly correlated to each other. To solve this issue, we need to systematically remove the variables of correlation coefficients higher than the threshold of 0.7 based on the recent research work (Dormann, C. F., J. Elith, S. Bacher, et al. 2013. Collinearity). After removing the variables with higher than 0.7 correlation coefficients, we reduce the variables from 95 to 61. It helps us to simplify the model and improve the performance. Finally, we obtain a dataset containing 6,819 instances with 61 input variables.

5.2 Handling Imbalanced Dataset

Meanwhile, we know the dataset is imbalanced, which only contains 220 bankruptcy cases but has 6,599 stable ones. In this project, we randomly split 80% of the dataset as training data and 20% as testing data. It’s very likely to have test data with no bankruptcy case due to the relatively small bankruptcy samples. To balance the training and testing datasets, we adopt stratified sampling to ensure they are in stratified fashions.



Most importantly, we don’t implement oversampling for this dataset, which leads to biased performance for testing data. The companies with the high risks of defaults are difficult to classify and each company has its unique financial ratios. The oversampling method creates a huge bias for variable importance due to the inappropriate modifications on the sample data.

5.3 Apply Classifications Models

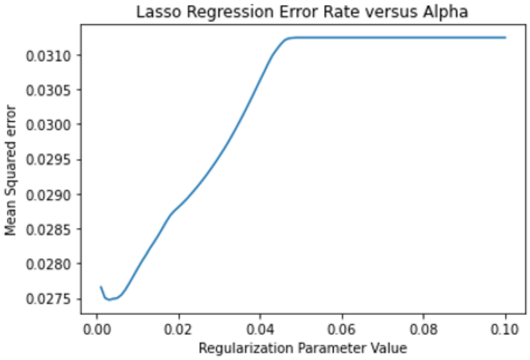
Before inputting data to the training models, it’s important to scale the data first. We scale the input variables but remain the output variable (bankruptcy or not) to stay unscaled. After that, we conduct a series of classification models including MLP Classifier, K-nearest neighbors, Gaussian process classification (GPC), Decision Tree, Random Forest, Naive Bayes, SVM, kernel SVM, and Logistic Regression Models. The following table shows the tuned parameters for these classifiers.

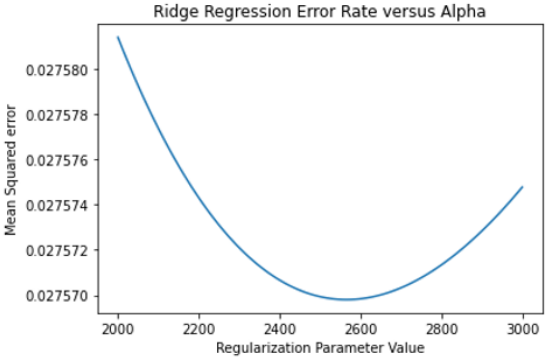
Classifiers	Parameters	Training/Testing	Converge
MLP Classifier	activation='logistic',max_iter=1000	80/20	Yes
KNN	n_neighbors=5	80/20	Yes
Gaussian Process	By defualt	80/20	Yes
Decision Tree	By defualt	80/20	Yes
Random Forest	By defualt	80/20	Yes
Naïve Bayes	By defualt	80/20	Yes
Logistic Regression	By defualt	80/20	Yes
SVM	kernel='rbf'	80/20	Yes
Kernel SVM	kernel='linear'	80/20	Yes

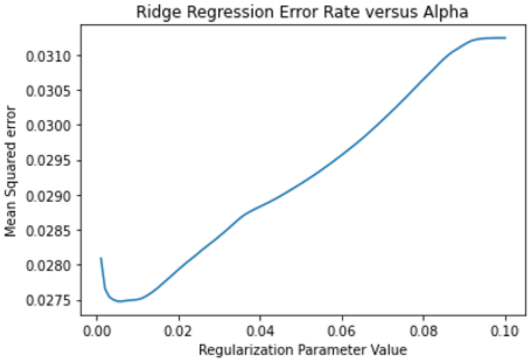
5.4 Regularized Regression (Feature Selections)

The primary goal of this project is not only to find the optimized algorithms to predict corporate bankruptcy but also to develop models that require fewer variable inputs. In reality, it is expensive to collect data from enterprise clients. If banks require up to 95 or 61 features, it will take the clients up to months to prepare and validate their internal financial data. To prevent from problems, we attempt to build a model that meets the above requirements of high performance and low costs.

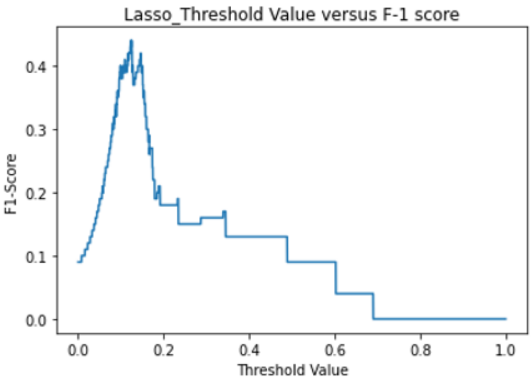
If so, we would conduct feature selections to filter variables that mainly contribute to the probability of bankruptcy. The regularized linear regression models become one of our potential solutions. We select Lasso, Ridge, and Elastic Net to be our candidate models. In this section, we run 10,000 combinations to search the regularization parameter by using 10-fold cross-validations. In this report, the regularization parameter is written as "Alpha" in accordance with the rules of python packages.

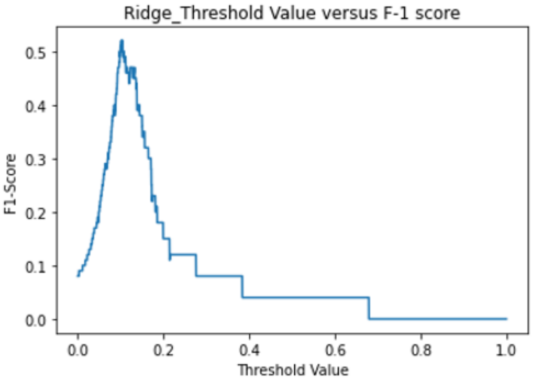


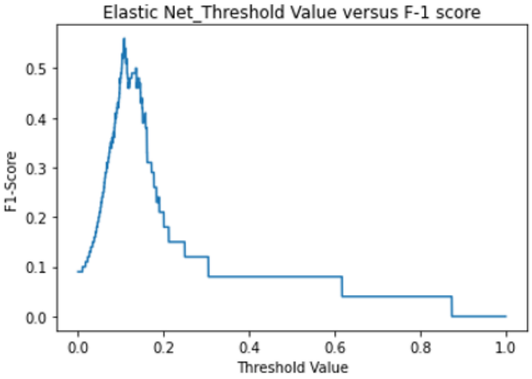




Furthermore, the outputs of the above regularized linear regression models are values instead of classification results. We need to set up a threshold to define if the samples will go bankrupt. By running the 1,000 thresholds from 0.001 to 1 for the training data, we select the threshold value having the best performance on the testing data as the final parameters.







By selecting the optimal parameters from the training part (80% of the dataset), we tune the regularized regression models to achieve the best performance on the testing test. In the next section, we select the optimal combinations of the following parameters and use the tuned models to predict the testing data (20% of the dataset).

Classifiers	**Parameters**	**Training/Testing**	**Optimal Alpha**	**Optimal Threshold**
Lasso	By defualt	80/20	0.003	0.1224
Ridge	By defualt	80/20	2565	0.1011
Elastic Net	By defualt	80/20	0.006	0.1075

6. Evaluation and Final Results

By implementing the above methodology, we adopt stratified random splits for the same 80% training and 20% testing dataset to evaluate each classifier's performance. The following table compares the performance of the classifier candidates. In the case of detecting corporation bankruptcy, we only focus on the F1 score for detecting bankruptcy because company defaults could cause huge losses for banks. Therefore, the penalty costs of the false positive are much larger than the false negative. This is the reason why we mainly focus on the performance of the F1 score for predicting bankruptcy.

Classifiers	**Accuracy**	**F1 score-stable**	**F1 score-bankruptcy**
Z-score	0.97	0.98	0.00
MLP Classifier	0.97	0.99	0.37
KNN	0.97	0.99	0.33

Classifiers	Accuracy	F1 score-stable	F1 score-bankruptcy
Gaussian Process	0.96	0.98	0.19
Decision Tree	0.95	0.97	0.22
Random Forest	0.97	0.98	0.19
Naïve Bayes	0.24	0.35	0.08
Logistic Regression	0.96	0.98	0.10
SVM	0.97	0.98	0.00
Kernel SVM	0.97	0.98	0.04
Lasso	0.97	0.98	0.48
Ridge	0.96	0.98	0.52
Elastic Net	0.96	0.98	0.56

Based on the above comparison table, the Elastic Net Model is the best performer in our classifier portfolio. It achieves the 0.56 F-1 score for detecting bankruptcy. To ensure the model performance, we conduct 100 runs on the stratified random sampling training and testing data and use the average F-1 score. The results indicate their performances are consistent.

Classifiers	Runs	Alpha	Threshold	Avg. F1 score-bankruptcy
Lasso	100	0.003	0.1224	0.4773
Ridge	100	2565	0.1011	0.5172
Elastic Net	100	0.006	0.1075	0.5556

The tests tell us that the performances of the selected models are consistent. Among them, the Elastic Net model is the best performer. What's more, it even outperforms the Kaggle and Z-score models. Next, we apply the tuned Elastic Net classifiers to obtain the confusion matrix:

Elastic Net Model		Predict	
		Stable	Bankrupt
True	Stable	1286	34
	Bankrupt	14	30

The above data implies that if banks use the Elastic Net Model to review their 1,364 corporate loan applications, they will approve 1,300 cases and reject 64 cases. Among its portfolio of 1,300 loans, 14 loans default at the end. It represents 95.30% approval rate and 1.08% default rate by number of cases, which is the good performance for corporate financing.

If we use the metric of loan amounts, the default rate is likely to drop further. It's because larger companies tend to borrow greater amounts of loans, and they are safer than those Small and Medium Enterprises.

Finally, we want to answer the question raised at the beginning: which input variables mainly contribute to the probability of bankruptcy. The Elastic Net Model has only 21 non-zero variables and 1 intercept. The coefficients can be written and reported as:

Input Variable	**Coefficients**	**Note**
Intercept	0.03147	NA
ROA(C) before interest and depreciation before interest	-0.02441	Negatively related to bankruptcy
Interest-bearing debt interest rate	-0.00723	
Non-industry income and expenditure/revenue	-0.00475	
Cash Flow Per Share	-0.00310	
Total Asset Growth Rate	-0.00199	
Tax rate (A)	-0.00184	
Revenue Per Share	-0.00163	
Realized Sales Gross Profit Growth Rate	-0.00134	
Operating Gross Margin	-0.00127	
Operating Expense Rate	-0.00022	
After-tax Net Profit Growth Rate	0.00031	Positively related to bankruptcy
Current Ratio	0.00092	
Total Asset Return Growth Rate Ratio	0.00258	
Operating Profit Rate	0.00260	
Net Value Per Share (B)	0.00347	
Operating Profit Growth Rate	0.00704	
Net Value Growth Rate	0.00783	
Continuous Net Profit Growth Rate	0.01146	
Cash Reinvestment %	0.01676	
Research and development expense rate	0.02231	
Cash flow rate	0.02409	

7. Conclusions

With the rapid evolutions of industries, economic environments, and financing tools, the traditional Z-score method is not sufficient to evaluate the corporate credit risks. To solve the potential corporation, our team wants to seek a better model with more powerful prediction power on potential default risks in the financial systems.

We analyze the UCI bankruptcy dataset of 6,819 instances with 95 input variables and 1 class label. They are data collected from 1999 to 2009 companies listed on Taiwan Stock Exchanges, where there are 220 bankrupt companies and 6,599 stable ones. In the first place, we apply the z-score methods to evaluate the potential bankruptcy cases, but it fails to deliver ideal results. The z-score model is not capable of detecting any bankrupt cases in the dataset.

To solve the issues and improve the model performance, we explore, analyze, and preprocess the dataset by scaling input variables, removing collinearity variables, and adopting randomly stratified sampling to split 80% data as the training set, and the 20% data as testing set. Then we test a variety of classification algorithms such as MLP, KNN, Logit, SVM, and so on. However, they don’t perform very well on the testing dataset. Finally, we tune the regularized linear regression of Lasso, Ridge, and Elastic Net combing the optimal threshold values based on the training dataset to meet our performance standards on testing data. Among the regularized regression candidates, the Elastic Net performs the best.

The result reveals that the Elastic Net Model combing the optimal scoring threshold can achieve up to 0.56 F1 score for detecting potential company bankruptcy. To be specific, the method can help banks to achieve up to a 95.30% loan approval rate and have only the 1.08% default rate by the number of cases.

Furthermore, the Elastic Net model only requires 21 variables Elastic Net to make predictions, which can save huge costs of data collection and validation to improve the efficiency of corporate financing operation better iteratively. Meanwhile, banks can improve their approval rate and reduce their default rate as well by leveraging this model.

References

WILL KENTON. Altman Z-Score. https://www.investopedia.com/terms/a/altman.asp. accessed 31 July 2021.

Dormann, C. F., J. Elith, S. Bacher, et al. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36:27–46

Marto93. Bankruptcy_Detection. Kaggle datasets https://www.kaggle.com/marto24/bankruptcy-detection. accessed 31 July 2021.

Giannopoulos, George & Sigbjørnsen, Sindre. (2019). Prediction of Bankruptcy Using Financial Ratios in the Greek Market. Theoretical Economics Letters. 09. 1114-1128. 10.4236/tel.2019.94072.

Altman, E. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance, 23(4), 589-609. https://doi.org/10.2307/2978933

Altman, E., Iwanicz-Drozdowska, M., Laitinen, E., & Suvas, A. (2017). Financial Distress Prediction in an International Context: A Review and Empirical Ana-lysis of Altman's Z-Score Model. Journal of International Financial Management and Accounting, 28(2), 131-171. https://doi.org/10.1111/jifm.12053

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. Expert Systems with Applications, 83, 405-417. https://doi.org/10.1016/j.eswa.2017.04.006