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
New York, New York, USA
gaweiss@fordham.edu

***Abstract*—This study generates data mining models to predict corporate bankruptcy in Poland and China, and then examines these models to determine the financial characteristics that are of the greatest predictive value. These financial features are then compared for the two countries. The study finds that there while there are some common financial indicators for bankruptcy between the two diverse financial markets, there are also key differences. In particular, asset-related features play a much larger role in predicting bankruptcy in China, while operations-related features play a larger role in predicting bankruptcy in Poland.**

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

I. Introduction

Corporate bankruptcy is an important topic in both the accounting and finance disciplines. The advent of data mining has provided a different set of methods for predicting these corporate bankruptcies. However, much of the relevant research in this area has focused more on the predictive value of the models than on which features they utilize to achieve this performance [1]. Although there have been attempts to explore the features associated with bankruptcy, these attempts generally rely on a statistical and econometrical approach rather than a data mining approach [2]. Furthermore, most work on studying bankruptcies is limited to a single country and there is very rarely any comparison between the importance of different financial features between dissimilar financial systems. This study utilizes data mining methods to build predictive models of corporate bankruptcy in Poland and China, and then compares the importance of the financial features utilized in these models.

This study utilizes corporate bankruptcy data from Poland and China that indicates the bankruptcy status after one year. The predictive models are built using classification algorithms that report feature importance. We only choose those algorithms which can report feature importance. Thus powerful algorithms like neural nnetworks are not utilized since they cannot easily determine feature importance [3]. Researchers have utilized feature selection and ranking methods in the context of data mining, but rarely in the financial field [4].

II. Background

This section provides necessary background information about bankruptcies in Poland and China. We start with a definition of bankruptcy. Bankruptcy is a legal process through which people or entities who cannot repay debts to creditors seek relief from some or all of their debts. It is a legal definition where the details vary between nations. However, within the financial field, the definition is simpler: bankruptcy occurs when insolvency appears, which means the entity is no longer able to repay its liabilities. This definition can be used to conduct studies that span nations despite the variations in the legal processes and definitions associated with bankruptcy [5].

The key financial indicators that impact bankruptcy include assets, liabilities, and available cash. Nonetheless, even with these financial indicators, a bankruptcy may not be seen ahead of time, which is quite problematic. This study, therefore, can make a practical contribution by identifying a better set of financial indicators associated with bankruptcy. Any financial warning would be invaluable to investors and shareholders.

III. Predicting Corporate Bankruptcies in Poland

This section describes the data, experiments, results, and features associated with corporate bankruptcies in Poland.

A. Data Description

The data set used in this study is “Polish Companies Bankruptcy Data Set,” which is available from the UCI Machine Learning Repository [1][6]. The data was collected from the Emerging Markets Information Service, which is a database containing information on emerging markets around the world. The bankrupt companies are from the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013. In order to predict whether a company will be bankrupt in the next year, only the fifth dataset is used for our task, which contains company financial information in the *t-1* year and the associated bankrupt status in the *t* year. This dataset is entitled “5thYear” in original format on the UCI Machine Learning Repository, which contains financial rates from the fifth year of the forecasting period, and corresponding class label that indicates bankruptcy status after one year.

The data set contains 5,910 total instances (i.e., financial statements), of which 410 (6.9%) represent bankrupted companies and 5500 (93.1%) represent companies that were not bankrupted during the forecasting period. The data set contains the 64 attributes that are enumerated in Table I. The missing values in the data set were imputed using the mean value. The one exception is feature P37, which had missing values for almost half of all records; this attribute was converted into a binary attribute that indicated if the features was missing or present. The target class has two values: “bankrupted” and “not bankrupted” (i.e., solvent). Feature engineering and dimensionality reduction techniques like principal components analysis were not used since they would obscure the importance of the original features.

TABLE I DESCRIPTION OF FEATURES IN POLISH DATA SET

ID	Description	ID	Description
P1	net profit / total assets	P33	operating expenses / short-term liabilities
P2	total liabilities / total assets	P34	operating expenses / total liabilities
P3	working capital / total assets	P35	profit on sales / total assets
P4	current assets / short-term liabilities	P36	total sales / total assets
P5	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365	P37	(current assets - inventories) / long-term liabilities
P6	retained earnings / total assets	P38	constant capital / total assets
P7	EBIT / total assets	P39	profit on sales / sales
P8	book value of equity / total liabilities	P40	(current assets - inventory - receivables) / short-term liabilities
P9	sales / total assets	P41	total liabilities / ((profit on operating activities + depreciation) * (12/365))
P10	equity / total assets	P42	profit on operating activities / sales
P11	(gross profit + extraordinary items + financial expenses) / total assets	P43	rotation receivables + inventory turnover in days
P12	gross profit / short-term liabilities	P44	(receivables * 365) / sales
P13	(gross profit + depreciation) / sales	P45	net profit / inventory
P14	(gross profit + interest) / total assets	P46	(current assets - inventory) / short-term liabilities
P15	(total liabilities * 365) / (gross profit + depreciation)	P47	(inventory * 365) / cost of products sold (COGS)
P16	(gross profit + depreciation) / total liabilities	P48	EBITDA (profit on operating activities - depreciation) / total assets
P17	total assets / total liabilities	P49	EBITDA / sales
P18	gross profit / total assets	P50	current assets / total liabilities
P19	gross profit / sales	P51	short-term liabilities / total assets
P20	(inventory * 365) / sales	P52	(short-term liabilities * 365) / COGS
P21	sales (n) / sales (n-1)	P53	equity / fixed assets
P22	profit on operating activities / total assets	P54	constant capital / fixed assets
P23	net profit/sales	P55	working capital
P24	gross profit (in 3 years) / total assets	P56	(sales - COGS) / sales
P25	(equity - share capital) / total assets	P57	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
P26	(net profit + depreciation) / total liabilities	P58	total costs /total sales
P27	profit on operating activities / financial expenses	P59	long-term liabilities / equity
P28	working capital / fixed assets	P60	sales / inventory
P29	logarithm of total assets	P61	sales / receivables
P30	(total liabilities - cash) / sales	P62	(short-term liabilities *365) / sales
P31	(gross profit + interest) / sales	P63	sales / short-term liabilities
P32	(current liabilities * 365) / COGS	P64	sales / fixed assets

B. Modeling

The classification models were induced from the data set using the Python-based scikit-learn data mining toolkit [7]. Since the primary goal of this study is to identify the features most responsible for predicting bankruptcy, only classification algorithms that measure/rank feature importance were utilized. This study used the following four classification methods:

- Logistic Regression (LR)
- The C4.5 decision tree model (DT) [8]
- Extreme Gradient Boosting (XGBoost) [9]
- Random Forest (RF) [10]

The XGBoost algorithm is not included in scikit-learn, so the xgboost library was used for this [9]. The experiments used a training/test set split of 70%/30%, where the instances were randomly selected. Unless otherwise specified default parameters were used for all experiments. Because mistakenly predicting a company that will become bankrupt as a solvent is much worse than predicting a solvent company will become bankrupt, instances belonging to the "bankrupt" class are weighted five times more than instances belonging to the "solvent" class. In the DT, XGBoost and RF models, when the algorithm calculated the gain ratio, it would regard one instance of the minor class as five instances in a node, so that adjusted the weighting. This weighting was not a default parameter in those models provided, so it was input manually. Meanwhile, we used to 10-fold cross-validation method to avoid possible over-fitting.

C. Prediction Results

The classification results for the four algorithms are provided in Table II. The evaluation metrics include precision, recall, F1-score, and accuracy. Note that precision, recall, and F1-score are with respect to the "bankrupt" class value.

TABLE II EVALUATION TABLES (POLAND)

Measure	LR	DT	RF	XGBoost
Precision	0.56	0.60	0.80	0.93
Recall	0.12	0.58	0.32	0.56
F1-Score	0.20	0.59	0.46	0.70
Accuracy	0.94	0.95	0.95	0.97

The results clearly demonstrate that XGBoost performs best overall since it has the highest F1-score and accuracy. Table III shows more detailed results by providing the confusion matrix results for XGBoost when evaluated on the test set. The biggest issue with the results is that only 56% of the bankrupted companies are identified, but given the relatively severe level of class imbalance (1:13.5), the results are nonetheless impressive.

TABLE III CONFUSION MATRIX OF XGBOOST MODEL (POLAND)

		Predicted Labels	
		Bankrupt	Solvent
Actual Labels	Bankrupt	64	50
	Solvent	5	1654

D. Feature Importance

This section describes the importance of the features with respect to classifying the Polish corporations. Logistic regression naturally generates feature importance based on the coefficient associated with each feature. The other three algorithms are all based on decision trees, since random forest is an ensemble of decision trees and XGBoost is based on boosted decision trees. Feature importance in decision trees can be calculated, although it is not as straightforward as for logistic regression. In decision trees, the closer the feature is to the root node, and the more often it appears in a boosted decision tree, the greater the weight and importance. Table IV provides the top-10 features for XGBoost, the best-performing model.

TABLE IV XGBOOST FEATURE IMPORTANCE SCORES (POLAND)

Rank	ID	Importance Score	Description
1	P22	0.1052	profit on operating activities / total assets
2	P35	0.0752	profit on sales / total assets
3	P41	0.0482	total liabilities / ((profit on operating activities + depreciation) * (12/365))
4	P34	0.0412	operating expenses / total liabilities
5	P26	0.0393	(net profit + depreciation) / total liabilities
6	P5	0.0389	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365
7	P46	0.0321	(current assets - inventory) / short-term liabilities
8	P21	0.0302	sales (n) / sales (n-1)
9	P39	0.0294	profit on sales / sales
10	P6	0.0286	retained earnings / total assets

IV. PREDICTING CHINESE CORPORATE BANKRUPTCIES

This section describes the data, experiments, results, and feature importance associated with corporate bankruptcies in China.

A. Data Description

The data set of Chinese corporations was formed by merging three datasets, each manually collected from the Wind Financial Database. The data for each record was obtained by querying the financial information one company at a time from WIND financial terminal interface. A total of 61 bankrupted companies were obtained; each was listed on the Chinese stock market and had gone bankrupt since 2006. The 620 solvent companies were collected from SH380 and SZ300 index companies. Hence, the data set contained 681 entries, of which 9.0% represents bankrupted companies and 91.0% represent solvent companies. The feature value information is from the last year of their bankruptcy, or the newest data if solvent.

Table V describes the 84 features that were collected. Missing values were imputed using the mean value. The C43 and C84 features were dropped because there often was no net debt, which led to a zero denominator. The C1 and C2 features were removed since they are identifiers and do not provide useful information.

TABLE V DESCRIPTION OF FEATURES IN CHINESE DATA SET

ID	Description	ID	Description
C1	Company Code	C43	Tangible assets / net debt
C2	Company Name	C44	Capital expenditure / depreciation and amortization
C3	Return on equity (ROE) (average)	C45	Cash received for sales of goods and services / Operating income (TTM)
C4	ROE (deducted / average)	C46	Net cash flow from operating activities / operating income (TTM)
C5	Return on Total Assets (ROA)	C47	Net cash flow from operating activities / operating profit (TTM)
C6	ROA (net)	C48	Proportion of net cash flow from operating activities
C7	Return on Human Investment (RHI)	C49	Proportion of net cash flow from investment
C8	ROE (annualized)	C50	Proportion of net cash flow from fundraising
C9	ROA (annualized)	C51	Net operating cash flow / total operating income
C10	ROA (net) (annualized)	C52	Cash operation index
C11	Sales margin	C53	Cash recovery rate of all assets
C12	Gross profit margin	C54	Asset-Liability Ratio
C13	Cost of sales ratio	C55	Asset-Liability Ratio (excluding advance receipts)
C14	Sales Period Expense Rate	C56	Asset-Liability Ratio (excluding advance receipts) (Announcement Based)
C15	Main business ratio	C57	Long-term debt ratio
C16	Net profit / Total operating income	C58	Long-term asset fit ratio
C17	Operating profit / total operating income	C59	Tangible assets / total assets
C18	EBIT / Total operating income	C60	Non-current debt ratio
C19	Total operating cost / Total operating income	C61	current assets / short-term liabilities
C20	Management expenses / Total operating income	C62	Current liabilities / total liability
C21	Financial costs / total operating income	C63	Capitalization ratio
C22	Asset impairment loss / total operating income	C64	Quick ratio
C23	ROA (Trailing Twelve Months (TTM))	C65	Conservative Quick Ratio
C24	ROA (net) (TTM)- Excluding Minority Shareholder Profit and Loss	C66	Cash ratio
C25	Return on invested capital (ROIC)	C67	Net Asset-Liability Ratio
C26	ROIC (TTM)	C68	Net debt ratio
C27	EBIT/ Total assets (TTM)	C69	Total equity attributable to shareholders of the parent company / liabilities
C28	Net sales margin (TTM)	C70	EBITDA / Total Liability
C29	Gross profit margin (TTM)	C71	Net cash flow from operating activities / Total Liability
C30	Sales Period Expense Rate (TTM)	C72	Net cash flow from operating activities / Current liabilities
C31	Operating profit / Total operating income (TTM)	C73	Net cash flow from operating activities / Non-current liabilities
C32	Total Operating Cost / Total Operating Revenue (TTM)	C74	Non-financing net cash flow / Non-current liabilities
C33	Operating profit / operating income (TTM)	C75	Non-financing net cash flow / Total Liability
C34	Tax / Total Profit (TTM)	C76	Proportion of long-term debt
C35	Net profit attributable to shareholders of the parent company / Operating income (TTM)	C77	Working capital / total assets
C36	Asset impairment loss / total operating income (TTM)	C78	Tangible net worth debt ratio
C37	Asset impairment loss / operating profit	C79	Retained earnings / total assets
C39	Operating profit / total profit (TTM)	C80	EBIT (TTM) / total assets
C39	Total profit / operating income (TTM)	C81	Total market value / liabilities for the day
C40	Cash received from sales of goods and services provided/operating income	C82	Total shareholders' equity (including minority) / total liabilities
C41	Net cash flow from operating activities / operating income	C83	Operating income / total assets
C42	Net profit cash rate	C84	Net cash flow from operating activities / Net debt

B. *Modeling*

The experiments for the Chinese data set are similar to those for the Polish data set, and the same four classification algorithms were used. The training and test sets were again partitioned using random sampling, but this time using a train/test split of 50%/50%, since the Chinese data set is smaller and has very few bankrupt instances, which makes accurate evaluation more difficult. As before, the bankrupt companies were given a weight of five times that of the solvent companies.

C. *Prediction Results*

The classification results are provided in Table VI, using the same format that was used previously for the Polish results. Random Forest and XGBoost both perform much better than logistic regression and decision trees. Random Forest and XGBoost perform equally well for both accuracy and F1-Score, with the only real difference being that Random Forest has a higher precision while XGBoost has a higher recall. In this case we give slight preference to precision over recall and choose Random Forest for the remainder of our analysis. Table VII shows the confusion matrix for Random Forest.

TABLE VI EVALUATION TABLES (CHINA)

Measure	**LR**	**DT**	**RF**	**XGBoost**
Precision	0.65	0.74	1.0	0.96
Recall	0.79	1.0	0.90	0.93
F1-Score	0.71	0.85	0.95	0.95
Accuracy	0.95	0.97	0.99	0.99

TABLE VII CONFUSION MATRIX OF RF MODEL (CHINA)

		Predicted Labels	
		Bankrupt	Solvent
Actual Labels	Bankrupt	26	3
	Solvent	0	312

D. *Feature Importance*

Random forest will generate different feature importance values for each run. The model is trained and tested many times and a specified number of the top features is extracted from each run. Ultimately the top features aggregated over all of the runs are collected. However, several features have a strong correlation with each other. For instance, C55, Asset-Liability Ratio (excluding advance receipts) is very similar to C54, Asset-Liability Ratio. In cases like this, where features are just slight variations of each other, we kept only the higher-ranked one and adjusted the ranks. Table VIII shows the top ranked features for the Chinese companies. The features that were removed for being variations of other features are denoted in Table by "-".

TABLE VIII FEATURE IMPORTANCE SCORES (CHINA)

Rank	**ID**	**Importance Score**	**Description**	**Adjusted Rank**
1	C66	0.0630	Cash ratio	1
2	C77	0.0477	Working capital / total assets	2
3	C79	0.0445	Retained earnings / total assets	3
4	C6	0.0432	ROA (net)	4
5	C69	0.0426	Total equity attributable to shareholders of the parent company/liabilities	5
6	C54	0.0415	Asset-Liability Ratio	6
7	C24	0.0398	ROA (net) (TTM)-Excluding Minority Shareholder Profit and Loss	-
8	C55	0.0378	Asset-Liability Ratio (excluding advance receipts)	-
9	C82	0.0375	Total shareholders' equity (including minority) / total liabilities	7
10	C56	0.0319	Asset-Liability Ratio (excluding advance receipts) (Announcement Based)	-
11	C17	0.0299	Operating profit / total operating income	8
12	C61	0.0292	current assets / short-term liabilities	9
13	C8	0.0244	ROE (annualized)	10

V. COMPARISON OF IMPORTANT FEATURES

This section compares the features that are important for identifying companies that are going to become bankrupt in Poland and China. Table IX shows similar features, which are ranked in the top 15 most important features, for both cases. If we only discussed perfect feature matches, then that would leave us with only P6 and C79. Each of these were one of the top 10 features in their respective lists. However, other pairs were quite similar. The main difference between P46 and C61 was that the inventory was subtracted in the numerator in P46. As far as the pair of P22 and C7, EBIT was earnings before interest and taxes, containing profit on not only the operating activities but also other profitable activities (this does not commonly make much difference). In terms of the pair of P28 and C77, there was little difference in the denominator, with one using fixed assets and the other total assets. Thus, we see that essentially there are four important features in common.

TABLE IX PAIRS OF RELEVANT FEATURES

Poland		**China**	
Description	*ID*	*ID*	*Description*
retained earnings / total assets	P6	C79	retained earnings / total assets
(current assets - inventory) / short-term liabilities	P46	C61	current assets / short-term liabilities
profit on operating activities / total assets	P22	C7	EBIT / total assets
working capital / fixed assets	P28	C77	working capital / total assets

There are also some important differences between the important features for the best classification models for the two countries. Asset and equity related features appeared eight times in the ranking list for China but only four times in the list for Poland. Meanwhile operation-related features, like profit and income, appeared five times in the ranking list for Poland but only one time in the list for China. In China, assets played an important role in bankruptcy prediction, because there are many state-owned enterprises (SOE) with enormous assets (which comprise a large part of the Chinese market. Moreover, operating-related attributes were not a big factor in predicting bankruptcy. SOEs generally have more assets and lower costs to get financing[11], which have crucial influence in financing like loans or bonds, so whatever they obtained high profits or not, they always had more possibility to get a loan to maintain operation to avoid bankruptcy. On the contrary, in Poland, the capacity to have an excellent operating situation may become an important factor affecting bankruptcy.

VI. Related Work

Even though there are many studies on bankruptcy in both the finance and machine learning areas, those studies tend not to focus on identifying the key financial characteristics. More importantly, there do not appear to be studies that compare the importance of features for predicting bankruptcy using data mining methods for multiple countries.

In earlier research, a logistic regression model was used to predict financial distress in China [2]. Although it only utilized six features of 139 records, the excellent performance was achieved with an accuracy of 0.94 for overall and an F1-score of 0.93 for class “distress.” However, the researchers did not explore features. They chose to conduct a univariate analysis and found the performance of ROA (net) was best. Several of the features in their study, including ROA and working capital to total assets ratio, were also important in our models as C6 and C77. Moreover, their asset turnover is very similar to P22 and P35. And the current ratio was similar to the P46.

Our study can also be compared to the prior research by Zieba et al. [1] that was conducted on the Polish data set. Table X shows the ranked list of important features found in that study.

TABLE X FEATURE IMPORTANCE SCORES (POLAND) (ZIEBA) [1]

Rank	ID	Importance Score
1	P25	0.0627
2	P22	0.0480
3	P27	0.0379
4	P15	0.0356
5	P52	0.0326
6	P53	0.0284
7	P14	0.0248
8	P40	0.0247
9	P42	0.0238
10	P36	0.0236

The P22 feature, profit on operating activities / total assets, which had the highest importance score in our prediction (see Table IV), was still ranked very high. Furthermore, it appeared in the logistic regression model we introduced as well. However, except for P22, there were no features in Zieba’s list that also appeared in our list. The difference was probably caused by that Zieba calculating the importance score on multiple models, while we relied only on the best performing model.

VII. Conclusion

The study described in this paper involved generating and evaluating classification models for predicting bankruptcy in companies in Poland and China, identifying the most important features, and then comparing and contrasting the features for the two nations. This comparison demonstrated that while there are some commonalities in the models for Polish and Chinese corporate bankruptcies, there are some significant differences. The common indicators include ROA and Retained Earnings to Total Assets Ratio (RE/TA), but in the Chinese markets the asset-related features were much more important than for the Polish markets, while the operations-related features were much more important in the Polish markets than in the Chinese markets. These differences were explained based on the role of large state-owned enterprises in China. This study is quite unusual in that it analyzed and compared the role of financial features in two very different markets.

There are several ways in which this research can be extended and improved. First, we only considered the financial index of the year before the bankruptcy. There was no longer-term or periodic consideration. Meanwhile, the study was based on the pre-existing list of companies, which may have introduced a bias. In particular, the Chinese data set is quite small, which made it more difficult to reliably evaluate, and which limits the generality of the results. Analysis using a much larger data set would be quite beneficial.

We hope that this study and its novel perspective will inspire future research. Data mining has not been widely adopted in the asset pricing field or in corporate finance, and we hope that this study will help to change that.

Acknowledgment

The authors would like to thank Yi Wang for her assistance in obtaining the data for China.

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.