Machine Learning on Bankruptcy

X Xiao

Jun 30, 2023

1. Introduction

Enterprise bankruptcy poses a significant risk to market agents, potentially leading to capital losses that necessitate early detection and avoidance by financial practitioners. This research unfolds in two segments: Firstly, it delves into classification within supervised learning to forecast bankruptcy trends among companies, utilizing 6,819 observations and chosen financial indicators from Taiwanese enterprises. Such predictions are crucial for the business and financial sectors, aiding pivotal decisions for borrowers and investors alike. Secondly, the study explores unsupervised learning through Principal Component Analysis (PCA), aiming to identify key financial indicators for effective classification.

2. Exploratory Data Analysis

The data were downloaded from Kaggle, which were initially collected from the Taiwan Economic Journal for the years 1999 to 2009. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange.

The original data set is in the structure of 6819 observations with 95 financial indicators and 1 response explaining the financial status of firms-bankruptcy or not. The number of observations is far larger than the dimension of variables. Moreover, due to all features are financial indicators originated from financial statements, they can have strong correlation with some others.

Bankrupt.	**Net.profit.before.tax.Paid.in.capital**	**Operating.Gross.Margin**	**Debt.ratio**
Min. :0.00000	Min. :0.0000	Min. :0.0000	Min. :0.00000
1st Qu.:0.00000	1st Qu.:0.1694	1st Qu.:0.6004	1st Qu.:0.07289
Median :0.00000	Median :0.1785	Median :0.6060	Median :0.11141
Mean :0.03226	Mean :0.1827	Mean :0.6079	Mean :0.11318
3rd Qu.:0.00000	3rd Qu.:0.1916	3rd Qu.:0.6139	3rd Qu.:0.14880
Max. :1.00000	Max. :1.0000	Max. :1.0000	Max. :1.00000
Quick.Ratio	**Interest.Coverage.Ratio**	**Inventory.Turnover.Rate..times**	**Total.Asset.Turnover**
Min. :0.000e+00	Min. :0.0000	Min. :0.000e+00	Min. :0.00000
1st Qu.:0.000e+00	1st Qu.:0.5652	1st Qu.:0.000e+00	1st Qu.:0.07646
Median :0.000e+00	Median :0.5653	Median :0.000e+00	Median :0.11844
Mean :8.377e+06	Mean :0.5654	Mean :2.149e+09	Mean :0.14161
3rd Qu.:0.000e+00	3rd Qu.:0.5657	3rd Qu.:4.620e+09	3rd Qu.:0.17691
Max. :9.230e+09	Max. :1.0000	Max. :9.990e+09	Max. :1.00000

Table 1

According to Accounting commonsense, financial indicators can generally be divided into 3 categories-*Profitability Indicators*, *Solvency Indicators* and *Asset Management Indicators*. In each

group of indicators, 2 or 3 features will be selected to analyze the relationship between each selected variable together. Here *Net Profit Before Tax Paid in Capital*, *Operating Gross Margin*, *Debt Ratio*, *Quick Ratio*, *Interest Coverage Ratio*, *Inventory Turnover* and *Total Asset Turnover* have been selected. Shown as summary information above, *Bankruptcy* is in proportion of 3% in firms. *Net Profit Before Tax Paid in Capital*, *Debt Ratio*, *Inventory Turnover* and *Asset Turnover* have more apparent skewness, which may require logarithm transformation. By contrast, *Operating Gross Margin* and *Interest Coverage Ratio* show less relatively skewness, and both data concentrate at 0.5-0.6.

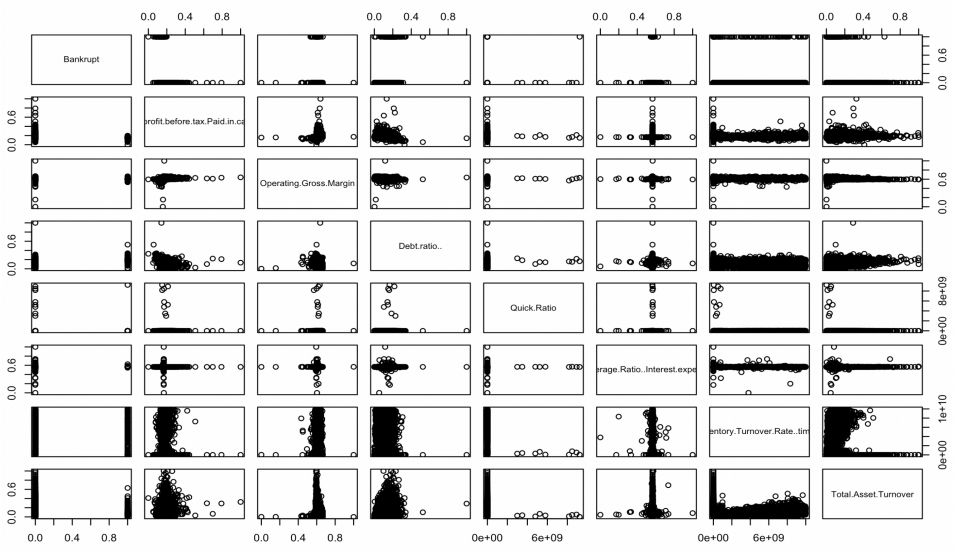

Figure 1

In the pairs plot, first, *Bankruptcy* tends to occur in enterprises with lower *Net Profit Before Tax Paid in Capital (PBT)*, because lower net profit means higher risk of loss in operating, and then higher risk of bankruptcy. However, in other indicators separately, we cannot find much palpable information and we can only know that bankruptcy can happen in any situations. Moreover, in the scatter plots of each financial indicators, there are no apparent positive or negative relationship between much of them, which means that, even if these plots might have low explanation value, it can have high prediction value due to no direct collinearity and low variance in reference to our econometrics theory.

By looking at the histogram, we can find that *Total Asset Turnover* and *Operating Gross Rate* are not normally distributed in their ranges. These situations might because of different average feature values between bankruptcy companies and non-bankruptcy companies. So, we do a separation between these two types of firms. And after that, we can obtain that the *total asset turnover* shows no obvious difference between firm types, so as other indicators like *quick ratio*. While *operating gross rate* of bankruptcy companies are relatively lower than that of non-bankruptcy companies, which may imply that less operating gross rate means less profitability and more risk of bankruptcy. And it can be included as a discriminant features.

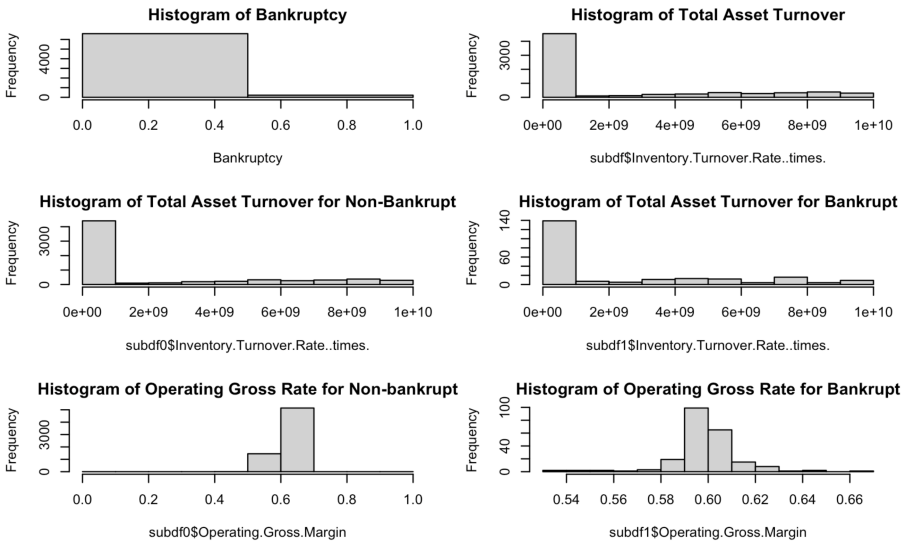

Figure 2

Before the creation of a model to predict the bankruptcy, we try to do a logarithm transformation on several indicators that show subjected to positive skewness and make a comparison between former and latter density function. On *total asset turnover* and *PBT*, a transformation does have benefit on correct the skewness. However, the transformation on *Debt ratio* shows a change from positive skewness to negative skewness, and we would better not transform it.

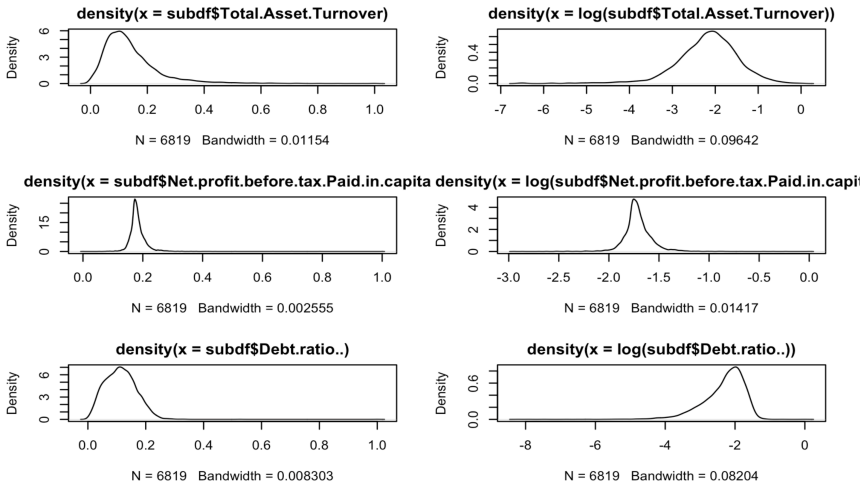

Figure 3

3. Unsupervised Learning

As the data set is composed of 95 financial indicators, all of which are extracted from financial statements of companies, the correlation between some of the variables can be strong.

In this part, it is aimed to analyze the importance and relationship of different financial indicators I selected above and find out a small number of the least correlated variables. In addition of 3 types of most essential indicators we frequently observe, I plan to take another kind of indicators into account, reflecting the prospect of an enterprise. So, I create a new type named *Growth Indicators*, including direct growth rates (e.g., total asset growth rate and cash reinvest ratio) and indirect growth indicators (e.g., *research and development expense rate*). Hence, based on the knowledge of accounting, I plan to undertake a dimension reduction on the sub dataset to obtain a valuable interpretation.

3.1. Principle Component Analysis

3.1.1. Generating Method

Principle component analysis (PCA) is a technique that reduces the dataset dimension and increases the interpretability of dataset constrained on minimized loss of information, aimed to find a $Z = (Z_1, \dots, Z_k)$, according to $X = (X_1, \dots, X_p)$, where $k < p$.

$$\boxed{cov(Z_j, Z_k) = 0 \; for \; all \; j \neq k}$$

$$\boxed{Var(Z_1) > Var(Z_2) > \cdots > Var(Z_k) \; where \; the \; variances \; are \; decreasing}$$

$$\boxed{Z_j = a_{1j}X_1 + \cdots + a_{pj}X_p = {a_j}^T a_j \;\; such \; that \; {a_j}^T a_j = 1}$$

We want to find a_1 that maximize $Var(Z_1)$ under the constraint ${a_1}^T a_1 = 1$. And then find a_2 that maximize $Var(Z_2)$ under the constraint ${a_2}^T a_2 = 1$ and $cov(Z_1, Z_2) = 0$. The procedure continues in the same way to generate till a_j.

3.1.2. Interpretation

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	1.251	1.174	1.027	1.004	0.996	0.984	0.955	0.924	0.840	0.750
Proportion of Variance	0.157	0.138	0.106	0.101	0.099	0.097	0.091	0.085	0.071	0.056
Cumulative Proportion	0.157	0.294	0.400	0.501	0.600	0.697	0.788	0.873	0.944	1.000

Table 2

Under PCA, there are 10 PCs attached with their standard deviation, proportion of variance and cumulative proportion, shown as Table 2.

Table 3 below shows the matrix of variable loadings (i.e., a matrix whose columns contain the eigenvectors). It gives weights of financial indicators on each principal component which we can use to compress the original 10 variables to a few of most essential components.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Net.profit.before.tax.Paid.in.capital	0.456	-0.456	-0.106	0.072	-0.027	-0.053	0.151	-0.309	0.313	0.590
Operating.Gross.Margin	0.548	0.044	-0.283	0.032	-0.095	0.044	0.058	-0.282	-0.680	-0.247
Debt.ratio..	-0.544	-0.249	0.041	-0.047	0.030	0.069	0.328	-0.066	-0.576	0.435
Quick.Ratio	-0.053	0.043	-0.518	-0.126	0.747	-0.288	0.237	0.038	0.080	-0.079
Interest.Coverage.Ratio..Interest.expense.to.EBIT.	-0.011	-0.065	-0.099	-0.825	-0.355	-0.421	-0.041	0.010	0.005	0.000
Inventory.Turnover.Rate..times.	0.058	0.422	0.079	0.099	-0.280	-0.123	0.824	-0.011	0.168	-0.045
Total.Asset.Turnover	-0.132	-0.679	0.068	0.002	-0.048	0.099	0.284	-0.134	0.140	-0.622
Total.Asset.Growth.Rate	0.109	0.192	0.497	-0.447	0.401	0.371	0.077	-0.446	0.036	-0.013
Cash.Reinvestment..	0.384	-0.194	0.213	-0.191	0.154	0.212	0.201	0.775	-0.138	0.069
Research.and.development.expense.rate	0.116	-0.089	0.571	0.221	0.200	-0.723	-0.049	-0.040	-0.185	-0.050

Table 3

In further analysis, to identify how many components that we determine to include when do supervised learning, we would use scree plot. The scree plot is a subjective way to select out variables. And here we would like to select components with variances larger than 1, so PC5, PC 6, PC7, PC8, PC9 and PC10 will be excluded.

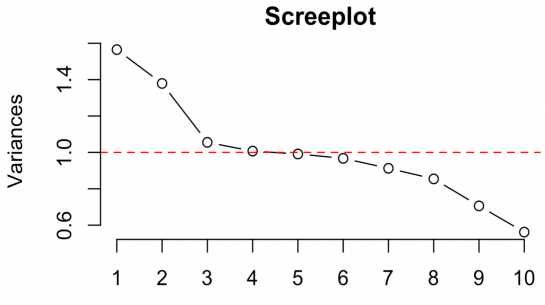

Figure 4

Nevertheless, by looking at the cumulative proportion variance explained, as we set a ceiling at 90, we can only exclude PC9 and PC10. It means that all components can take a large part in variation and there is no rationale to exclude them from the data set.

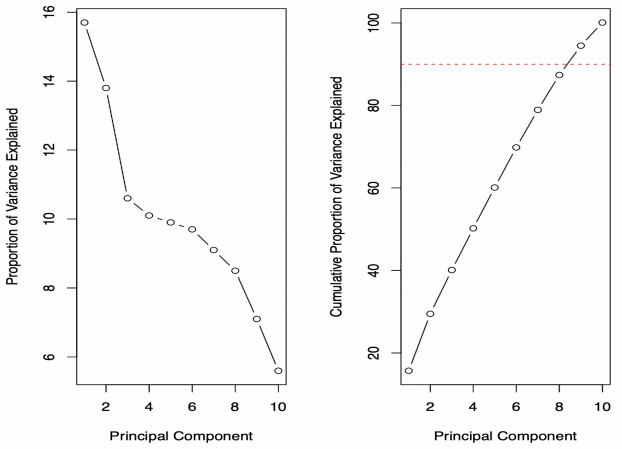

Figure 5

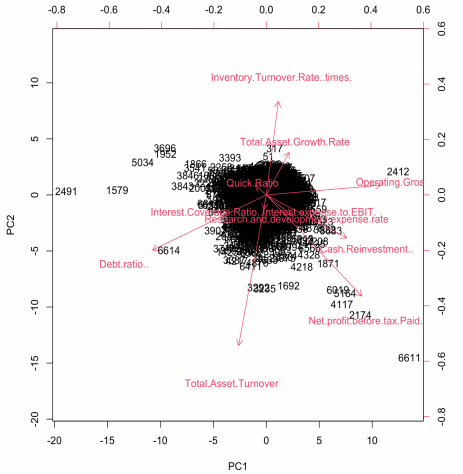

Figure 6

Then, we will generate a biplot. From figure 6, total asset turnover and inventory turnover are two vectors have smallest angle with axis PC2, which means that they are contributed mostly to PC2. To our surprise, the angle between total asset turnover and inventory turnover are nearly 180 degrees, reflecting their highly negative correlation. Additionally, operating gross margin contributes mostly to PC1 with a longer vector than any others, and it also shows highly positive correlation with cash reinvestment rate, total asset growth rate and PBT. Furthermore, other financial indicators-quick ratio, interest coverage ratio and so on-have short length vector in this 2-dimension space, which means that they may have longer length in other dimensions and contribute more to other principal components rather than PC1 and PC2.

3.2. Limitation

The scree plot analysis reveals two inflection points, which presents a challenge in selecting PCs based solely on the 'elbow method'. This complexity arises because, despite identifying components above the elbow for retention, the significance of PCs below this threshold remains substantial. This phenomenon could be attributed to our preliminary selection of key financial indicators, all of which contribute significantly to the variance explained.

Upon deeper examination, alternative analytical approaches such as Factor Analysis and Independent Component Analysis (ICA) warrant consideration. The utility of ICA, in particular, stems from its efficacy in handling non-Gaussian distributions, a characteristic prevalent among our financial indicators as demonstrated in the Exploratory Data Analysis section. These distributions tend to be irregular, suggesting that ICA may offer additional insights into the underlying data structure.

Furthermore, the PCA process has illuminated the potential oversight in pre-selecting subsets of data. A more encompassing approach, involving the entire dataset of 95 financial indicators in the PCA, might yield a comprehensive identification of all critical components. This realization underscores the importance of an inclusive strategy in data preprocessing to ensure no vital information is inadvertently excluded from the analysis.

4. Supervised Learning – Classification

In supervised learning, we aim to design a model that can predict the bankruptcy of firms most accurately. We separate observations into test data set and train data set. And we use 10 financial indicators to fit 4 different models and evaluate which one has highest predict ability on bankruptcy with out-of-sample data.

4.1. Linear Discriminant Analysis (LDA)

In LDA, we assume two classes of individuals share a common covariance matrix as below:

$$X_i \sim \begin{cases} N(\mu_0, \Sigma) \; if \; y_i = 0 \\ N(\mu_1, \Sigma) \; if \; y_i = 1 \end{cases}$$

Table 4 shows the coefficients of linear discriminants we obtained from R. The prior probabilities of groups are set as binomial distribution. In LDA and QDA we cannot do t-test to infer the statistical significance of coefficients. The test error rate of LDA is approximately 0.031.

Prior probabilities of groups:	0 1
	0.9625 0.0375
Coefficients of linear discriminants:	
	LD1
Net.profit.before.tax.Paid.in.capital	-1.398911e+01
Operating.Gross.Margin	-9.604871e+00
Debt.ratio..	1.405115e+01
Quick.Ratio	-3.477506e-10
Interest.Coverage.Ratio..Interest.expense.to.EBIT.	-6.746687e+00
Inventory.Turnover.Rate..times.	-2.111452e-11
Total.Asset.Turnover	-3.962375e+00
Total.Asset.Growth.Rate	-2.497502e-11
Cash.Reinvestment..	4.993970e+00
Research.and.development.expense.rate	-1.995915e-11

Table 4

4.2. Quadratic Discriminant Analysis (QDA)

In QDA, the covariance matrices are different between different classes.

$$X_i \sim \begin{cases} N(\mu_0, \Sigma_1) \text{ if } y_i = 0 \\ N(\mu_1, \Sigma_2) \text{ if } y_i = 1 \end{cases}$$

After fitting data on QDA via R, we got the same prior distribution as LDA due to the same train data set. And the test error rate of QDA is approximately 0.94.

By looking at the group means below, we can find that ***PBT***, ***debt ratio*** and ***quick ratio*** have significant difference between two classes. Bankruptcy enterprises more often to have lower ***PBT***, higher ***debt ratio*** and lower ***quick ratio***, notwithstanding several data of ***quick ratio*** might be wrong when they are collected.

Group means:	0	1
Net.profit.before.tax.Paid.in.capital	0.184	0.148
Operating.Gross.Margin Debt.ratio	0.608	0.599
Debt.ratio	0.111	0.181
Quick.Ratio	8290909	0.005
Interest.Coverage.Ratio.	0.565	0.564
Inventory.Turnover.	2.23E+09	2.15E+09
Total.Asset.Turnover	0.142	0.101
Total.Asset.Growth.Rate	5.58E+09	5.21E+09
Cash.Reinvestment	0.38	0.379
Research.and.development.expense.rate	2E+09	1.5E+09

Table 5

4.3. Logistical Regression

Model for (y_i, X_i):

$$y_i = Bernoulli(\pi(c_k|X_i))$$

$$\pi(c_k|X_i) = \sigma(X_i\beta) \; or \; else \log\left(\frac{\pi(c_k|X_i)}{1-\pi(c_k|X_i)}\right) = X_i\beta$$

```
Coefficients:
                                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                                     3.277e+00  6.364e+00   0.515 0.606568
Net.profit.before.tax.Paid.in.capital          -5.756e+01  8.216e+00  -7.005 2.47e-12 ***
Operating.Gross.Margin                          3.892e+00  7.770e+00   0.501 0.616427
Debt.ratio..                                    1.412e+01  2.922e+00   4.833 1.35e-06 ***
Quick.Ratio                                    -3.738e-09  1.815e-07  -0.021 0.983565
Interest.Coverage.Ratio..Interest.expense.to.EBIT. -1.832e+00  7.723e+00  -0.237 0.812435
Inventory.Turnover.Rate..times.                -5.452e-11  4.615e-11  -1.181 0.237448
Total.Asset.Turnover                           -8.407e+00  2.304e+00  -3.649 0.000264 ***
Total.Asset.Growth.Rate                         7.018e-11  6.110e-11   1.149 0.250704
Cash.Reinvestment..                             1.613e+00  6.208e+00   0.260 0.795025
Research.and.development.expense.rate          -2.861e-11  5.961e-11  -0.480 0.631199
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 639.66  on 1999  degrees of freedom
Residual deviance: 434.21  on 1989  degrees of freedom
AIC: 456.21

Number of Fisher Scoring iterations: 14
```

Table 6

In Table 6, we can look at the P-value, among the financial indicators, ***PBT***, ***Debt ratio*** and ***Total asset turnover*** show high significance level. The test error rate of logistical regression is 0.0299.

4.4. K-Nearest Neighbor (K-NN)

K-nearest neighbor algorithm, commonly referred to as KNN, is an instance-based supervised learning algorithm. It classifies or predicts the label of an unknown sample based on the labels of its nearest neighbors.

Here we set k=1, 10 and 100 separately to test the performance of the model. And when k=1, the test error rate is 0.05209; when k=10, the test error rate is 0.03050; when k=100, the test error rate is 0.03009. The test error rate when k=10 is approximately equal to that when k=100

4.5. Comparison and Evaluation

When comparing error rate on test data set, we find that LDA, Logistic Regression and K-NN all have good performance on test data and can predict bankruptcy of enterprises with only 3% error rate, while QDA has poor performance when working on out-of-sample with more than 9% error rate.

Technique	Test Error Rate
LDA	0.031
QDA	0.094
Logistic Regression	0.03
K-NN	0.03

Table 7

Discriminative model applies regression without specifying the distribution of independent variables. Logistic regression usually performs better as it has less parameters and makes fewer assumptions. Generative models are used more often when we want to understand X itself and generate from it.

4.6. Limitation

In our classification analysis, a deeper exploration into classification tree methodologies is warranted to potentially enhance predictive accuracy. Evaluating model performance should encompass a broader array of metrics, including sensitivity, specificity, ROC curves, and scoring rules. Prioritizing models with high sensitivity is crucial, as it minimizes the risk of misclassifying actual bankruptcies (true positives) as non-bankruptcies (false negatives). This approach aims to safeguard the interests of both investors and borrowers by ensuring a more accurate identification of firm bankruptcies.

$$true\ positive\ rate = sensivity = \frac{true\ positives}{true\ positives + false\ negatives}$$

$$false\ positive\ rate = (1 - specificity) = \frac{false\ positives}{false\ positives + true\ negatives}$$

5. Conclusion

This project endeavors to provide actionable insights for financial practitioners, facilitating the prediction of corporate bankruptcy. Our exploratory data analysis revealed that the chosen indicators exhibit low correlation, enhancing the robustness of our predictive model. Subsequent principal component analysis efficiently distilled the financial indicators into four principal components, aligning with the anticipated categorization of three traditional and one novel financial metric. In the realm of supervised learning, our comparative analysis across four models established that both Logistic Regression and K-NN models exhibit superior performance in forecasting firm bankruptcy. These findings hold significant promise for advancing predictive accuracy in financial risk assessment.