A Study on Comparison between Important Features when Predicting Corporates Bankruptcy in Poland's and China's Markets

YIFAN REN
Fordham University
New York, NY, USA
yren50@fordham.edu

***Abstract*—This paper aims to attempt a novel method to figure out those critical financial characteristics in companies' bankruptcy, and, the 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 and found some common indicators in the credit risk field can always be referred, such as Return on Assets (ROA) and Retained Earnings to Total Assets Ratio (RE/TA). And the assets elements were significant in bankruptcy in China's market, but operation situations were more important in Poland's market.**

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

I. INTRODUCTION

Company's 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 prove one or some models and algorithms perform well in financial field, but the 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 more rare 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 provide one more localized method to study corporate finance.

Based on our two datasets including totally 6,591 entries, we separately conduct two predictions by various classification algorithms. The datasets have contains 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, bust also to find sets of important features from the most reliable model as its excellent prediction results. So we only choose those algorithms which are able to report feature importance clearly. As far as those black-box algorithms, like Neural Network, they will not engage in this paper due to their expensive 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. Such feature selections and ranking methods have been utilized by several academia but not in financial field [4].

II. BACKGROUND

A. Introduction to Bankruptcy

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. In most jurisdictions, bankruptcy is imposed by a court order, often initiated by the debtor. It is an essential lawful definition but cannot be further expanded as actually different nations or regions do not have the same exact explanation because of different legal provisions.

As far as the details in two countries we will have exemplified are concerned, in Poland, the bankruptcy is declared against a debtor who has become insolvent, a debtor shall be insolvent if he has lost the ability to fulfill his matured pecuniary liabilities (Polish Parliament, 2003); in China, bankruptcy is a situation where, for a company, if the company cannot pay off its debts due and the assets are not sufficient to pay off its entire debt (China Congress, 2006).

However, financial field, compared to definitions in law, usually simply 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 of kinds of unique law rules all around the world [5].

B. Serveral Significant Financial Factors

Assets, liabilities and cash all play important roles in Financial status of the company.

As we introduced before, the simplest bankrupt circumstance is insolvency. Thus, the relationship between total assets and total liabilities is critical. If the total assets are less than total liabilities, namely the insolvency appears, we can say

the company has been bankrupt. The company must always pay attention to their asset-liability ratio to avoid insolvency.

On the other hand, sometimes the bankruptcy starts from shortage of cash. A company is impossible to borrow all the money at once and return all of them at another time. It is more common to present a rolling process, a company tend to borrow and repay at different multiple times. However, if a company does not have enough cash for one debt, even though the debt may be a very small sum of money or equivalent, negative chain reactions can happen and lead to eventual bankruptcy of the company. It is so-called capital chain rupture.

Nonetheless, sometimes only when the bankruptcy occurs you can see these obvious circumstances happen. This is also exactly the significance of our study. Financial warning beforehand is that investors and shareholders care most about.

III. Experiment

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, in addition, most vital things in our study, the feature importance.

A. Bankruptcy Prediction based on the Poland's Dataset

1) Data Collection and Description

This dataset was collected from UC Irvine Machine Learning Repository, and created by Sebastian Tomczak. 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. But 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.

The data contained 5910 instances (financial statements), 410 represented bankrupted companies, 5500 firms that did not bankrupt in the forecasting period. And 64 attributes related to were included, as shown below in Table I.

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
X16	(gross profit + depreciation) / total liabilities	X48	EBITDA (profit on operating activities - depreciation) / total assets
X17	total assets / total liabilities	X49	EBITDA (profit on operating activities - depreciation) / 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) / cost of products sold)
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 - cost of products sold) / 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) / cost of products sold	X64	sales / fixed assets

2) Data Preprocessing

a) Missing value processing

The dataset had some missing values so that we firstly should handle them. The most missing values appearing in Attribute 37, which is (current assets - inventories) / long-term liabilities, had 2548 rows of data. This number was close to half of the total, so we could not simply impute these missing values or delete records with missing value. We decided to turn this variable into a binary class, having true value or having missing value. As for other missing values, we filled in the missing values in each column with the its mean.

b) Data conversion

Nevertheless, the class variable was still in string format like "b'0'" and "b'1'", we should also convert them to integer to look suitable, as 0 and 1. The class 0 meant no bankruptcy, the class 1 means bankruptcy.

Because the object we studied was these features, we would not drop any columns like ordinary feature engineering based on direct business sense or some statistical measurement. Data dimensionality reduction methods, like Principal Components Analysis, would not be adopted as well, because they would mask the features themselves, and, our purpose was not high prediction performance itself.

c) Data partition

The tool provided by scikit-learn was used to randomly split our dataset into 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.

3) Modeling

a) Model Selection

The goal of the experiment was to identify the best classification model and furthermore based on the best model explore feature importance. Hence, we should consider the 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) [6];

Boosted trees trained with Extreme Gradient Boosting (XGBoost);

Random Forest (RF) [7]

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 library of scikit-learn in Python. But XGBoost model was not included in scikit-learn, so we also used the xgboost library to generate XGB classifier.

b) Model Training

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 absolutely worse than predicting a company without bankruptcy as bankrupt. Thus, in our experiment, we set the weight of class 0 to five times that of class 1 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.

4) Evaluation and Output

a) Prediction and Evaluation

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 Root Mean Squared Error (RMSE), R Squared (R2), precision rate, recall rate, F1-Score, accuracy and average cross validation score (ACVS).

TABLE II. EVALUATION TABLES 1 (POLAND)

Class	Precision	Recall	F1-Score	Support
LR				
0	0.93	1.0	0.96	1641
1	0.54	0.05	0.10	132
Macro Average	0.73	0.52	0.53	1773
Weighted Average	0.90	0.93	0.90	1773
DT				
0	0.97	0.97	0.97	1641
1	0.58	0.58	0.58	132
Macro Average	0.77	0.77	0.77	1773
Weighted Average	0.94	0.94	0.94	1773
RF				
0	0.94	0.99	0.97	1641
1	0.72	0.21	0.33	132
Macro Average	0.83	0.60	0.65	1773
Weighted Average	0.92	0.94	0.92	1773
XGBoost				
0	0.94	0.99	0.97	1641
1	0.72	0.21	0.33	132
Macro Average	0.83	0.60	0.65	1773
Weighted Average	0.92	0.94	0.92	1773

TABLE III. EVALUATION TABLES 2 (POLAND)

Model	RMSE	R2	Accuracy	ACVS
LR	0.27	-0.07	0.93	0.93
DT	0.25	0.09	0.94	0.95
RF	0.25	0.06	0.94	0.94
XGBoost	0.19	0.46	0.96	0.96

Based on our experiment results, it was obviously accepted that the XGBoost model perform best as its lowest RMSE and highest R2 and Accuracy. It also performed better in prediction in class1. So, we would choose XGBoost as our next steps' basis.

The Figure 1 was the confusion matrix of XGBoost model's performance on testing set.

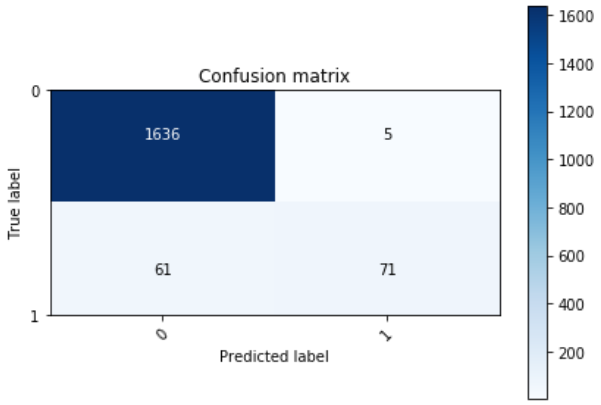

Fig. 1. Confusion Matrix of XGBoost Model (Poland)

b) 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 important 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 weighting 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 greater the weight; selected by the more the boosting trees, the more important 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)

Ranking	ID	Importance Score
1	X22	0.1052
2	X35	0.0752
3	X41	0.0482
4	X34	0.0412
5	X26	0.0393
6	X5	0.0389
7	X46	0.0321
8	X21	0.0302
9	X39	0.0294
10	X6	0.0286

B. *Bankruptcy Prediction based on the China's Dataset*

1) *Data Collection and Description*

This dataset was merged from 3 smaller 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. The corresponding attributes information were 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 biggest financial data provider in China.

Please be careful about the difference between the target variables in Poland and China datasets. In Poland dataset, class 0 meant non-bankruptcy, but in China dataset, class 0 meant bankruptcy.

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

TABLE V. DESCRIPTION OF ATTRIBUTES IN CHINA DATASET

ID	Description	ID	Description
X1	Company Code	X44	Capital expenditure / depreciation and amortization
X2	Company Name	X45	Cash received for sales of goods and services / Operating income (TTM)
X3	Return on equity (ROE) (average)	X46	Net cash flow from operating activities / operating income (TTM)
X4	ROE (deducted / average)	X47	Net cash flow from operating activities / operating profit (TTM)
X5	Return on Total Assets (ROA)	X48	Proportion of net cash flow from operating activities
X6	ROA (net)	X49	Proportion of net cash flow from investment
X7	Return on Human Investment (RHI)	X50	Proportion of net cash flow from fundraising
X8	ROE (annualized)	X51	Net operating cash flow / total operating income
X9	ROA (annualized)	X52	Cash operation index
X10	ROA (net) (annualized)	X53	Cash recovery rate of all assets
X11	Sales margin	X54	Asset-Liability Ratio
X12	Gross profit margin	X55	Asset-Liability Ratio (excluding advance receipts)
X13	Cost of sales ratio	X56	Asset-Liability Ratio (excluding advance receipts) (Announcement Based)
X14	Sales Period Expense Rate	X57	Long-term debt ratio
X15	Main business ratio	X58	Long-term asset fit ratio
X16	Net profit / Total operating income	X59	Tangible assets / total assets
X19	Total operating cost / Total operating income	X62	Current liabilities / total liability
X20	Management expenses / Total operating income	X63	Capitalization ratio

X21	Financial costs / total operating income	X64	Quick Ratio
X22	Asset impairment loss / total operating income	X65	Conservative Quick Ratio
X23	ROA (Trailing Twelve Months (TTM))	X66	Cash ratio
X24	ROA (net) (TTM)-Excluding Minority Shareholder Profit and Loss	X67	Net Asset-Liability Ratio
X25	Return on invested capital (ROIC)	X68	Net debt ratio
X26	ROIC (TTM)	X69	Total equity attributable to shareholders of the parent company / liabilities
X27	EBIT/ Total assets (TTM)	X70	EBITDA / Total Liability
X28	Net sales margin (TTM)	X71	Net cash flow from operating activities / Total Liability
X29	Gross profit margin (TTM)	X72	Net cash flow from operating activities / Current liabilities
X30	Sales Period Expense Rate (TTM)	X73	Net cash flow from operating activities / Non-current liabilities
X31	Operating profit / Total operating income (TTM)	X74	Non-financing net cash flow / Non-current liabilities
X32	Total Operating Cost / Total Operating Revenue (TTM)	X75	Non-financing net cash flow / Total Liability
X33	Operating profit / operating income (TTM)	X76	Proportion of long-term debt
X34	Tax / Total Profit (TTM)	X77	Working capital / total assets
X35	Net profit attributable to shareholders of the parent company / Operating income (TTM)	X78	Tangible net worth debt ratio
X36	Asset impairment loss / total operating income (TTM)	X79	Retained earnings / total assets
X37	Asset impairment loss / operating profit	X80	EBIT (TTM) / total assets
X39	Operating profit / total profit (TTM)	X81	Total market value / liabilities for the day
X39	Total profit / operating income (TTM)	X82	Total shareholders' equity (including minority) / total liabilities
X40	Cash received from sales of goods and services provided / operating income	X83	Operating income / total assets
X41	Net cash flow from operating activities / operating income	X84	Net cash flow from operating activities / Net debt
X42	Net profit cash rate		
X43	Tangible assets / net debt		

2) Data Preprocessing

a) Missing value processing

The China dataset also had some missing values so that we firstly should handle them. However, some columns had many missing values, which X43 and X84 both had 313 missing values. When we looked at them in details, we could find they were related to net debt. It may be caused by no net debt, the net debt was zero, but zero is never the denominator. We decided to drop these columns, as we had a large enough number of columns already. As for other missing values, we filled in the missing values in each column with the its mean as same as the Poland dataset.

b) Data conversion

Like what we did on Poland dataset, we would not like to make further feature engineering except the two columns we dropped just now. But the two columns, X1 and X2, which were the code and name of firms, should be removed, as they were not useful for our task.

c) Data partition

We still used the same tool to make partition. 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.

In our RF model, we got the top 10 features shown below, based on their importance scores.

3) Modeling

a) Model Selection

We still selected the same models we used before on Poland dataset:

Logistic Regression (LR);

C4.5, decision tree model (DT) [6];

Boosted trees trained with Extreme Gradient Boosting (XGBoost);

Random Forest (RF) [7]

b) Model Training

In this step, we continued to use our previous parameter settings. The weight of class 0 to five times that of class 1 and 10-fold cross validation.

4) Evaluation and Output

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 1 (CHINA)

Class	Precision	Recall	F1-Score	Support
LR				
0	0.65	0.79	0.71	29
1	0.98	0.96	0.97	312
Macro Average	0.81	0.87	0.84	341
Weighted Average	0.95	0.95	0.95	341
DT				
0	0.74	1.0	0.85	29
1	1.0	0.97	0.98	312
Macro Average	0.87	0.98	0.92	341
Weighted Average	0.98	0.97	0.97	341
RF				
0	1.0	0.90	0.95	29
1	0.99	1.0	1.0	312
Macro Average	1.0	0.95	0.97	341
Weighted Average	0.99	0.99	0.99	341
XGBoost				
0	0.96	0.93	0.95	29
1	0.99	1.0	1.0	312
Macro Average	0.98	0.96	0.97	341
Weighted Average	0.99	0.99	0.99	341

TABLE VII. EVALUATION TABLES 2 (CHINA)

Model	RMSE	R2	Accuracy	ACVS
LR	0.23	0.30	0.95	0.96
DT	0.17	0.61	0.97	0.95
RF	0.09	0.89	0.99	0.98
XGBoost	0.09	0.88	0.99	0.97

As the model had lowest RMSEs and highest R2 and Accuracy, RF also perform very well in prediction class 1. Hence, we would choose RF for next steps.

The Figure 2 was the confusion matrix of RF model's performance on testing set.

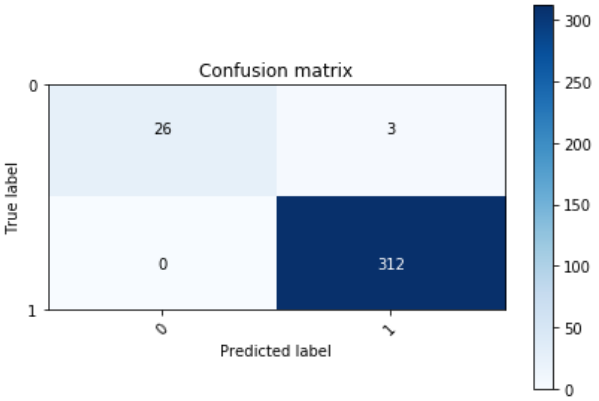

Fig. 2. Confusion Matrix of Random Forest Model (China)

b) 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 an important contribution to the impact of classification tasks.

TABLE VIII. FEATURE IMPORTANCE SCORES (CHINA)

Ranking	ID	Importance Score
1	X66	0.0630
2	X77	0.0477
3	X79	0.0445
4	X6	0.0432
5	X69	0.0426
6	X54	0.0415
7	X24	0.0398
8	X55	0.0378
9	X82	0.0375
10	X56	0.0319

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

TABLE IX. ADJUSTED FEATURE IMPORTANCE SCORES RANKING(CHINA)

Ranking	ID	Importance Score
1	X66	0.0630
2	X77	0.0477
3	X79	0.0445
4	X6	0.0432
5	X69	0.0426
6	X54	0.0415
7	X82	0.0375
8	X17	0.0299
9	X61	0.0292
10	X8	0.0244

IV. RESULTS

Based on the above experiment, we could start to compare the features importance between the two datasets in Table X.

TABLE X. FEATURE IMPORTANCE SCORES COMPARISON

Poland		Ranking	China	
Description	*ID*		*ID*	*Description*
profit on operating activities / total assets	X22	1	X66	Cash ratio
profit on sales / total assets	X35	2	X77	Working capital / total assets
total liabilities / ((profit on operating activities + depreciation) * (12/365))	X41	3	X79	Retained earnings / total assets
operating expenses / total liabilities	X34	4	X6	ROA (net)
(net profit + depreciation) / total liabilities	X26	5	X69	Total equity attributable to shareholders of the parent company / liabilities
[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365	X5	6	X54	Asset-Liability Ratio
(current assets - inventory) / short-term liabilities	X46	7	X66	Total shareholders' equity (including minority) / total liabilities
sales (n) / sales (n-1)	X21	8	X17	Operating profit / total operating income
profit on sales / sales	X39	9	X61	Current debt equity ratio
retained earnings / total assets	X6	10	X23	ROE (annualized)

When We had a look at these attributes, it was feasible to find some related features among these important features in each dataset.

If we only discussed direct relationships, X6 (Poland) and X79 (China) were both one of 10 features in their ranking list. And they were the absolutely same feature, retained earnings / total assets. But it looked like nothing interesting else except this relation. Was it the truth?

Apparently, it was not. When we paid attention to those ordinary and well-known features, we introduced before, like liability and assets, we found they really appeared many times in the ranking lists. Our models preferred to make their prediction based on these features related to the two basic attribute, liability and assets. And the operating activities, which could generate another significant index we introduced, cash, also appeared frequently.

In addition, actually assets or equity appeared in 8 attributes in ranking list of China, but only in 4 attributes in list of Poland. Income-related elements, like profit or income, appeared in only 1 attribute in list of China, but in 5 attributes in list of Poland.

Hence, in China's market, assets played an important role in bankruptcy prediction, because many state-owned enterprises (SOE) with large assets, which was main part of China's market, did not easily go bankrupt. And 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 [8], it meant whatever they had better profits, they had more possibility to get loan to maintain operation to avoid bankruptcy. On the contrary, in Poland, whether it had excellent operating situation may become an important factor affecting bankruptcy.

V. RELATED WORK

In earlier research, a logistic regression model was used to predict financial distress, but 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. And the final result was as following table.

TABLE XI. 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	

Actually, 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 performance of ROA (net) was the best one [2].

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

Next, let's 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 XII. 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 own prediction, was still ranked very high. And it appeared in the logistic regression 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-workers calculated the importance score based on multiple model, but we only relied on the model having best performance.

VI. 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). And the assets elements were significant in bankruptcy in China's market, but operation situations were more important in Poland's market.

But 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 conduct the experiment with a larger dataset in the future. At the same time, our study only stays on the surface phenomenon without deeper 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 would like to thank my professor Gary Weiss for his contribution in building my data mining knowledge; my girlfriend Yi Wang for her help in obtaining China's data from Wind database and identification of those 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] J. R. Quinlan, C4. 5: programs for machine learning, 1st ed. Morgan Kaufmann, 2014.

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

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