

Bankruptcy Prediction Based on Financial Data using PCA

Zhou Li

Department of Accounting, Xiamen University, Fujian, China

32920192201076@stu.xmu.edu.cn

Abstract. The survival of a company is a problem faced by every company operator and investor. It is very important to effectively predict the bankruptcy of a company. There are two problems in predicting company bankruptcy, one is the selection of eigenvectors, the other is the selection of prediction models. In this paper, based on the financial data of listed companies, the principal component analysis method is used to reduce dimensions and build a classification model to predict the bankruptcy of companies. It is found that the profitability, debt and liquidity of the company are significantly related to the bankruptcy of the company. In addition, it is found that the decision tree model has excellent prediction ability.

Keywords: bankruptcy prediction; financial data; classifier.

1 Introduction

The number of corporate bankruptcies worldwide is expected to rise 15% year on year in 2022, according to a recent report survey, as the unexpected economic impact of the COVID-19 pandemic and the historic, even expected, continuing impact on the production and sales of many companies are facing a crisis, and the trend of bankruptcies will be greatly accelerated. The impact of bankruptcy and potential bankruptcy is huge, which will affect the investment returns of investors and the future decisions of enterprise managers. Therefore, we need to predict the future bankruptcy of enterprises.

In 1966, Beaver first proposed that financial indicators have high research value in the study of enterprise bankruptcy prediction search related to corporate bankruptcy prediction [1]. In terms of selecting characteristics, Tian et al. used LASSO method to set three financial ratio indicators as variables to build a model and conduct research on bankruptcy prediction of Japanese enterprises [2]. Kou et al. proposed a two-stage multi-object feature selection algorithm for feature extraction to achieve similar performance of classifiers [3]. In terms of model selection, Le et al. used an improved extreme gradient lifter, which has certain advantages [4]. Shen et al. found that random forest classifier is superior to some other models in the task of imbalanced data classification [5]. Based on the above research, we find that there are some problems such as feature screening and model selection in enterprise forecasting.

This paper collects the financial data set of listed companies, and conducts correlation analysis and classification model construction and comparison through PCA dimensionality reduction method. Through data analysis and model construction, investors and investment companies can evaluate through the model, which can be used as a reference for investment decisions, and managers of the company can predict the business future of the company through modeling.

2 Method

2.1 PCA

Principal component analysis (PCA), first raised by Pearson in 1901, was used to fit the best lines and planes to some points in space. Subsequently, a large number of literatures have made in-depth study on it. PCA can reduce data, compress data, model, detect singular value and so on.

The principle of principal component analysis is as follows [6]:

Suppose there is a data set X , each row of which represents the sampling of process variables at different moments, and its columns correspond to S process variables. The data array can be decomposed by SVD as follows:

$$X = \sum_{i=1}^s \bar{\sigma}_i r_i p_i^T \quad (1)$$

Here, p_i and r_i are the eigenvectors of $X^T X$ and XX^T respectively, which are standard orthogonal vectors, and $\bar{\sigma}_1 > \bar{\sigma}_2 > \dots > \bar{\sigma}_s$ is singular value, that is, the non negative square root of the x eigenvalue. Let $t_i = \bar{\sigma}_i r_i$, then:

$$X = \sum_{i=1}^s t_i p_i^T \quad (2)$$

t_i is the score vector, each element of which corresponds to the score of a sampling point, and the variable corresponding to t_i is called the i th principal element; p_i represents the direction of the principal component, that is, the load vector. If $R = \frac{1}{N-1} X^T X$ is set, the relationship between its eigenvalue and singular value is as follows:

$$\lambda_i = \frac{1}{N-1} (\bar{\sigma}_i)^2 \quad (3)$$

In actual analysis, in order to eliminate the impact of dimensions on the results and extract the maximum change information (maximum variance) in the process, the original process data is often standardized before the above analysis.

2.2 Analysis steps

This paper analyzed the data according to the process shown in Fig.1, mainly including: data description, data preprocessing, model building, analysis.

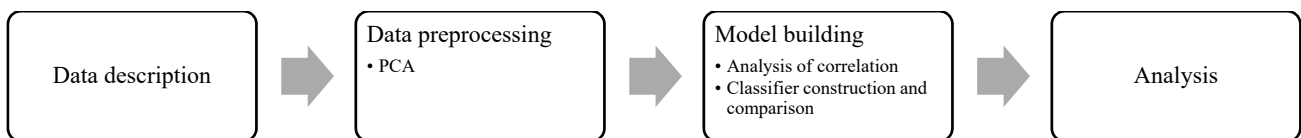


Fig. 1 Analysis steps

3 Experiment

In order to evaluate the prediction of the company's financial situation on bankruptcy, we applied the above methods in practice.

3.1 Data description

The data are derived from Taiwan Economic Journal for the years 1999 to 2009, which contain 95 financial indicators and 1 bankrupt indicator. Whether the company is bankrupt was determined by the standard of the Taiwan Stock Exchange. Dataset are cited from UCI Machine Learning Repository.

The Company Bankruptcy Prediction Dataset contains 6819 rows describing company bankrupt statement and financial data, each row with a single dumb variable as 1 dependent variable and 95 indexes as independent variables. The general information is displayed in the Fig.2 below. The dependent 'Bankrupt?' is a dumb variable labeled with 0 and 1, 1 representing bankrupt statement and vice versa. The original dataset contains 6599 "0" and 220 "1" in "Bankrupt?". (Fig.2) The rest are company's financial data. Part of the data description is shown in the Fig.4. All the variable's class is 'int' or 'float'.

Table 1 Original dataset

	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	...	Net Income to Total Assets
0	1	0.370594	0.424389	0.40575	0.601457	0.601457	0.998969	0.796887	0.808809	0.302646	...	0.716845
1	1	0.464291	0.538214	0.51673	0.610235	0.610235	0.998946	0.79738	0.809301	0.303556	...	0.795297
2	1	0.426071	0.499019	0.472295	0.60145	0.601364	0.998857	0.796403	0.808388	0.302035	...	0.77467
3	1	0.399844	0.451265	0.457733	0.583541	0.583541	0.9987	0.796967	0.808966	0.30335	...	0.739555
4	1	0.465022	0.538432	0.522298	0.598783	0.598783	0.998973	0.797366	0.809304	0.303475	...	0.795016

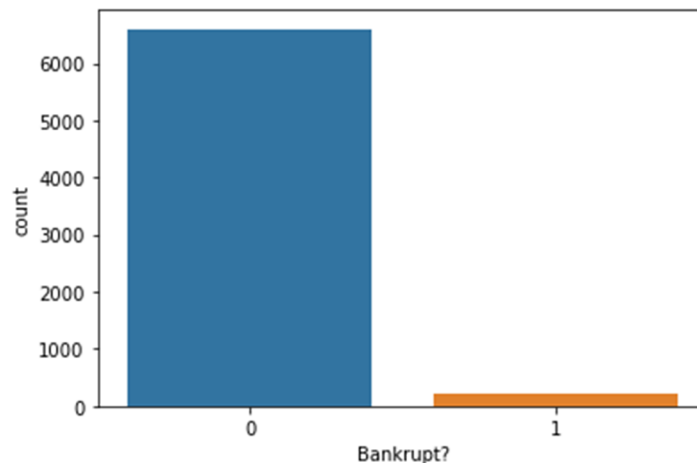


Fig. 2Dependent variable

Table 2 Data describe

	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	...	Net Income to Total Assets
count	6819.000	6819.000	6819.000	6819.000	6819.000	6819.000	6819.000	6819.000	6819.000	6819.000	...	6819.000
mean	0.032	0.505	0.559	0.554	0.608	0.608	0.999	0.797	0.809	0.304	...	0.808
std	0.177	0.061	0.066	0.062	0.017	0.017	0.013	0.013	0.014	0.011	...	0.040
min	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	...	0.000
25%	0.000	0.477	0.536	0.527	0.600	0.600	0.999	0.797	0.809	0.303	...	0.797
50%	0.000	0.503	0.560	0.552	0.606	0.606	0.999	0.797	0.809	0.304	...	0.811
75%	0.000	0.536	0.589	0.584	0.614	0.614	0.999	0.798	0.809	0.304	...	0.826
max	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	...	1.000

3.2 Data Processing

In this part, the dataset had been processed by the following process. 1) Check and handle missing values. 2) Check and handle outliers. 3) Dimensionality reduction by PCA method.

The dataset is very clean and normalized, there are no unwanted or unnecessary columns, either no missing data.

The outliers were processed by the boxplot method, and the outliers in each feature were eliminated, and 6270 fields were finally obtained.

Since this dataset contains up to 95 feature variables, the dimensionality will be reduced through principal component analysis in the following part of this paper [7]. Before PCA, all independent variables are standardized. Then, PCA algorithm is used to reduce independent variables. Through PCA, 30 components can explain about 79% of the total variance, while the first 3 components can explain 27% of total variance. (Fig.3)

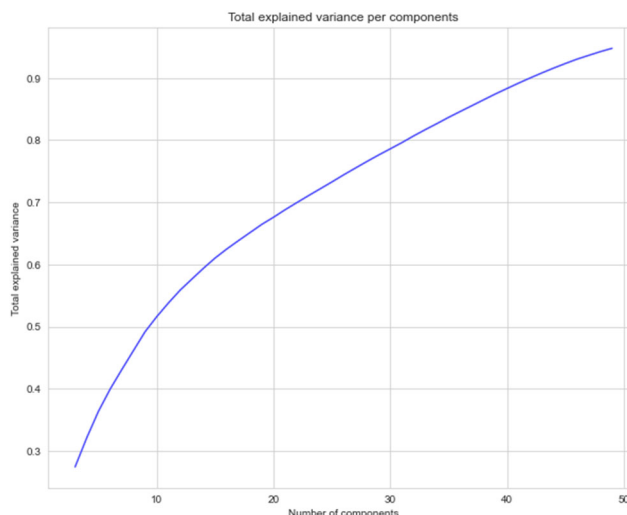


Fig. 3 PCA result

3.3 Modeling

The modeling process contains two parts. The dataset will be modeled with 3 components, which explain 27% of total variance and 30 components, which explain 79% of total variance. First, through only using 3 components, the data can be interpreted quite well when using business knowledge, though the total variance is not as high as 30 components. The 3 components extracted do not share any dominant features, while it also has strong internal consistency and is closely related to existing business concepts. Second, through 30 components, the data is modeled by the following classifiers: Logistic Regression Model, SVM Model, Naive Bayes, Decision Tree, K-NN, and Random Forest Classification, then their precision is compared, in order to build an appropriate model to forecast bankrupt.

3.3.1 Analysis based on 3 components

In this part, by analyzing the characteristic variables contained in the first 3 principal components and through linear analysis, we will analyze the correlation between financial indicators and company bankruptcy.

From the heat map (Fig.4), each component is constructed by 4 - 5 dominant features (those with lightest or darkest color). It's reasonable to consider them to be the more meaningful metrics to look at when monitoring company's likelihood to when bankrupt.

There's some internal correlation between the feature of one component as well. The features index will be mapped with their name to have a closer look at what they are. In details:

Profitability (PC1): All features making up these components are related to return, earning, profit. The main differences in each feature are how tax, interest, depreciation, capital and seasonality are taken into account. Profitability represents the profitability and financial situation of the enterprise.

Liability (PC2): 4/5 features are related to how much debt (long/short term) the company bares. Liability represents the liabilities of the business.

Liquidity (PC3): 2/5 features are clear indicator of the company liquidity (Current Assets/ Total Assets, Quick Assets/ Total Assets), while the other 3 indicate the demand for liquid assets to finance company operations (Total Asset Turnover) or debt repayment (Current Liability to Assets). Liquidity represents the liquidity and solvency of the enterprise.

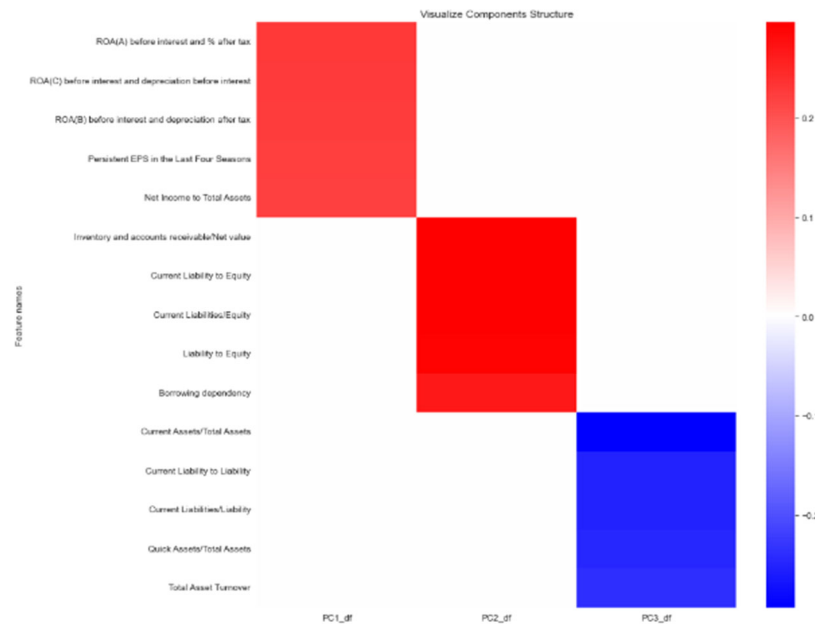


Fig. 4 Three components structure

From the visualizing effort (Fig.5), it's quite clear that PC1 - Profitability does have some predictive power as most bankrupt cases have PC1 value lower than 0.

There are opposite linear trends in the Fig.5.

For surviving companies, their profitability and liability components follow a linear positive relationship.

For bankrupt companies, their profitability and liability components follow a negative linear relationship.

Some statistic tests will be taken to see if this relationship is meaningful.

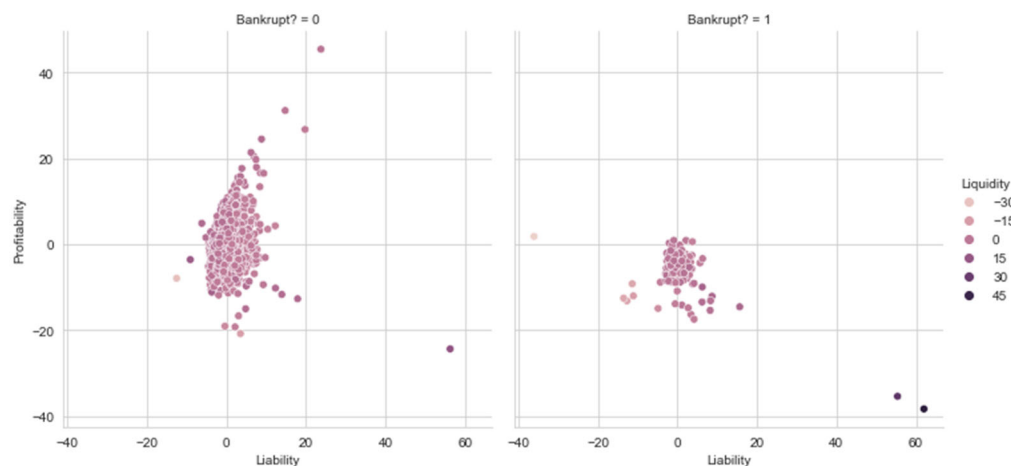


Fig. 5 Relation chart

From the regression result (Table 3), the relationship in both models is statistically significant (Prob (F-statistic)) < 0.005). It seems that for bankrupt companies, liability can predict profitability better than for surviving companies (R-squared is 24.6% versus 2.1%).

Table 3. Regression result

	R-squared	coef	P> t
surviving company	0.021	0.2316	0.000
bankrupt company	0.246	-0.4856	0.000

3.3.2 Analysis based on 30 components

In this part, in order to build an appropriate model to forecast bankrupt, the first 30 components, which can explain about 79% of the total variance, are use to build models through Python. Final data were divided into a train set of 70% and a test set of 30%, with a “random_state = 15”. SVM, Naive Bayes, KNN, Logistic Regression, Decision Tree, Random Forest Classification classifiers will be modeled by “Sklearn” library [8][9]. Their parameters are shown below:

Logistic Regression: random_state = 0;

SVM: random_state = 0;

KNN: n_neighbors = 5, metric = 'minkowski', p = 2;

Naive Bayes: GaussianNB();

Decision Tree: criterion = 'entropy', random_state = 0;

RandomForest: n_estimators = 10, criterion = 'entropy', random_state = 0.

Subsequently, to evaluate the predictive performance of each model, we calculated their accuracy, recall, precision and f1-score [10].

Given total samples, if we set:

the number of positive samples that is correctly predicted is TP; the number of positive samples that is wrongly predicted is FN; the number of negative samples that is wrongly predicted is FP; the number of negative samples that is correctly predicted is TN.

Then:

$$Accuracy = \frac{TP + TN}{Total\ samples} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$F1 - score = \frac{2 * Recall * Precision}{Recall + Precision} \quad (7)$$

The accuracy is shown in the Fig.6.

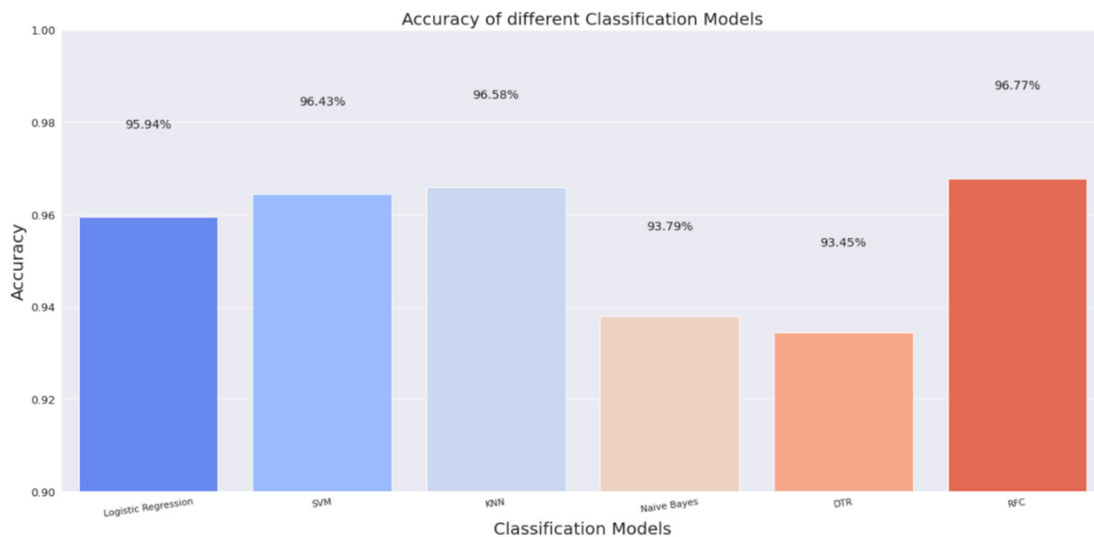


Fig. 6 Accuracy

As we can see from the results, the accuracy of all the models is very high. This is mainly due to the large difference in the distribution of variables in the original data set, with the number of non-bankrupt accounting for more than 90% of the total, which will lead to the situation of high accuracy in the case of inaccurate prediction of the possible model. Therefore, accuracy is not very helpful for our model selection, and it doesn't apply to a bankrupt company. So, we tried to use recall, precision. The recall is mainly used to judge how much probability the company will have when it is finally

bankrupt. The precision mainly determines how many probabilities are predicted when the company is finally not bankrupt. The results of recall and precision are shown in the Fig.7, Fig.8.



Fig. 7 Recall

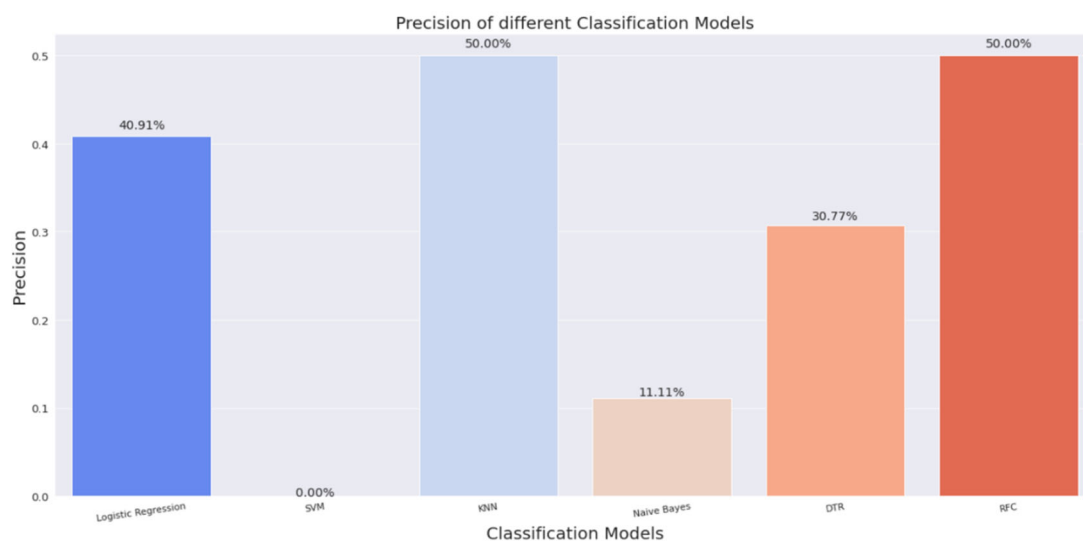


Fig. 8 Precision

As can be seen from the Fig.7, decision tree (DTR) has the highest recall, that is 24.24%. This is at a higher level when the housing exchange rate is generally low and the sample size is small, so it has better prediction performance.

As can be seen from the Fig.8, both KNN and random forest (RFC) have the highest recall, that is 50%.

Since different models have different levels of recall and precision, we consider calculating their F1-score for comparison and evaluation. The F1-score takes into account both the precision and recall ratio, so that they can reach the highest and take a balance.

The result of f1-score is shown in Fig.9.

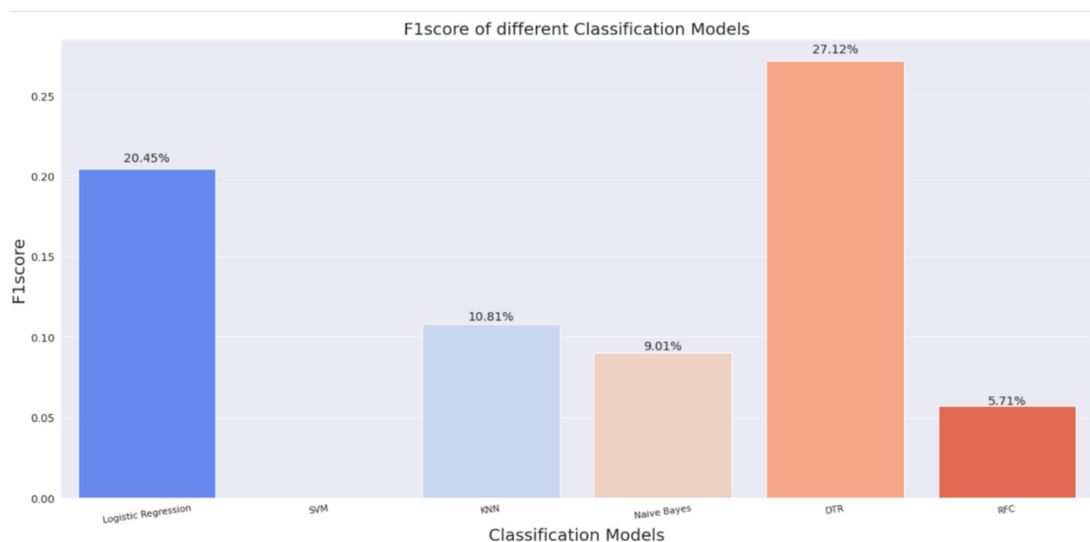


Fig. 9 F1score

It can be seen from the figure that after considering recall and precision, DTR has the highest F1-score, that is 27.12%, and the second one is logistic regression, that is 20.45%.

4 Conclusion

Based on the corporate financial data set of Taiwan magazine, this paper uses PCA method to reduce dimensions, and selects 3 and 30 principal components for analysis respectively. For the three principal components, this paper conducts a correlation analysis, and believes that the profitability, debt and liquidity of the company are significantly related to the bankruptcy of the company. For the part of 30 principal components, this paper constructs a variety of classification models, and compares each model through F1 score evaluation. It is found that the decision tree model has the highest performance in predicting corporate bankruptcy. Compared with previous studies by scholars, this study has two advantages. First, it emphasizes the influence of the company's profitability, debt and liquidity on the company's bankruptcy. Second, it finds that the decision tree model is more convenient and effective for predicting the company's bankruptcy based on financial data.

In terms of practical application, companies and investors can not only predict whether a company can make a decision tree model, but also predict the life of a company by analyzing its profitability, debt and liquidity. This will help company managers to operate the company better, and also help investors to make rational investment decisions.

References

- [1] BEAVER W H. Financial Ratios as Predictors of Failure [J]. Journal of Accounting Research,1966,4:71-111.
- [2] TIAN S N, YU Y. Financial Ratios and Bankruptcy Predictions: An International Evidence [J]. International Review of Economics & Finance, 2017, 51: 510-526.
- [3] KOU G, XU Y, PENG Y, et al. Bankruptcy Prediction for Smes Using Transactional Data and Two-Stage Multiobjective Feature Selection [J]. Decision Support Systems,2020(publish), <https://doi.org/10.1016/j.dss.2020.113429>.
- [4] LE T, VO B, FUJITA H, et al. A Fast and Accurate Approach For Bankruptcy Forecasting Using Squared Logistics Loss with GPU-Based Extreme Gradient Boosting [J].Information Sciences, 2019, 494: 294-310.
- [5] SHEN F, LIU Y Y, WANG R, et al. A Dynamic Financial Distress Forecast Model with Multiple Forecast Results Under Unbalanced Data Environment [J]. Knowledge-Based Systems,2020,192:105365.

- [6] Pearson, K. (1901) LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2, 559-572.
- [7] Josse, J., Wager, S., & Husson, F. (2016). Confidence Areas for Fixed-Effects PCA. Journal of Computational and Graphical Statistics, 25(1), 28–48.
- [8] SHEN F, LIU Y, WANG R, et al. A dynamic financial distress forecast model with multiple forecast results under unbalanced data environment [J] . Knowledge-based Systems, 2020, 192: 105365.
- [9] CHOI H, SON H, KIM C. Predicting financial distress of contractors in the construction industry using ensemble learning [J] . Expert Systems with Applications, 2018, 110: 1
- [10] Luo L, Chen X. Integrating piecewise linear representation and weighted support vector machine for stock trading signal prediction [J]. Applied Soft Computing, 2013, 13(2): 806-816.