A Comparison of Important Features for Predicting Polish and Chinese Corporate Bankruptcies

YIFAN REN
Dept. of Information, Technology, and Operations
Fordham University
New York, New York, USA
yren50@fordham.edu

GARY M. WEISS
Dept. of Computer & Information Science
Fordham University
Bronx, New York, USA
gaweiss@fordham.edu

Abstract*—This paper aims to attempt a novel method to figure out those critical financial characteristics in companies' bankruptcy and, most importantly, whether there are some differences between such classification predictions in Poland and in China. Based on classification prediction for Poland's and China's markets separately, we analyzed the feature importance. We found some common indicators in the credit risk field can always be referred to, such as Retained Earnings to Total Assets Ratio (RE/TA). Moreover, the assets elements were significant in bankruptcy in China's market, but operation situations were more critical in Poland's market.

***Keywords*—Classification, Finance, Bankruptcy, Feature Importance**

I. INTRODUCTION

Companys' bankruptcy issue has always been a hot topic in the field of accounting and finance. Nowadays, with the development of data mining, the prediction of this problem has become achievable. However, according to previous research, data mining techniques tend to be examined by their prediction performance, so that proves one or some models and algorithms perform well in the financial field. However, scholars usually spend less time on feature importance in the predictions [1]. Even though some attempts were made to explore these features, they are commonly based on the statistical and econometrical approaches, rather than data mining techniques [2]. Furthermore, the comparison between the feature importance in bankruptcy predictions in two dissimilar financial systems is even rarer to talk about.

Hence, we expect to adopt data mining techniques in such study, be means of comparison feature importance in prediction tasks in two different financial systems, to survey whether the financial factors affecting corporations' bankruptcy in two countries are similar with each other, so that provides one more localized method to study corporate finance.

Based on our two datasets, including a total of 6,591 entries, we separately conduct two predictions by various classification algorithms. The datasets have financial rates and corresponding class label that indicates bankruptcy status after one year. However, unlike other studies, we use various models here not only for high predicted performance, but also to find sets of important features from the most reliable model as its excellent prediction results. So we only choose those algorithms which can report feature importance. As far as those black-box algorithms, like Neural Network, they will not engage in this paper due to their high computing power need for getting feature importance, even if they are available [3].

In the light of feature importance from the above experiments, we can investigate the regional divergence of the financial characteristics in firms' bankruptcy. Several researchers have utilized such feature selections and ranking methods but not in the financial field [4].

II. BACKGROUND

This section will provide the necessary background into the definition and role of bankruptcies in Poland and China and will also discuss the most relevant financial factors.

Bankruptcy is a legal process through which people or other entities who cannot repay debts to creditors may seek relief from some or all of their debts. It is an essential lawful definition but cannot be further expanded as different nations or regions do not have the same explanation because of different legal provisions. However, the financial field, compared to definitions in law, usually regard firms as bankrupt status if insolvency appears, which means the firms are no longer able to repay their liabilities. It can be commonly used to conduct studies despite kinds of unique law rules all around the world [5].

In terms of some financial indicators, assets, liabilities, and cash all have an important influence in the financial status of the company, and hence play a critical role in bankruptcies. Nonetheless, sometimes people can see these indicators are leading to bankruptcy only when the bankruptcy occurs. Thus, this is also precisely the significance of our study. Financial warning beforehand is that investors and shareholders care most about.

III. PREDICTING POLISH CORPORATE BANKRUPTCIES

In this section, we will introduce how our experiments were conducted separately in Poland's and China's dataset, including data collection, data preprocessing, prediction and evaluation, besides, the most vital things in our study, the feature importance.

A. Data Description

This dataset, named "Polish companies bankruptcy data Data Set," was collected from the UC Irvine Machine Learning Repository, and created by Sebastian Tomczak at https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data[1][6]. The data was collected from Emerging Markets Information Service, which was a database containing information on emerging markets around the world. The

bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013. Nevertheless, in order to predict whether a company will be bankrupt in the next year, we only picked the 5th dataset for our prediction, which contains companies' financial information in the *t-1* year and their bankrupt status in the *t* year. This dataset was entitled "5thYear" in original format on the UCI Machine Learning Repository, which contained financial rates from the 5th year of the forecasting period and corresponding class label that indicates bankruptcy status after one year.

The data contained 5910 instances (financial statements), 410 represented bankrupted companies, 5500 firms that did not bankrupt in the forecasting period. It meant that the bankrupted instances occupied 6.937%, and the others occupied 93.063%, of the aggregate data. Furthermore, 64 attributes related to were included, as shown below in Table I.

However, the dataset had some missing values so that we firstly should handle them by imputation with the mean value of each column, except Attribute 37. As almost half of the records had missing value in Attribute 37, so turning this variable into a binary class may make more sense. We also identify the target class as "B" (initially being 1, meaning bankrupt) and "S" (initially being 0, meaning solvent). To avoid potential influence in feature importance, we would not take further feature engineering by dropping or dimensionality reduction like principal components analysis, as our purpose was not the high prediction performance itself.

B. Modeling

In order to complete our experiment successfully, we would utilize the scikit-learn to conduct our modeling steps. Scikit-learn exposes a wide variety of machine learning algorithms, both supervised and unsupervised, using a consistent, task-oriented interface, thus enabling easy comparison of methods for a given application. Since it relies on the scientific Python ecosystem, it can easily be integrated into applications outside the traditional range of statistical data analysis [7].

The goal of the experiment was to identify the best classification model, and based on the best model, explore feature importance. Hence, we should consider computing effectiveness when getting feature importance as well. Eventually, we selected the following methods in our experiment:

Logistic Regression (LR);

C4.5, decision tree model (DT) [8];

Boosted trees trained with Extreme Gradient Boosting (XGBoost) [9];

Random Forest (RF) [10]

The four models tend to perform well steadily in binary class prediction tasks. We would generate four different classifiers based on these algorithms directly supported by a library of scikit-learn in Python. However, the XGBoost model was not included in scikit-learn, so we also used the xgboost library to generate an XGB classifier [9].

TABLE I. DESCRIPTION OF ATTRIBUTES IN POLAND DATASET

ID	Description	ID	Description
X1	net profit / total assets	X33	operating expenses / short-term liabilities
X2	total liabilities / total assets	X34	operating expenses / total liabilities
X3	working capital / total assets	X35	profit on sales / total assets
X4	current assets / short-term liabilities	X36	total sales / total assets
X5	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365	X37	(current assets - inventories) / long-term liabilities
X6	retained earnings / total assets	X38	constant capital / total assets
X7	EBIT / total assets	X39	profit on sales / sales
X8	book value of equity / total liabilities	X40	(current assets - inventory - receivables) / short-term liabilities
X9	sales / total assets	X41	total liabilities / ((profit on operating activities + depreciation) * (12/365))
X10	equity / total assets	X42	profit on operating activities / sales
X11	(gross profit + extraordinary items + financial expenses) / total assets	X43	rotation receivables + inventory turnover in days
X12	gross profit / short-term liabilities	X44	(receivables * 365) / sales
X13	(gross profit + depreciation) / sales	X45	net profit / inventory
X14	(gross profit + interest) / total assets	X46	(current assets - inventory) / short-term liabilities
X15	(total liabilities * 365) / (gross profit + depreciation)	X47	(inventory * 365) / cost of products sold (COGS)
X16	(gross profit + depreciation) / total liabilities	X48	EBITDA (profit on operating activities - depreciation) / total assets
X17	total assets / total liabilities	X49	EBITDA / sales
X18	gross profit / total assets	X50	current assets / total liabilities
X19	gross profit / sales	X51	short-term liabilities / total assets
X20	(inventory * 365) / sales	X52	(short-term liabilities * 365) / COGS
X21	sales (n) / sales (n-1)	X53	equity / fixed assets
X22	profit on operating activities / total assets	X54	constant capital / fixed assets
X23	net profit/sales	X55	working capital
X24	gross profit (in 3 years) / total assets	X56	(sales - COGS) / sales
X25	(equity - share capital) / total assets	X57	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
X26	(net profit + depreciation) / total liabilities	X58	total costs /total sales
X27	profit on operating activities / financial expenses	X59	long-term liabilities / equity
X28	working capital / fixed assets	X60	sales / inventory
X29	logarithm of total assets	X61	sales / receivables
X30	(total liabilities - cash) / sales	X62	(short-term liabilities *365) / sales
X31	(gross profit + interest) / sales	X63	sales / short-term liabilities
X32	(current liabilities * 365) / COGS	X64	sales / fixed assets

The tool provided by scikit-learn was used to randomly split our dataset into a training set and testing set, given the test size of 30%. After partition, the training set had 4137 entries, and the testing set had 1773 instances.

Compared to default parameters, we adjusted the class weight in all the tree models. Because in the real-world, predicting a company with bankruptcy as non-bankrupt, is worse than predicting a company without bankruptcy as bankrupt. Thus, in our experiment, we set the weight of class "B" to five times that of class "S" to make the model more oriented in the direction we wanted. Meanwhile, we used to 10-fold cross-validation method to avoid possible over-fitting.

C. Prediction Results

After training on the training set, we got a trained model to conduct prediction in our testing set, then evaluated and got tables shown below. We used R-squared, precision rate, recall rate, F1-Score, accuracy.

TABLE II. EVALUATION TABLES (POLAND)

Index	LR	DT	RF	XGBoost
Precision (Class "B")	0.56	0.60	0.80	0.93
Precision (Class "S")	0.94	0.97	0.96	0.97
Recall (Class "B")	0.12	0.58	0.32	0.56
F1-Score (Class "B")	0.20	0.59	0.46	0.70
Overall Accuracy (Weighted)	0.94	0.95	0.95	0.97

Based on our experiment results, it was accepted that the XGBoost model performed best as its highest Accuracy. It also performed better in prediction in class1. So, we would choose XGBoost as our next steps' basis.

Table III was the confusion matrix of the XGBoost model's performance on the testing set.

TABLE III. CONFUSION MATRIX OF XGBOOST MODEL (POLAND)

	Predicted Labels		
True Labels		Class "B"	Class "S"
	Class "B"	64	50
	Class "S"	5	1654

D. Feature Importance

Generally speaking, the feature importance score measures the value of features in boosting decision tree construction. The more an attribute is used to build a decision tree in a model, the more critical it is.

Attribute importance is calculated by sorting each attribute in the data set. In a single decision tree, the attribute importance is calculated by the amount of each property split point to improve the performance measurement, and the node is responsible for weighing and recording the number of times. That is to say, the larger an attribute's improvement performance measure for the split point (the closer it is to the root node), the higher the weight; selected by the more the boosting trees, the more critical the attribute is.

Finally, the results of one attribute in all the boosting trees are weighted and summed and then averaged to obtain the importance score. In our XGBoost model, we got the top 10 features shown below, based on their importance scores.

TABLE IV. FEATURE IMPORTANCE SCORES (POLAND)

Rank	ID	Importance Score	Description
1	X22	0.1052	profit on operating activities / total assets
2	X35	0.0752	profit on sales / total assets
3	X41	0.0482	total liabilities / ((profit on operating activities + depreciation) * (12/365))
4	X34	0.0412	operating expenses / total liabilities
5	X26	0.0393	(net profit + depreciation) / total liabilities
6	X5	0.0389	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365
7	X46	0.0321	(current assets - inventory) / short-term liabilities
8	X21	0.0302	sales (n) / sales (n-1)
9	X39	0.0294	profit on sales / sales
10	X6	0.0286	retained earnings / total assets

IV. PREDICTING CHINESE CORPORATE BANKRUPTCIES

A. Data Description

This dataset was merged from 3 smaller datasets, manually collected from Wind, a financial information services company in China We queried the financial information by its financial terminal one firm by one firm, and download the results as our datasets. One was bankruptcy dataset, containing 61 listed companies in China stock market, finally going bankrupt since 2006. The second dataset was from SH380 index firms, and the third one was from SZ300 index firms. The number of non-bankruptcy firms was 620. Hence, our dataset had 681 entries in total, in which 8.957% were bankrupted instances, and 91.043% were non-bankrupted ones, which the class "B" meant bankrupt, and "S" meant solvent. The corresponding attributes information was the last year of their bankruptcy or the newest data if non-bankruptcy. All the data were collected from Wind Financial Database, which was the most prominent financial data provider in China.

We got 80 attributes, including firms' necessary information and accounting data, as shown in Table V.

We filled in the missing values in each column with its mean. However, unlike Poland dataset, X43 and X84 were dropped because of their missing value caused by no net debt, leading to no denominator. X1 and X2 were removed as well, as they are the code and name of firms, not useful for our task.

B. Modeling

We still selected the same models we used before on Poland dataset: LR, DT [8], XGBoost [9], and RF [10], and the same tool to make partition [7]. However, we chose the test size as 50% this time, because the China dataset was not large enough with very few bankrupt instances, we would like to have enough negative instances in test set to evaluate the model better. After partition, the training set had 340 entries, and the testing set had 341 instances.

TABLE V. DESCRIPTION OF ATTRIBUTES IN CHINA DATASET

ID	Description	ID	Description
X1	Company Code	X43	Tangible assets / net debt
X2	Company Name	X44	Capital expenditure / depreciation and amortization
X3	Return on equity (ROE) (average)	X45	Cash received for sales of goods and services / Operating income (TTM)
X4	ROE (deducted / average)	X46	Net cash flow from operating activities / operating income (TTM)
X5	Return on Total Assets (ROA)	X47	Net cash flow from operating activities / operating profit (TTM)
X6	ROA (net)	X48	Proportion of net cash flow from operating activities
X7	Return on Human Investment (RHI)	X49	Proportion of net cash flow from investment
X8	ROE (annualized)	X50	Proportion of net cash flow from fundraising
X9	ROA (annualized)	X51	Net operating cash flow / total operating income
X10	ROA (net) (annualized)	X52	Cash operation index
X11	Sales margin	X53	Cash recovery rate of all assets
X12	Gross profit margin	X54	Asset-Liability Ratio
X13	Cost of sales ratio	X55	Asset-Liability Ratio (excluding advance receipts)
X14	Sales Period Expense Rate	X56	Asset-Liability Ratio (excluding advance receipts) (Announcement Based)
X15	Main business ratio	X57	Long-term debt ratio
X16	Net profit / Total operating income	X58	Long-term asset fit ratio
X17	Operating profit / total operating income	X59	Tangible assets / total assets
X18	EBIT / Total operating income	X60	Non-current debt ratio
X19	Total operating cost / Total operating income	X61	Current ratio
X20	Management expenses / Total operating income	X62	Current liabilities / total liability
X21	Financial costs / total operating income	X63	Capitalization ratio
X22	Asset impairment loss / total operating income	X64	Quick ratio
X23	ROA (Trailing Twelve Months (TTM))	X65	Conservative Quick Ratio
X24	ROA (net) (TTM)- Excluding Minority Shareholder Profit and Loss	X66	Cash ratio
X25	Return on invested capital (ROIC)	X67	Net Asset-Liability Ratio
X26	ROIC (TTM)	X68	Net debt ratio
X27	EBIT/ Total assets (TTM)	X69	Total equity attributable to shareholders of the parent company / liabilities
X28	Net sales margin (TTM)	X70	EBITDA / Total Liability
X29	Gross profit margin (TTM)	X71	Net cash flow from operating activities / Total Liability
X30	Sales Period Expense Rate (TTM)	X72	Net cash flow from operating activities / Current liabilities
X31	Operating profit / Total operating income (TTM)	X73	Net cash flow from operating activities / Non-current liabilities
X32	Total Operating Cost / Total Operating Revenue (TTM)	X74	Non-financing net cash flow / Non-current liabilities
X33	Operating profit / operating income (TTM)	X75	Non-financing net cash flow / Total Liability
X34	Tax / Total Profit (TTM)	X76	Proportion of long-term debt
X35	Net profit attributable to shareholders of the parent company / Operating income (TTM)	X77	Working capital / total assets
X36	Asset impairment loss / total operating income (TTM)	X78	Tangible net worth debt ratio
X37	Asset impairment loss / operating profit	X79	Retained earnings / total assets
X39	Operating profit / total profit (TTM)	X80	EBIT (TTM) / total assets
X39	Total profit / operating income (TTM)	X81	Total market value / liabilities for the day
X40	Cash received from sales of goods and services provided/operating income	X82	Total shareholders' equity (including minority) / total liabilities
X41	Net cash flow from operating activities / operating income	X83	Operating income / total assets
X42	Net profit cash rate	X84	Net cash flow from operating activities / Net debt

In this step, we continued to use our previous parameter settings. The weight of class "B" was higher five times than that of class "S" and 10-fold cross-validation. The class weight worked when the algorithms calculated the gain ration; it gave a more extensive adjusted index for the class with fewer instances and a smaller index for another class. The validation method is not different from the traditional method, because although the data were collected from different years, the chronological relations between independent and target variables were stable.

C. Prediction Results

We got the same format tables shown below as Poland dataset so that we could also know about how our models did on China dataset.

TABLE VI. EVALUATION TABLES (CHINA)

Index	LR	DT	RF	XGBoost
Precision (Class "B")	0.65	0.74	1.0	0.96
Precision (Class "S")	0.98	1.0	0.99	0.99
Recall (Class "B")	0.79	1.0	0.90	0.93
F1-Score (Class "B")	0.71	0.85	0.95	0.95
Accuracy (Overall)	0.95	0.97	0.99	0.99

As the model had the highest Accuracy, RF also performed very well in prediction class 1. Hence, we would choose RF for the next steps.

Table VII was the confusion matrix of the RF model's performance on the testing set.

TABLE VII. CONFUSION MATRIX OF XGBOOST MODEL (CHINA)

	Predicted Labels		
True Labels		Class "B"	Class "S"
	Class "B"	26	3
	Class "S"	0	312

D. Feature Importance

Due to the inherent randomness of the random forest, the model may give different importance weights to the features each time. By training the model multiple times, that is, each time by selecting a certain number of features and retaining the intersection of the previous features, this cycle is repeated a certain number of times, so we can finally get a certain number of features that have a significant t contribution to the impact of classification tasks.

However, some features had a strong correlation with each other. For instance, X55, Asset-Liability Ratio (excluding advance receipts), was very similar to and related to X54, Asset-Liability Ratio. Thus, in these cases, we kept only the higher-ranked one and got a new adjusted rank.

TABLE VIII. FEATURE IMPORTANCE SCORES (CHINA)

Rank	ID	Importance Score	Description	Adjusted Rank
1	X66	0.0630	Cash ratio	1
2	X77	0.0477	Working capital / total assets	2
3	X79	0.0445	Retained earnings / total assets	3
4	X6	0.0432	ROA (net)	4
5	X69	0.0426	Total equity attributable to shareholders of the parent company/liabilities	5
6	X54	0.0415	Asset-Liability Ratio	6
7	X24	0.0398	ROA (net) (TTM)-Excluding Minority Shareholder Profit and Loss	-
8	X55	0.0378	Asset-Liability Ratio (excluding advance receipts)	-
9	X82	0.0375	Total shareholders' equity (including minority) / total liabilities	7
10	X56	0.0319	Asset-Liability Ratio (excluding advance receipts) (Announcement Based)	-
11	X17	0.0299	Operating profit / total operating income	8
12	X61	0.0292	Current debt equity ratio	9
13	X8	0.0244	ROE (annualized)	10

V. COMPARISON OF FEATURE IMPORTANCE

Based on the above experiment, we could start to compare the features importance between the two datasets. When We had a look at these attributes, it was feasible to find some related features among these critical features in each dataset. In Table IX, we figured out some relevant features in the two datasets, ranked at least the top 15 in feature importance.

TABLE IX. PAIRS OF RELEVANT FEATURES

Poland		China	
Description	*ID*	*ID*	*Description*
retained earnings / total assets	X6	X79	Retained earnings / total assets
(current assets - inventory) / short-term liabilities	X46	X61	Current ratio
profit on operating activities / total assets	X22	X7	EBIT / total assets
working capital / fixed assets	X28	X77	Working capital / total assets

If we only discussed direct relationships, X6 (Poland) and X79 (China) were both one of 10 features in their ranking list. Moreover, they were the same feature, retained earnings / total assets.

Other pairs, although they did not have such robust similarities, their relationship was still impressive. X46 in Poland's dataset was similar to X61 in China's dataset, as the current ratio was current assets / short-term liabilities. In this case, their difference was that the inventory was subtracted in the numerator in Poland's X46. As far as the pair of X22 and X7, EBIT was earnings before interest and taxes, containing profit on not only the operating activities but also other profitable activities. Thus, they commonly did not make much difference. In terms of the pair of X27 and X77, there was little difference in the denominator, which one was the fixed assets, and another one was total assets. Nevertheless, they were relatively similar. This similarity was mainly reflected in their numerator, working capital, which was not a general financial indicator.

Also, Their different important attributes allowed us to obtain some interesting findings. Features related to assets or equity appeared eight times in the ranking list of China, but only four times in the list of Poland. Operation-related elements, like profit or income, appeared in only one attribute in the list of China, but in 5 attributes in the list of Poland.

Hence, in China's market, assets played an important role in bankruptcy prediction because many state-owned enterprises (SOE) with a large scale of assets, which was the main part of China's market, did not smoothly go bankrupt. Moreover, operating-related attributes also did not lead to whether it would have bankruptcy. SOEs generally had more assets, and at the same time, lower financing costs [11], it meant whether they had high profits, they had more possibility to get a loan to maintain operation to avoid bankruptcy. On the contrary, in Poland, whether it had an excellent operating situation may become an important factor affecting bankruptcy.

VI. RELATED WORK

In earlier research, a logistic regression model was used to predict financial distress. However, it only selected six independent variables, earnings growth rate, ROA, current ratio, long-term debt to shareholder equity ratio, working capital to total assets ratio, and asset turnover. Furthermore, the final result was as the following table.

TABLE X. CONFUSION MATRIX

True Label		Predicted Label		Total	Error Rate (%)
		0 (No Distress)	*1(Distress)*		
Count	*0*	66	4	70	6.47
	1	5	64	69	
Normalized (%)	*0*	94.29	5.71	100	
	1	7.25	92.75	100	

This study had achieved excellent prediction performance already, but the researchers did not explore further feature importance in the prediction. They chose to conduct a univariate analysis and found the performance of ROA (net) was the best one [2].

And we could find that these features playing roles in the previous study, such as ROA and working capital to total assets ratio, were still significant in our predictions, like X6 (China) and X77 (China). Moreover, the asset turnover was very similar to X22 (Poland) and X35 (Poland). The current ratio was mainly the X46 (Poland).

Next, let us take a look at M. Zieba and his co-workers' previous research on the Poland dataset. When they analyzed the feature importance, they got a table shown below [1].

TABLE XI. FEATURE IMPORTANCE SCORES (POLAND) (ZIEBA) [1]

Ranking	ID	Importance Score
1	X25	0.0627
2	X22	0.0480
3	X27	0.0379
4	X15	0.0356
5	X52	0.0326
6	X53	0.0284
7	X14	0.0248
8	X40	0.0247
9	X42	0.0238
10	X36	0.0236

The X22, profit on operating activities / total assets, which had the highest importance score in our prediction, was still ranked very high. Furthermore, it appeared in the logistic regression model we introduced as well.

However, except X22, there were no features in Zieba's list appearing in our Poland list. The difference was probably caused by that Zieba and his co-authors calculated the importance score based on multiple models. However, we only relied on the model having the best performance.

VII. CONCLUSION

This paper, according to two classification experiments in Poland and China datasets, studied in feature importance in the best model of each dataset, so that discussed the similarity and dissimilarity between financial index in different markets. Given the result, although there are objective differences between different markets, some common indicators in the credit risk field can always be referred, such as ROA and Retained Earnings to Total Assets Ratio (RE/TA). Moreover, the assets elements were significant in bankruptcy in China's market, but operation situations were more critical in Poland's market.

Nevertheless, our research still has many deficiencies. We only considered the financial index of the year before the bankruptcy. There was no longer-term or periodic consideration. Meanwhile, we have only selected listed companies in the China dataset, which may cause some bias. This data set is also not large enough, and we expect to experiment with a larger dataset in the future. At the same time, our study only stays on the surface phenomenon without more in-depth exploration. These are the directions that can be improved in future research.

In general, we hope to provide a novel perspective to explore corporate finance and hope this will inspire future research. Data mining does not necessarily shine in the asset pricing field. It can be used in corporate finance as well.

ACKNOWLEDGMENT

I want to thank my professor Gary Weiss for his contribution to building my data mining knowledge, my friend Yi Wang for her helpful work obtaining China's dataset and financial characteristics.

REFERENCES

[1] M. Zieba, S. Tomczak, J. Tomczak, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction," Expert Systems with Applications, vol. 58, pp. 93-101, Apr 2016.

[2] S. Wu, X. Lu, "A Study of Models for Predicting Financial Distress in China's Listed Companies," Economic Research Journal, vol. 36, pp. 4, Jun 2001.

[3] R. Setiono, H Liu, "Neural-network feature selector," IEEE Transactions on Neural Networks, vol. 8, pp. 654-662, May 1997.

[4] R. Genuer, J. Poggi, C. Tuleau-Malot, "Variable selection using random forests," Pattern Recognition Letters, vol. 31, pp. 2225-2236, Oct 2010

[5] C. Lu, L. Xu, L. Zhou, "Comparative Analysis of Corporate Financial Distress and Financial Bankruptcy," Economic Research Journal, vol. 39, pp. 64-73, Aug 2004.

[6] D. Dua and C. Graff, "UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]," Irvine, CA, USA.

[7] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, Oct 2011.

[8] J. R. Quinlan, C4. 5: programs for machine learning, 1st ed. Morgan Kaufmann, 2014.

[9] T. Chen and C. Guestrin, "XGBoost," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16, San Francisco, CA, USA, Aug 2016, pp. 758-794.

[10] T. K. Ho. "Random decision forests," in Proceedings of 3rd international conference on document analysis and recognition, Montreal, Quebec, Canada, 1995, pp. 278-282.

[11] G. Ferri, L. Liu, "Honor thy creditors beforan thy shareholders: are the profits of Chinese state-owned enterprises real?," Asian Economic Papers, vol. 9, pp. 50-71, Oct 2010.