Automation and Machine Learning (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2022.030101
ISSN 2516-5003 Vol. 3 Num. 1

Predicting Financial Fraud in Listed Companies Based on Machine Learning Models

Jiacheng Lan[1], Peisen Li[1], Chengzeya Chen[1], Jianwen Deng[1], Sixi Gong[2]

[1]Department of Math, Guangdong University of Education, Guangdong, 510303, China
[2]School of Electronic Information and Electrical Engineering, Huizhou University, Guangdong, 516007, China

Keywords: Financial fraud, Industry classification, Machine learning, Characteristic selection, Feature similarity

Abstract: "Financial fraud" refers to listed companies falsifying financial statements, falsely reporting or concealing part of the financial data of the company. Based on the training data processing of missing value, exception value, standardized processing or other data processing on the training data set, among the 19 major industries in society, processing unbalanced data for 8 industries. Using SVM to select the optimal equilibrium method for each industry, and the remaining 11 industry data do not perform any equalization operations. Subsequently, The Weight, Filter, Wrapper, Embedded and other methods are used to select data characteristics, and combines the actual economic significance to obtain the final characteristics. Finally, through seven models of prediction, using AUC score as evaluation index to predict which listed companies may have fraud. On the premise of not deviating from the actual laws and practical significance, through the self-constructing function and using it to make a heat map. In this way, the similarities and differences of data indicators related to financial fraud of listed companies in different industries are obtained.

1. Introduction

With the rapid development of China's economy, the securities market has continued to expand and the number of listed companies in different industries and of different sizes has increased and now exceeds 4,000[1]. However, in recent years, cases of financial data fraud have emerged from time to time among listed companies, and liquidity crises and credit debt defaults have also emerged in recent years[2]. Unlawful companies have arrived at producing false financial statements by means of false transactions, false assets, false income recognition in advance, and the use of transitional accounts to reconcile profits to raise funds illegally from investors[3]. The core content of the problem is about the analysis of the financial data of listed companies, the main purpose is to distinguish whether the financial data of listed companies are fraudulent or not, to dig out the characteristic factors[4], so as to build a mathematical model to find the listed companies that may have financial fraud, and to reduce and avoid the risk of mines for investors as far as possible[5].

2. Establishment and Solution of Model

2.1 Data pre-processing

2.1.1 Missing value processing

The missing values were first observed in the raw data, and the proportion of missing values in the total data was calculated for each industry, and a bar chart was produced (Figure 1), in which the missing values of each characteristic factor were treated differently in terms of their proportion. The data of each industry is grouped and processed with the purpose of adopting different feature selection methods for different industries, and their feature screening is also different.

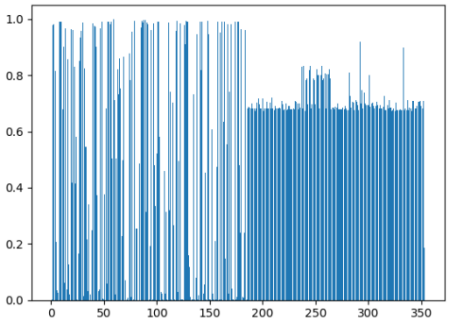

Figure 1 Proportion of missing data by sector

Based on the proportion of missing data for each industry in Figure 1, the following three treatments were adopted.

For data with a proportion of missing data greater than 0.5, the data were processed by way of elimination between them. For features with a percentage of missing data between 0.5 and 0.2, they will be filled in. Therefore, in the missing value filling part, the zero value is treated as a null value, and the data to be filled is classified by industry and the mean value is filled for that part of the data. For data where the proportion of missing values is in the range of less than 0.2, we use the random forest fill process, which, as a relatively new machine learning method, is a classifier that contains multiple decision trees.

The result is 22213 rows and 101 columns of data after the missing value process, including the 'whether falsified in the year' factor removed before filling.

2.1.2 Outlier processing

The specific treatment method has different results for different industries, and the following is an example of a bill receivable from the manufacturing industry. A box plot of the manufacturing industry data is shown in Figure 2 below, which shows that there are outliers that need to be dealt with. However, the size of the outliers is bound to vary greatly or slightly depending on the data of each industry. Therefore, it is necessary to make box plots according to different industries and then make appropriate processing of the outliers, replacing the upper and lower boundaries of the box plots of the industry with the outliers of the corresponding industry data.

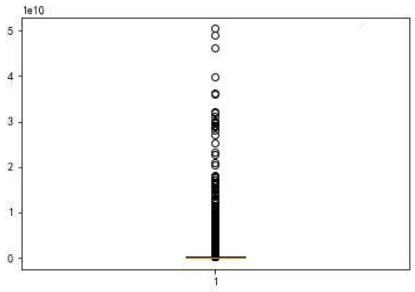

Figure 2 Manufacturing 'Notes Receivable' box plot

2.1.3 Standardization

Before machine learning, there is more or less variability in the data of each feature after the outliers and missing values have been processed. In order to account for the variability in the data of each feature, the variance of each data needs to be standardized first in order to filter the feature factors. The standardized variances are as follows:

$$\bar{x} = \frac{x - \bar{x}}{std(x)} \tag{1}$$

where $\bar{x}$ is the mean; $std(x)$ is the standard deviation.

2.2 Data imbalance treatment

Based on the bookkeeping data of listed companies in 19 industries, this includes 11 industries where the number of falsified training sets does not exceed two such that no balancing treatment can be applied to the industry and therefore no balancing treatment is applied to these 11 industries.

The industries for which no balancing treatment is applied include: accommodation and food services; health and social work; residential services, repairs and other services; construction; education; culture, sports and recreation; water, environment and utilities management; scientific research and technical services; rental and business services; general; and mining, total 11.

The other eight industries were randomly under-sampled, randomly over-sampled, ADASYN over-sampled, and AUC was used as the evaluation indicator, and the final imbalance treatment was chosen as follows in Figure 3:

AUC value of each industry in four imbalance treatment methods

	Direct prediction	Naive random undersampling	Naive random oversampling	Adasyn oversampling
Transportation, storage and postal services	0.5	0.75	1	0.981012658
Accommodation and catering	0	0	0	0
Information transmission, software and information technology services	0.479310345	0.714285714	0.973170732	0.983796296
Agriculture, forestry, animal husbandry and fishery	0.96875	0.5	0.985714286	0.954545455
manufacturing	0.53333333	0.529655172	0.715037114	0.788269951
Health and social work	0	0	0	0
Residential services, repair and other services	0	0	0	0
construction	0	0	0	0
real estate	0.583425414	0.55	0.948275862	0.960784314
Wholesale and retail	0.573252688	0.5	0.972	0.995934959
education	0	0	0	0
Culture, sports and entertainment	0	0	0	0
Water conservancy, environment and public facilities management	0	0	0	0
Electricity, heat, gas and aquaculture and supply industry	0.497175141	0.5	1	0.995238095
Scientific research and technology services	0	0	0	0
Leasing and business services	0	0	0	0
comprehensive	0	0	0	0
Mining	0	0	0	0
finance	0.496875	0.5	0.965909091	0.969201807

Figure 3 19 industries

The highest AUC values for each sector trained on the four imbalance treatments are shown in blue.

Analysis of feature selection results:

There are four general feature selection methods as follows: weighting method, filtering method, wrapping method and embedding method, where the filtering method, wrapping method and embedding method data are taken as an example for manufacturing industry(Tables 1-4).

Table 1 Characteristics of the filtering method selection

feature selection method	selected result
interrelationship metric	2 4 5 7 9 11 17 18 20 21 23 27 28 50 55 68 69 78 82 87
chi-square test	7 13 14 15 37 53 57 58 60 61 65 67 79 80 81 83 85 90 91 96

Table 2 Characteristics of the wrapping method selection

feature selection method	selected result
logistic regression recursive feature elimination method	6 12 13 16 31 33 36 37 43 44 48 63 67 74 79 80 82 88 90 96
random forest regression recursive feature elimination method	9 11 19 21 23 27 36 44 47 53 54 57 63 67 69 78 82 87 93 94
logistic regression models	7 13 16 17 32 34 38 44 45 62 66 72 79 80 81 82 87 91 92 97

Table 3 Features selected by the penalty term-based embedding method

feature selection method	selected result
SVM Model	7 13 14 17 32 38 44 45 61 62 66 79 80 81 82 86 87 91 92 97
Lasso Model	7 17 32 34 38 44 45 50 59 61 62 66 77 80 82 85 86 87 91 97
logistic regression models	7 13 16 17 32 34 38 44 45 62 66 72 79 80 81 82 87 91 92 97

Table 4 Features selected by the tree model-based embedding method

feature selection method	selected result
random forest model	1 3 10 12 15 18 22 24 25 29 43 48 59 61 67 76 79 83 88 91
decision tree model	7 15 16 17 34 37 38 44 45 55 58 61 62 66 79 80 81 91 93 97

The features obtained by all methods were counted (taking manufacturing industry as an example) and those with a number of occurrences greater than or equal to four were selected as the final features, with a total of 19 features selected. The results are shown in Table 5, which contains features such as long-term equity investments, other non-current assets, short-term borrowings, deferred income tax liabilities, total liabilities, total owners' equity attributable to the parent company, total owners' equity (or shareholders' equity), total liabilities and owners' equity (or shareholders' equity), and payments of other cash related to financing activities, all of which are economically meaningful and are therefore final features.

Table 5 Final selection of features

selection of feature results		
long-term equity investments	other non-current assets	short-term borrowings
deferred income tax liabilities	total liabilities	total equity attributable to owners of the parent
total owner's equity (or shareholders' equity)	total liabilities and owners' equity (or shareholders' equity)	other comprehensive income
other cash paid in relation to financing activities	cash paid for investments	tax refunds received
total comprehensive income attributable to minority shareholders	monetary funds	other receivables
inventory	construction work in progress	intangible assets
deferred income tax assets		

To match the features selected by each method introduced in the feature selection method, the following formula is constructed and a heat map is made in order to better reflect the correlation between the features,

$$a_{ij} = \frac{b_{ij}}{c_i} \tag{2}$$

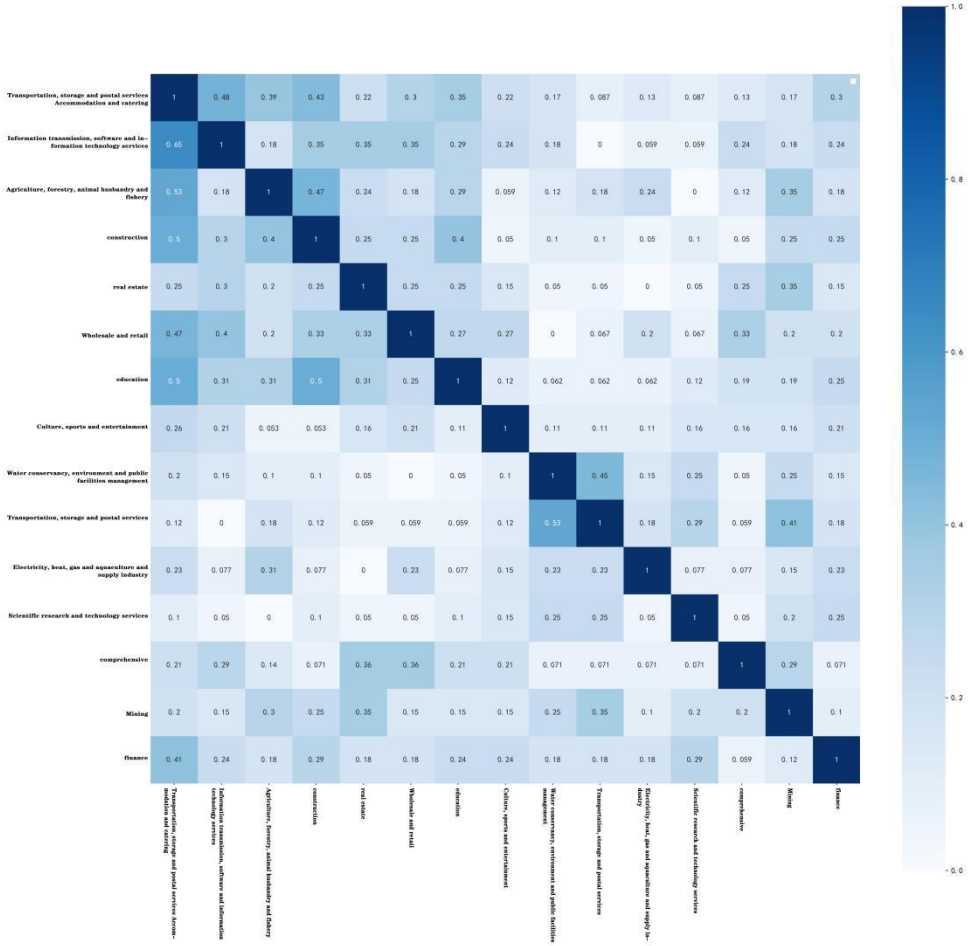

Figure 4 Heat map of correlations of characteristic factors

Where, is the feature similarity of industry i relative to industry j; is the number of identical features between industry i and industry j; is the number of specially selected features of industry i

From Figure 4, we can find that the features selected by various methods have a certain degree of overlap, which can well reflect the following two points.

1) The data indicators related to financial fraud in each industry

2) The similarities and differences of data indicators related to listed companies in different industries.

3. Hyper-parameter adjustment

In this paper, the grid search is mainly chosen and the AUC value after tuning the parameters is used as the evaluation indicator. Taking the transport, storage and postal industry as an example, various model parameters are tuned, where the initial values, ranges and tuning results set are shown in the detailed table of the results of each algorithm framework parameter(Table 6).

Table 6 Model parameters tuning

MLP Tuning			
parameter name	scope	results	post-referencing AUC
'max_iter	[100,5000]	1770	0.665094
Random_state	[1,30]	18	
hidden_ layer_sizes	[1,200]	87	
ADAboost Tuning			
parameter name	scope	results	post-referencing AUC
n_estimators	[10,1000]	70	0.721698
learning_rate	[0.01,1]	0.218421	
Random Forest Tuning			
parameter name	scope	results	post-referencing AUC
n_estimators	[10,1000]	20	0.778302
max_depth	[1,17]	11	
Logistic Regression			
parameter name	scope	results	post-referencing AUC
'max_iter	[100,1000]	640	0.613208
C	[0.001,1]	0.143714	
Rupport Vector Machine Tuning			
parameter name	scope	results	post-referencing AUC
C	[0.001,1]	0.053579	0.814725
gamma	[0.1,1]	0.816327	
Xgboost Tuning			
parameter name	scope	results	post-referencing AUC
learning_rate	[0.1,1]	0.9	0.127359
n_estimators	[10,1000]	830	
max_depth	[1,20]	6	
GBDT Tuning			
parameter name	scope	results	post-referencing AUC
learning_rate	[0.1,1]	0.3	0.857436
n_estimators	[10,1000]	220	
max_depth	[1,20]	15	
random_state	[5,30]	16	

AUC values of the optimal models for each industry

Table 7 below shows the optimal tuning models and AUC values for various industries, including eight industries assigned a value of -1, with subsequent forecasting using the weighting method.

Table 7 Optimal tuning models by industry

Industry	Model	AUC value
transport, storage and postal services	GBDT	0.857436
information transmission, software and information technology services	ADAboost	0.668750
agriculture, forestry, livestock and fisheries	GBDT	0.737500
manufacturing	GBDT	0.650000
construction	NULL	-1.000000
real estate	ADAboost	0.995763
wholesale and retail	RF	0.680723
education	NULL	-1.000000
culture, sport and entertainment	NULL	-1.000000
Water, Environment and Public Facilities Management	NULL	-1.000000
electricity, heat, gas and water production and supply	SVM	0.948718
scientific research and technical services	NULL	-1.000000
general	NULL	-1.000000
mining	NULL	-1.000000
financial services	RF	0.5800000

4. Conclusion

This paper uses a machine learning model to predict the existence of financial fraud in listed companies in various industries, with the main objective of obtaining information on companies with potential financial fraud before investing, so as to avoid losses in time. The following conclusions were eventually drawn: based on the bookkeeping data of listed companies in 19 industries, after pre-processing operations, data imbalance processing and feature selection, seven machine learning models are used for training and prediction, and the optimal model prediction is screened by comparing the evaluation results of each model with a certain degree of reasonableness.

The overall average accuracy of the best prediction model for all industries was 0.751, with an overall average AUC of 0.76486. The results of predicting the financial fraud of listed companies in the manufacturing sector in the following year: the number of financial frauds was 47. The number of non-fraudulent companies is 2613, with a reasonable falsification ratio of about 1.77%. The same applies to other industries.

In summary, the strategies and methods adopted in this paper have certain accuracy and practical significance, and can be used as an effective model to analyse the existence of financial falsification of listed companies in various industries.

References

[1] Li Hang. Statistical learning methods [M]. Beijing: Tsinghua University Press, 2012.

[2] Yuan Xianzhi, Zhou Yunpeng, Yan Chengxing, Liu Haiyang, Zeng Tu. A new approach to corporate financial fraud early warning and risk feature screening: based on artificial intelligence algorithm[C]. Proceedings of the 15th Annual China Management Conference, 2020: 709-724

[3] He Qing, Li Ning, Luo Wenjuan, Shi Zhongzhi. A review of machine learning algorithms under big data[J]. Pattern Recognition and Artificial Intelligence, 2014: 10-12

[4] Liu X.Y., Nong G.C. A comparison of several different missing value filling methods[J]. Journal of Nanning Normal College of Higher Education, 2007: 148-150

[5] Yesha. A review of missing data and its processing methods[J]. Network and Information Engineering,2017:48-50