Monitoring corporate credit risk with multiple data sources

Du Ni[2], Ming K. Lim[4], Xingzhi Li[3], Mei Yang[1]

[1] School of Economics and Business Administration, Chongqing University, Chongqing, 400030, PR China
[2] School of Management, Nanjing University of Posts and Telecommunications, Jiangsu, 210003, PR China
[3] School of economics and management, Chongqing Jiaotong University, Chongqing, 400074, PR China
[4] Adam Smith Business School, University of Glasgow, Glasgow G14 8QQ, United Kingdom

Abstract

Purpose: Monitoring corporate credit risk (CCR) has traditionally relied on such indicators as income, debt, and inventory at a company level. These data are usually released on a quarterly or annual basis by the target company and include, exclusively, the financial data of the target company. As a result of this exclusiveness, the models for monitoring credit risk usually fail to account for some significant information from different sources or channels, like the data of its supply chain partner companies and other closely relevant data yet available from public networks, and it is these seldom used data that can help unveil the immediate CCR changes and how the risk is being propagated along the supply chain.

Design: Going beyond the existing CCR prediction data, this study intends to address the impact of supply chain data and network activity data on CCR prediction, by integrating machine learning technology into the prediction to verify whether adding new data can improve the predictability.

Findings: The results show that the predictive errors of the datasets after adding supply chain data and network activity data to them are made the ever least. Moreover, intelligent algorithms like support vector machine (SVM), compared to traditionally used methods, are better at processing nonlinear datasets and mining complex relationships between multi-variable indicators for CCR evaluation.

Value: This study indicates that bringing in more information of multiple data sources combined with intelligent algorithms can help companies prevent risk spillovers in the supply chain from causing harm to the company, and, as well, help customers evaluate the creditworthiness of the entity to lessen the risk of their investment.

Keywords

corporate credit risk, prediction, multiple data sources, machine learning, support vector machine

1. Introduction

One major characteristic of corporate credit risk (CCR) is that investors must bear certain risks if chipping in financial securities or purchasing bonds of some business companies. If a bond issuer fails to repay the principal and interest when due, the investors will suffer the losses (Bazarbash, 2019; Zhang et al., 2022). Then CCR assessment is applied, in a way, to help ensure the worthiness of the purchase or cushion the investors from the losses. However, several studies suggest that the CCR prediction is greatly affected by the evaluation standards (Fracassi et al., 2016; Wang et al., 2020), and It is not easy to include all factors leading to corporate default by a single evaluation method (Li et al., 2020). Therefore, when predicting CCR, it is necessarily better to consider the effect of its evaluation and try a broader coverage of predicative data.

In general, the traditional indicators for CCR prediction mainly come from the target company's financial data (Bonsall IV et al., 2017; Wu et al., 2015), which are usually released by the target company on a quarterly or annual basis. However, these traditional indicators have certain intrinsic deficiencies; for instance, some companies' financial data are not released regularly (Grover et al., 2018; Wu et al., 2022). Even worse, their financial reports are easily manipulated that there are often seen overstating assets, improper disclosures and immoral financial frauds, yet it is impossible to testify these manipulations as easily (Lev, 2018; Zhuang et al., 2021). Luckily, the deficiencies find their remedies in the development of artificial intelligence and communication technology which has rendered a more transparent financial and information flow between companies and institutions (AlShamsi et al., 2021; Huang et al., 2021). The CCR can be revealed sooner by a careful study of some data available on public networks before the target company suffers the financial losses, and even before the target company realizes a credit risk may occur (Li et al., 2021; Moat et al., 2016). Besides, some researchers have found that credit risks are able to spread along the supply chain networks and adversely affect the target companies through the supply chain (Agca et al., 2021; Lei et al., 2021), and therefore, in considering the company's own financial situation, the data expansion by including data of network activities along the supply chains are proposed to be employed to predict the CCR in a more timely and more precise way.

This study will first review the previous studies on CCR prediction in Section 2. Based on the gaps explored in the literature review, Section 3 is to present the methods concerning data acquisition, compilation and processing for the CCR model analysis deployed by this study. Section 4 is for the results of the data analysis and an elaborate discussion about them. Finally, the conclusions of this study are presented in Section 5.

2. Literature review

Traditional indicators for CCR prediction are mainly collected from data of a target company like cash flow (Wu & Brynjolfsson, 2015), financial proportion heterogeneity (Niemann et al., 2008), debt

cost (Mansi et al., 2012), corporate governance (Bonsall IV et al., 2017), to name some of them. Generally, these data are released by a single company either quarterly or annually. This exclusiveness, to some extent, has tarnished the credibility of the data, and an increase in financial frauds over the years has furthered undermined the credibility of publicized financial data The trick is that the researchers for CCR prediction can't verify whether the company has exaggerated its financial data or not (Lev, 2018; Zhuang et al., 2021), and some small and medium-sized enterprises or private companies do not release their financial data regularly at all, as a result of which CCR institutions obtain no valuable information for CCR prediction (Grover et al., 2018; Wu et al., 2022), and also as a result of which the institutions have turned to the other data sources and research methods for a trustworthy and effective CCR prediction.

In regard of the valuable data sources for CCR prediction, several studies claim that rapid advances in technologies such as artificial intelligence have made financial and information flows between companies and the institutions for evaluation more transparent. Non-financial data of network activities like the negative news (Martín et al., 2021), social media evaluation (Bazarbash, 2019; Ghasemkhani et al., 2015) and Internet search (Liu, 2020) can also have an impact on CCR prediction, as such data on the Internet can reveal the corresponding preferences of corporate managers. In other words, there should be a correlation between online activities and economic performance. By quantifying the search on Internet with the help of Google and Wikis for financial-related information, the researchers hold that they can capture the real-time CCR changes ahead of the risk report in its traditional financial statements (Moat et al., 2016; Preis et al., 2013). Moreover, recent studies suggest that supply chain data could be helpful for analyzing economic behaviours, for the credit risk has been verified to be able to spread along supply chain, and adversely affect the target companies through the supply chain (Agca et al., 2021; Lei et al., 2021). Therefore, the researchers hope to employ network activity data in monitoring sudden inventory, sales, and other fluctuations along the supply chain, or so to speak, they intend to optimize the structure of the supply chain with the network data, such as customer reviews, sales data, sudden traffic changes, and inventory information (Wang et al., 2016). That Lee et al. (2021a) succeeded in timely predicting the CCR by monitoring the company's supply chain data and the Internet data on Twitter is one case in point.

More cases of recent studies and an overview of some detectable transition in CCR prediction are presented in Table 1. As seen from the table, financial data and loan records, one variety of financial data, have dominated the studies on CCR prediction, and in terms of the number of indicators for CCR prediction, the researchers have used an average of 26 of them in the studies, which seems not bad for a fair CCR assessment but somewhat dubious considering the high uniformity of their sources. Since 2015, non-financial data out of network and supply chain has entered as valuable data for CCR prediction (Ghasemkhani et al., 2015), and there has been evidence that the expansion by adding non-financial indicators can improve the accuracy of CCR prediction (Moat et al., 2016; Teles et al., 2020),

but the cases are few and there are even fewer cases where are seen indicators featuring all sources of finance, network and the supply chain.

Table 1. Recent studies on the prediction of cooperate credit risk

Authors & Year	Indicators Source	Number of Indicators
Trustorff et al. (2011)	Financial data	19
Wang et al. (2011)	Financial data	18
Oreski et al. (2012)	Financial data	33
Hajek et al. (2013)	Financial data	14
Zhang et al. (2014)	Loan records	24
Wu et al. (2014)	Financial data	18
Ghasemkhani et al. (2015)	Financial data; Supply chain & Network activity data	37
Florez-Lopez et al. (2015)	Financial data	20
Yang et al. (2016)	Financial data; Social media text	15
Zhu et al. (2016)	Financial data	18
Gu et al. (2017)	Financial data; Supply chain data	30
Figini et al. (2017)	Financial data	43
Bequé et al. (2017)	Loan records; Demographics	51
Zhang et al. (2018)	Loan records; Historical Loan Information	44
Jiang et al. (2018)	Loan records	15
Huang et al. (2018)	Financial data	7
Zhang et al. (2019)	Loan records	14
Papouskova et al. (2019)	Loan records	78
Tang et al. (2019)	Loan records	13
Shen et al. (2019)	Loan records	15
Pławiak et al. (2020)	Loan records	20
Wang et al. (2020)	Financial data	30
Martín-Oliver et al. (2020)	Loan records	16
Lee et al. (2021b)	Financial data	22
Wang et al. (2021)	Financial data	28
Lee et al. (2021a)	Network activity data	8
Wang et al. (2022)	Financial data; Negative information	40
Yu et al. (2022)	Macroeconomic Variables; Financial data	11
Zhang et al. (2022)	Financing behavior data; Demographics	53

As for the research methods, the CCR assessment used to go with traditional statistical methods such as linear discriminant analysis and multivariate discriminant analysis, and the like (Chen et al., 2016; Mahmoudi et al., 2015; Ul Hassan et al., 2017), meaning to find the best linear correlation of input indicators. However, there are assumptions that linear separability, variable independence, and multivariate normality cannot handle complex relationships between multiple variables (Chen et al., 2016), therefore many recent studies have shifted to intelligent methods which can mine complex

relationships between variables without relying on the restrictive assumptions, and support vector machine (SVM), in particular, has been verified as a very powerful intelligent algorithm in that it allows for complex decision boundaries and performs well on the non-linear datasets with high dimensional variables (Cervantes et al., 2020; Ghaddar et al., 2018).

In light of the literature reviewed above on data sources and research methods, here are the indentified research gaps: (1) Studies in this area have attempted to expand the dada sources for CCR prediction, but the expansion is always lopsided, and none of them has truly established a full-feature database so far. As much done as Ghasemkhani et al. (2015) included Wikipedia and Google hits when predicting CCR, their study ignored data from public opinions. Actually, a clearer capture of the CCR should be based on a well-rounded data expansion, reaching out for multi-level indicators from sources as various as they can be. (2) Traditional statistical methods like linear discriminant analysis and multivariate discriminant analysis have been used in predicting CCR (Mahmoudi & Duman, 2015; Ul Hassan et al., 2017), but these methods require the dataset to be linearly separable and variable independent (García et al., 2019), so they often fail to interpret the nonlinear relationship among the corresponding variables, as the number of indicators increases when the data sources are expanded.

To fill these two gaps, this study intends first to expand the data into more sources and manage them in a well-rounded way by adding both network activity data and supply chain data, and then to verify the effects of the expansion. To be specific, this study is to address two research questions. One is the feasibility of expansion of the data sources; the other is the verification of the expansion of the data sources.

3. Data and Methods

This section is to introduce the database sources, the selection of data, the compilation of datasets both as dependent and independent variables in the CCR model, and the processing of the data with the specific courses and the rationale underlying them.

3.1 Data sources

Actually, into the most conventional corporate financial data, this study adds two more categories: supply chain data, network activity data as independent variables, and credit risk data, the dependent one, and they are retrieved from the sources as follows.

(1) Corporate financial data. The financial data for this study were mainly from the Standard & Poor's Accounting Database (https://www.spglobal.com/ratings/en/). The database covers more than 50,000 public and delisted companies worldwide, including annual financial data, quarterly financial data, and other phase financial data of incorporated companies. These data can reflect the corporate overall market policy and strategy changes and the financial stability or instability on the corporate performance. This study selected 441 companies and obtained their financial data, weekly, quarterly and annual, dating from January 1, 2009, to December 31, 2019, and specifically including weekly

working capital ratio, interest compensation ratio, retained earnings ratio, return on assets, tax leverage ratio, cash inverse ratio, debt to capital ratio, and some other indicators.

(2) Network activity data. The data for search trends, website visits, and other network activity data have a strong timeliness and can be used as a supplementary source to corporate financial data with a weak timeliness (Moat et al., 2016). This study collected data from the page texts of Wikipedia, Google Trends, and Facebook, with the search period of January 1, 2009, through December 31, 2019. The main reasons for selecting these websites were 1) Google is the most widely used search engine in the world, and googling is the most commonly used method for investors to search for the information of a company when choosing corporate bonds. Google Trends data present the popularity of popular queries in Google searches in different regions and languages (Preis et al., 2013; Sulyok et al., 2021). 2) Wikipedia is the world's largest popular science website (Moat et al., 2016), which allows any user to freely edit its content and make comments. Despite some concerns about the reliability and accuracy of Wikipedia's content, this decentralized model of knowledge building draws on the wisdom of the masses, making Wikipedia similar to a social media platform--the first place for a quick understanding of superficial information (Samoilenko et al., 2014), just as Moat et al. (2013) claimed in their study that the changes of search frequency of Wikipedia were able to reflect the fluctuation of stock markets. 3) Facebook has some 2.91 billion active users, making it a vital platform for small businesses in particular, where potential customers or real investors often refer to the news about the business, products, services, up-coming events and user comments posted on a business company's Facebook as emotional factors (Ladhari et al., 2019).

(3) Supply chain data. The supply chain data for this study were collected from the Bloomberg Database (https://www.bloombergchina.com/solution/data-content/). The database contains more than 20,000 pieces of quantitative supply chain data, filed by the target companies to their regulators of the U.S. Securities and Exchange Commission, mainly in the form of company reports and electronic revenue records (Scott III, 2010), and the Database will automatically work out an average of each supplier's revenue and the cost of its trades with the target company. The point is these Bloomberg data can help trace the upstream of the supply chain to find about suppliers and all customers downstream too (Schwieterman et al., 2020). This study obtained 19 indicators of supply chain data on the 441 companies from the Bloomberg Database.

(4) Credit risk data. There is always a risk that a company cannot fulfill its responsibilities completely due to various reasons, and a major sign of its credit risk is the fluctuation of bond price (Gilchrist et al., 2018; Zamore et al., 2018). Therefore, this study refers to the bond price in different variations as the actual value of corporate credit risk, as against the expected CCR value out of the evaluation algorithms, the data of which were mainly from the Standard & Poor's Accounting Database (https://www.spglobal.com/ratings/en/) dating from Jan. 1, 2009, to Dec. 31, 2019. A total of 6 indicators were used in this study, including Credit rank, Weekly/daily sum volume reported on the trade, weekly/daily standard deviation of reported bond price, weekly/daily mean standard deviation of

return earned on a security, weekly/daily mean reported bond price, and weekly/daily mean return earned on a security.

A neat summary of the database sources was presented in Table 2: 19 corporate financial indicators, 30 network activity indicators, 38 supply chain indicators, and 6 credit risk indicators, 93 indicators in total, out of which 87 serve as predicative variables or the independent variables, the biggest ever seen in this research field, and of cross-channel quality.

Table 2. Summary of the data sources

Category		Number of indicators	Data sources
Corporate financial data		19	Standard & Poor's Accounting Database
Supply chain data	Upstream company	19	Bloomberg Database
	Downstream company	19	Bloomberg Database
Network activity data	Upstream company	10	Wikipedia, Google Trends, Facebook
	Downstream company	10	Wikipedia, Google Trends, Facebook
	Target company	10	Wikipedia, Google Trends, Facebook
Credit risk data		6	Standard & Poor's Accounting Database
Total		93	——

3.2 Evaluation methods

Having expanded the data source and gathered new prediction indicators, next is to verify the validity of the practice. In this study, the datasets for the running CCR model are to be, first, compiled and tested through the Delphi method, a conventional technique for data evaluation, on the validity of the newly added indicators, and varied combinations among these data, to be exactly, in order to single out the combination that may have the optimal result in CCR prediction. Delphi method, to arrive at a group decision on the data by surveying a panel of experts, is considered a popular social research technique (Belton et al., 2019), which asks experts respond to questionnaire questions, and after rounds of responses, plus consultation, induction, and revision, a group of the most reliable consensus of expert opinions are finally obtained (Brady, 2015). And alongside Delphi method, intelligent algorithms, known as a powerful tool for handling datasets, both linear and non-linear ones, is then applied to double test the effect of the introduction of the new data from network and supply chain into the evaluation. In other words, this study will combine the qualitative and quantitative methods to verify whether the addition of supply chain data and network activity data will affect CCR prediction. To be more specific, this study will adopt the Delphi method to sort out datasets, and the OLS (Ordinary Least Squares) and SVM algorithms to verify the predictability of datasets with MAPE (Mean Absolute Percentage Error), MSE (Mean Squared Error), and RMSE (Root Mean Square Error) as the evaluation index.

3.2.1 Delphi method

To precisely illustrate the effect of the two newly included types of data, supply chain data and network activity data on CCR prediction, a careful comparison is necessary between the effects of different indicator combinations on CCR, and for that matter, the amalgamation of indicator variables

from varied sources was done first and manually by using the Delphi method. The process was divided into the following three steps. Step One: Three CCR experts were invited for their opinion on classification of the data in the first round, where three major categories, were identified of the 87 variables: financial data, network activity data and supply data. Step Two: Following the identification, 25 combinations of these variables for the CCR comparison test were proposed to the experts. Actually the 87 indicators varied much in their sources and types as they were data out of the target or the focal companies, their suppliers and customers up and down their supply lines, or network activities, and differed in nature from financial data to Google trends and online news relevant to the company, there could have been 63 combinations if randomly done, and considering that the 63 datasets, either unidentical or redundant in categorizing features, were unfriendly for the comparison test, 25 of them were left and arranged in order for more opinion of the experts. Step Three: Resulted from the second round of experts' response, 11 index datasets were derived, and the 11 datasets were forwarded to the experts for the third time, along with reasons for this amalgamation, and indicator variables of each specific dataset, and with more consideration and some adjustment, they arrived, unanimously, at the final version of the selected datasets for running the CCR model.

3.2.2 Intelligent algorithms

The intelligent algorithms that are good at processing complex data and not limited by statistical assumptions have received extensive attention from the researchers across fields (Ni et al., 2020, 2021; Sarker, 2021), and again conventionally in quantitative analysis, OLS algorithm is one of the most common techniques to measure the influence of some factors on dependent variables (Bilginol et al., 2015; Zhang et al., 2005). Following the convention, OLS was used in this study too as a multi-variable linear fitting method to evaluate the influence of different indicator groups on the CCR prediction. However, considering the fact that the expansion of data will cause the nonlinear characteristics of data indicators to increase, and to solve the nonlinear problems that may be caused by the increase in data dimensions, the SVM algorithm was used, for of frequently applied algorithms, SVM is particularly well known for its remarkable advantages in dealing with nonlinear and high-dimensional features (Cervantes et al., 2020; Ghaddar & Naoum-Sawaya, 2018). SVM has a strong generalizability for unknown data samples, and its recognition method can be transformed into the processing of convex quadratic programming problems by solving the local optimal point, getting the best support vector, and the use of the kernel function in SVM is able to cleverly avoid the problems of nonlinear data (Lu et al., 2004; Yang et al., 2020), converting the linearly inseparable data into linearly separable ones.

As shown in Figure 1, the black and white dots represent two different categories of attributes. Lines H_1 and H_2 are parallel to each other. The formation conditions of the two lines are as follows: 1) The two lines must pass through points representing different attributes, and the distance between points of different categories passing through is required to be the minimum; 2) Both lines are parallel to the optimal classification line H. The purpose of the two lines is to ensure maximum spacing. The generalized optimal classification hyperplane is formed when the two-dimensional data is extended to

multiple dimensions. The function of the optimal classification hyperplane is to segment the data of different categories effectively and to the maximum extent (Chang et al., 2011).

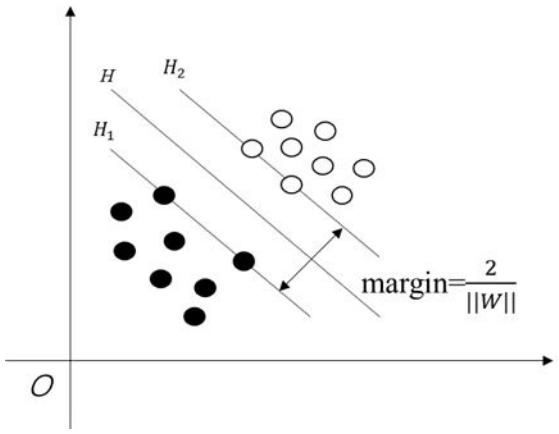

Figure 1. Schematic diagram of a two-dimensional classifier

For H there is:

$$\omega^T x + b = 0 \tag{1}$$

In formula (1), ω represents the normal vector of the hyperplane and b represents the intercept. If the training sample (x_i, y_i) can be properly divided by the optimal classification hyperplane, it is required to satisfy formula (2).

$$\begin{cases} \omega^T x_i + b \geq +1 & y_i = +1 \\ \omega^T x_i + b \leq -1 & y_i = -1 \end{cases} \tag{2}$$

$y_i = +1$ are positive samples and $y_i = -1$ are negative. Here, +1 and -1 are the two types of samples in a theoretical sense just for convenient calculation. Through the equivalent transformation of formula (2), the model formula (3) of the classifier can be obtained.

$$y_i(\omega^T x_i + b) \geq +1 \tag{3}$$

This, in turn, can get interval $= \frac{1-b+1+b}{||w||} = \frac{2}{||w||}$, according to the principle of the maximum interval, and launch formula (4).

$$\begin{cases} max \ \frac{2}{||\omega||} \\ s.t. \ y_i(\omega^T x_i + b) \geq +1 \end{cases} \tag{4}$$

Maximization is equivalent to minimization. For the simplicity of calculation, formula (4) can be converted into Formula (5).

$$\begin{cases} min\ \frac{1}{2}\ ||\omega||^2 \\ s.t.\ y_i(\omega^T x_i + b) \geq +1 \end{cases} \quad (5)$$

The Lagrange function is expressed in formula (6) as follows.

$$L(\omega, b, \alpha) = \frac{1}{2}\ ||\omega||^2 + \sum_{i=1}^{m} \alpha_i [1 - \ y_i(\omega^T x_i + b)] \quad (6)$$

The partial derivative of ω, b is obtained by formula (7).

$$\begin{cases} \frac{\partial L}{\partial \omega} = \omega - \sum_{i=1}^{m} \alpha_i y_i x_i \\ \frac{\partial L}{\partial b} = \sum_{i=1}^{m} \alpha_i y_i \end{cases} \quad (7)$$

If they are equal to 0, formula (8) can be obtained.

$$\begin{cases} \omega = \sum_{i=1}^{m} \alpha_i y_i x_i \\ \sum_{i=1}^{m} \alpha_i y_i = 0 \end{cases} \quad (8)$$

Into the Lagrange function, the formula (9) can be obtained.

$$\begin{cases} L(\omega, b, \alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i\ \alpha_j y_i y_j x_i x_j\ [1 - \ y_i(\omega^T x_i + b)] \\ s.t. \sum_{i=1}^{m} \alpha_i y_i = 0, \alpha_i \geq 0, i = 1,2, \ldots . m \end{cases} \quad (9)$$

After α, ω and b can be solved, and then the model expression (10) can be obtained.

$$f(x) = \omega^T x + b = \sum_{i=1}^{m} \alpha_i y_i x^T x + b \quad (10)$$

Formula (10) is the theoretical optimal classification hyperplane obtained for solving the linear separable problems, which can be used to complete the classification and recognition of data samples. With nonlinear problems, SVM introduces kernel function to solve the problems of the linear inseparability of original spatial data by mapping the vector X from n-dimensional original space to higher-dimensional space (El Kafrawy et al., 2021; Ming, 2015).

3.2.3 Evaluation indicators

In this study, multiple indexes such as MAPE, MSE, and RMSE were used to comprehensively evaluate the model. The calculation formulas are as follows.

$$MAPE = \sum_{t=1}^{n} \left| \frac{observed_t - predicted_t}{observed_t} \right| \times \frac{100}{n} \quad (11)$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (observed_t - predicted_t)^2 \quad (12)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(observed_t - predicted_t)^2} \quad (13)$$

Observed$_t$: observed value, which is the actual bond price; *Predicted*$_t$: the expected value of the model, which is the expected bond price; *n*: Number of samples.

4. Results and discussion

This section will carefully analyze the CCR predictability after adding network data and supply chain data. It also discusses how the introduction of new information will affect CCR prediction.

4.1 Delphi method

This section presents the 12 types in much detail of indicator combinations obtained through Delphi method introduced in Section 3.2.1. Among the 12 types, 11 were decided by experts' consensus and 1 was the dataset with full indicators as shown in Table 3.

Table 3. 12 Datasets based on the Delphi method

Number	Datasets	Number of Indicators
1	Full feature	87
2	Full feature_focal	20
3	Financial_focal	30
4	Trends_wiki_news_focal	11
5	Full feature_customers	30
6	Basic financial_focal	15
7	Financial_trends_wiki_news_focal	25
8	Full feature_suppliers	30
9	Financial_focal_customers	34
10	Full feature_splc	59
11	Financial_focal_suppliers	34
12	Financial_focal_splc	53

As can be seen from Table 3, the original dataset selected by Delphi method has a maximum of 87 indicators, and on average, each, of the 12 types, has 36 predictors. 36, it far exceeds the average of 26 predictors reviewed in Table 1, and these added data form different networks, at different ends of supply chain and of different nature are able to tamper with the biases in CCR assessment caused by the shortage of data or embellished financial data offered by the target companies themselves.

4.2 Intelligent algorithms

Of the all the selected data, the first 80% was taken as the training set and the rest 20% as the test set. OLS and SVM algorithms were used to analyze the predictability of the datasets respectively.

4.2.1 OLS evaluation

In this section, the OLS algorithm was used to compare the predictability of 12 types of datasets. The MAPE value out of R (R-studio) was taken as the evaluation indicator for predictability and Table 4 shows the specific results.

Table 4. The results of OLS evaluation

Datasets	Number of Indicators	MAPE	Ranking
Full feature	87	0.709	1
Full feature_focal	30	0.715	2
Financial_focal	20	0.780	3
Trends_wiki_news_focal	11	0.964	4
Full feature_customers	30	1.086	5
Basic financial_focal	15	1.121	6
Financial_trends_wiki_news_focal	25	1.162	7
Full feature_suppliers	30	1.171	8
Financial_focal_customers	34	1.240	9
Full feature_splc	59	1.253	10
Financial_focal_suppliers	34	1.298	11
Financial_focal_splc	53	1.385	12

As seen in Table 4, the top five types of indicator combination ranking downwards in terms of the minimum prediction errors for the test set are the *full feature* datasets, the *full feature_focal* datasets, the *financial_focal* datasets, the *trends_wiki_news_focal* datasets, and the *full feature_customers* datasets, where the actual error of the *full feature* datasets is made the smallest, that is, 0.709. This suggests that the largely expanded dataset that has the biggest number and most varied indicator data derived from multiple sources yields the best predictability. In other words, the addition of the indicators of the supply chain data and network activity data is of great significance to the CCR prediction, even though CCR is measured by the conventional method.

However, as mentioned before, more indictors or more dimensions of the target datasets may diminish the predictability of the traditional linear prediction models because a target dataset with more dimensions has indicators with more nonlinear relationship, so OLS is not an ideal analyzing tool for the datasets with more nonlinear indictors. It is right the case here that the combinations with less errors resulted from OSL tended to have a less number of indicators, between 11 to 30 out of the top five databases, with only one exception, presented in Table 4, while the databases with more indicators did not rank high in the predictability, not as they had been assumed in theory, and too in theory, this failure is likely caused by the constraints of OSL to the linear fitting between variables in processing data and then to counter the constraints and upgrade their predictability, SVM algorithm was introduced, for its being more capable of working with the datasets of high-dimensional and nonlinear features.

4.2.2 SVM evaluation

Displayed in Table 5 are the R results of SVM analysis of the predictability of the12 indicator combinations.

Table 5. The results of SVM evaluation

Datasets	Number of Indicators	MAPE	Ranking
Full feature	87	0.399	1
Full feature_suppliers	30	0.419	2
Full feature_focal	20	0.42	3
Financial_focal	30	0.53	4
Financial_trends_wiki_news_focal	25	0.53	5
Full feature_splc	59	0.533	6
Financial_focal_customers	34	0.538	7
Financial_focal_suppliers	34	0 .542	8
Financial_focal_splc	53	0.542	9
Basic financial_focal	15	0.584	10
Full feature_customers	30	0.594	11
Trends_wiki_news_focal	11	0.632	12

In regard of the errors by the SVM algorithm, there is some alteration among these databases in the ranking places, as seen form Table 4 and Table 5, but the *full feature* dataset again tops the list, and compared with its OLS result (0.709), there is a big drop in the MAPE value (0.399), which strongly manifests that the expanded datasets have a better predictability. And what's more, each and every SVM error of these 12 combinations are seen a great reduction compared to their OLS results, and the rank of the *full feature_splc* dataset, which has the second largest number of indicators (59) out of the 12 datasets, shifted from the 10th place by the OLS algorithm to the 6th by the SVM algorithm, and these are remarkably forceful evidence that the SVM algorithm is better at processing the nonlinear datasets and mining the significant indicators of high-dimensional datasets and thus does highly improve their CCR predictability.

Table 6. Three Indexes presented by SVM algorithm

Datasets	MAPE		MSE		RMSE	
	Error	Ranking	Error	Ranking	Error	Ranking
Full feature	0.399	1	59.513	1	7.714	1
Full feature_suppliers	0.419	2	112.07	10	10.586	10
Full feature_focal	0.42	3	77.29	3	8.792	3
Financial_focal	0.53	4	74.32	2	8.621	2
Financial_trends_wiki_news_focal	0.53	5	112.68	11	10.615	11
Full feature_splc	0.533	6	109.23	4	10.451	4
Financial_focal_customers	0.538	7	119.19	12	10.917	12
Financial_focal_suppliers	0.542	8	111.79	8	10.573	8
Financial_focal_splc	0.542	9	111.17	6	10.544	6
Basic financial_focal	0.584	10	111.97	9	10.581	9

Full feature_customers	0.594	11	111.31	7	10.55	7
Trends_wiki_news_focal	0.632	12	110.28	5	10.501	5

For the last time to verify in what way the CCR prediction of the 12 datasets can be impacted by the expansion of indicators and their source dimensions, an enhanced yet quite straightforward comparison was accomplished between their MAPE, MSE and RMSE values, as shown in Table 6.

Whether it is out of MAPE, MSE and or RMSE evaluation, the full feature dataset outperforms all other datasets ranked by the value of error. Moreover, the lopsided datasets, the ones without the indicators from either supply chain data or network activity data showed relatively low predictability. For example, *financial_focal_customers* falls to the bottom of the list in both MSE and RMSE error value. This indicates that the indicators from either network activity data or supply chain data may improve the CCR predictability, or arguably, the data from supply chain or network activity may make up for the absence of financial reports that should have been regularly released but quite often have not, by some small and medium-sized enterprises in particular, when any institution or investor is in need of a CCR profile of considerable credibility.

5. Conclusions

This study has revealed with solid clarity that the expansion of the data sources from the network activity data and the supply chain of the target company will upgrade the predictability of its CCR, and this effect of expansion of the data sources is verified by integrating the Delphi method with intelligent algorithms, of which SVM outperforms OSL in the predicative capability.

First, a more inclusive database (the full feature database of 87 indicative variables as a typical one in this study) that has data of supply chains and networks apart from financial sources, or the financial data of partners apart from the target company exclusively, i.e. data of more varied features, into the CCR prediction of the target company is indeed effective in greatly reducing the errors of the predictability, and it is applicable, more meaningfully, to companies with limited financial information. In cases where some small and medium enterprises do not disclose their financial data regularly or worse, some companies manipulate their financial data, the supply chain data and network data proposed in this study can be used as alternatives or bias countermeasure in their CCR evaluation. Second, more advanced intelligent algorithms are found to help improve the CCR prediction when compared to the traditionally used ones, for the traditional method requires the datasets obey the assumptions of linear separability and independence of variables, thus incompetent in nonlinear or more complex relationships between variables that are very likely to happen with the expansion of the databases. A rationally designed SVM algorithm model was run in this study on largely expanded databases and its results showed that, SVM would offer a much less predictive errors than OLS, a conventionally used model, in CCR prediction. Finally, in selecting the predicative variables and arranging them into logical combinations for running the algorithm model, Delphi method was

introduced and by the way, the study completed itself from both the quantitative and qualitative perspective. Thus far this study has made its claim that adding relevant information and using powerful intelligent algorithms are well proved to help predict the CCR and remarkably, the full dataset containing financial data, network activity data and supply chain data has the best CCR predictability. And this enhanced CCR analysis is surely of great significance in practice, for instance, network activity data can, help business companies to be alerted by early warnings of some possible impairment to its financial stability, and supply chain data can help protect the extensive corporate structure of the target company in case that there would be risk spillovers along the supply chain, and for individual bonder buyers on the market, the efficient and timely risk warning indicators of any type must be the best insurance for their investment, sustaining their interest in investing.

While a novel model or approach of incorporating supply chain data and network activity data into CCR prediction was proposed and effectively run, this study suffered from some limitations. First of all, part of the main purpose of this study was to verify whether the supply chain data would improve the CCR predictability, and when exploring the supply chain indicators, a simple weighted average was calculated in the dataset, but in practice, even a small improvement in forecast accuracy would significantly eliminate the credit risk. Therefore, more fine-grained supply chain data could be used to help companies predict CCR in the future. Secondly, although this study used the SVM algorithm to process datasets and verified that the algorithm could achieve less computational errors, the role of other machine learning algorithms (such as deep learning and reinforcement learning) in improving predictability is worthy of further attention. Finally, there might be more data types out there for future research attention and for the even better CCR predictability.

References

Agca, S., Babich, V., et al. (2021). Credit shock propagation along supply chains: Evidence from the CDS market. *Management Science, 0*(0).

AlShamsi, M., Salloum, S. A., et al. (2021). Artificial intelligence and blockchain for transparency in governance. In *Artificial intelligence for sustainable development: Theory, practice and future applications* (pp. 219-230): Springer

Bazarbash, M. (2019). *Fintech in financial inclusion: machine learning applications in assessing credit risk*: International Monetary Fund.

Belton, I., MacDonald, A., et al. (2019). Improving the practical application of the Delphi method in group-based judgment: A six-step prescription for a well-founded and defensible process. *Technological Forecasting and Social Change, 147*, 72-82.

Bequé, A., & Lessmann, S. (2017). Extreme learning machines for credit scoring: An empirical evaluation. *Expert Systems with Applications, 86*, 42-53.

Bilginol, K., Denli, H. H., et al. (2015). *Ordinary Least Squares Regression Method Approach for Site Selection of Automated Teller Machines (ATMs).* Paper presented at the Spatial Statistics Conference, Avignon, France.

Bonsall IV, S. B., Holzman, E. R., et al. (2017). Managerial ability and credit risk assessment. *Management Science, 63*(5), 1425-1449.

Brady, S. R. (2015). Utilizing and Adapting the Delphi Method for Use in Qualitative Research. *International Journal of Qualitative Methods, 14*(5), 1609406915621381.

Cervantes, J., Garcia-Lamont, F., et al. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing, 408*, 189-215.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A Library for Support Vector Machines. *Acm Transactions on Intelligent Systems and Technology, 2*(3), 1-27.
Chen, N., Ribeiro, B., et al. (2016). Financial credit risk assessment: a recent review. *Artificial Intelligence Review, 45*(1), 1-23.
El Kafrawy, P., Fathi, H., et al. (2021). An Efficient SVM-Based Feature Selection Model for Cancer Classification Using High-Dimensional Microarray Data. *IEEE Access, 9*, 155353-155369.
Figini, S., Bonelli, F., et al. (2017). Solvency prediction for small and medium enterprises in banking. *Decision Support Systems, 102*, 91-97.
Florez-Lopez, R., & Ramon-Jeronimo, J. M. (2015). Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal. *Expert Systems with Applications, 42*(13), 5737-5753.
Fracassi, C., Petry, S., et al. (2016). Does rating analyst subjectivity affect corporate debt pricing? *Journal of Financial Economics, 120*(3), 514-538.
García, V., Marqués, A. I., et al. (2019). Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction. *Information Fusion, 47*, 88-101.
Ghaddar, B., & Naoum-Sawaya, J. (2018). High dimensional data classification and feature selection using support vector machines. *European Journal of Operational Research, 265*(3), 993-1004.
Ghasemkhani, H., Reichman, S., et al. (2015). *Using Predictive Analytics to Reduce Uncertainty in Enterprise Risk Management*. Paper presented at the Thirty Sixth International Conference on Information Systems, Fort Worth, the United States. .
Gilchrist, S., & Mojon, B. (2018). Credit risk in the euro area. *The Economic Journal, 128*(608), 118-158.
Grover, V., Chiang, R. H., et al. (2018). Creating strategic business value from big data analytics: A research framework. *Journal of Management Information Systems, 35*(2), 388-423.
Gu, J., Xia, X., et al. (2017). An approach to evaluating the spontaneous and contagious credit risk for supply chain enterprises based on fuzzy preference relations. *Computers & Industrial Engineering, 106*, 361-372.
Hajek, P., & Michalak, K. (2013). Feature selection in corporate credit rating prediction. *Knowledge-Based Systems, 51*, 72-84.
Huang, M.-H., & Rust, R. T. (2021). A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science, 49*(1), 30-50.
Huang, X., Liu, X., et al. (2018). Enterprise credit risk evaluation based on neural network algorithm. *Cognitive Systems Research, 52*, 317-324.
Jiang, H., Ching, W.-K., et al. (2018). Stationary Mahalanobis kernel SVM for credit risk evaluation. *Applied Soft Computing, 71*, 407-417.
Ladhari, R., Rioux, M. C., et al. (2019). Consumers' motives for visiting a food retailer's Facebook page. *Journal of Retailing and Consumer Services, 50*, 379-385.
Lee, C.-H., Yang, H.-C., et al. (2021a). Enabling Blockchain Based SCM Systems with a Real Time Event Monitoring Function for Preemptive Risk Management. *Applied Sciences, 11*(11), 4811.
Lee, J. W., Lee, W. K., et al. (2021b). Graph convolutional network-based credit default prediction utilizing three types of virtual distances among borrowers. *Expert Systems with Applications, 168*, 114411.
Lei, J., Qiu, J., et al. (2021). Credit risk spillovers and cash holdings. *Journal of Corporate Finance, 68*, 101965.
Lev, B. (2018). The deteriorating usefulness of financial report information and how to reverse it. *Accounting and Business Research, 48*(5), 465-493.
Li, S., Shi, W., et al. (2021). A deep learning-based approach to constructing a domain sentiment lexicon: a case study in financial distress prediction. *Information Processing & Management, 58*(5), 102673.
Li, W., Liang, Y., et al. (2020). *Research on Security Risk Assessment Based on the Improved FAHP.* Paper presented at the IOP Conference Series: Materials Science and Engineering, Wuhan, China.
Liu, X. (2020). A visualization analysis on researches of internet finance credit risk in coastal area. *Journal of Coastal Research, 103*(SI), 85-89.
Lu, S. X., Wang, X. Z., et al. (2004, Aug 26-29). *A comparison among four SVM classification methods: LSVM, NLSVM SSVM and NSVM.* Paper presented at the International Conference on Machine Learning and Cybernetics, Shanghai, China.
Mahmoudi, N., & Duman, E. (2015). Detecting credit card fraud by modified Fisher discriminant analysis. *Expert Systems with Applications, 42*(5), 2510-2516.
Mansi, S. A., Maxwell, W. F., et al. (2012). Bankruptcy prediction models and the cost of debt. *The Journal of Fixed Income, 21*(4), 25-42.
Martín-Oliver, A., Ruano, S., et al. (2020). How does bank competition affect credit risk? Evidence from loan-level data. *Economics Letters, 196*, 109524.
Martín, A. G., Fernández-Isabel, A., et al. (2021). Suspicious news detection through semantic and sentiment measures. *Engineering Applications of Artificial Intelligence, 101*, 104230.

Ming, Z. (2015). *Research on Least Squares Support Vector Machines Algorithm.* Paper presented at the International Industrial Informatics and Computer Engineering Conference (IIICEC), Xian, China.
Moat, H. S., Curme, C., et al. (2013). Quantifying Wikipedia Usage Patterns Before Stock Market Moves. *Scientific reports, 3*(1), 1801.
Moat, H. S., Olivola, C. Y., et al. (2016). Searching choices: Quantifying decision-making processes using search engine data. *Topics in cognitive science, 8*(3), 685-696.
Ni, D., Xiao, Z., et al. (2020). A systematic review of the research trends of machine learning in supply chain management. *International Journal of Machine Learning and Cybernetics, 11*(7), 1463-1482.
Ni, D., Xiao, Z., et al. (2021). Machine learning in recycling business: an investigation of its practicality, benefits and future trends. *Soft Computing, 25*(12), 7907-7927.
Niemann, M., Schmidt, J. H., et al. (2008). Improving performance of corporate rating prediction models by reducing financial ratio heterogeneity. *Journal of Banking & Finance, 32*(3), 434-446.
Oreski, S., Oreski, D., et al. (2012). Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. *Expert Systems with Applications, 39*(16), 12605-12617.
Papouskova, M., & Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision Support Systems, 118*, 33-45.
Pławiak, P., Abdar, M., et al. (2020). DGHNL: A new deep genetic hierarchical network of learners for prediction of credit scoring. *Information Sciences, 516*, 401-418.
Preis, T., Moat, H. S., et al. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific reports, 3*(1), 1-6.
Samoilenko, A., & Yasseri, T. (2014). The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics. *EPJ data science, 3*(1), 1-11.
Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*(3), 1-21.
Schwieterman, M. A., Miller, J., et al. (2020). Do Supply Chain Exemplars Have More or Less Dependent Suppliers? *Journal of Business Logistics, 41*(2), 149-173.
Scott III, R. H. (2010). Bloomberg 101. *Journal of Financial Education*, 80-88.
Shen, F., Zhao, X., et al. (2019). A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation. *Physica A: Statistical Mechanics and its Applications, 526*, 121073.
Sulyok, M., Ferenci, T., et al. (2021). Google Trends Data and COVID-19 in Europe: Correlations and model enhancement are European wide. *Transboundary and Emerging Diseases, 68*(4), 2610-2615.
Tang, L., Cai, F., et al. (2019). Applying a nonparametric random forest algorithm to assess the credit risk of the energy industry in China. *Technological Forecasting and Social Change, 144*, 563-572.
Teles, G., Rodrigues, J., et al. (2020). Artificial neural network and Bayesian network models for credit risk prediction. *Journal of Artificial Intelligence and Systems, 2*(1), 118-132.
Trustorff, J.-H., Konrad, P. M., et al. (2011). Credit risk prediction using support vector machines. *Review of Quantitative Finance and Accounting, 36*(4), 565-581.
Ul Hassan, E., Zainuddin, Z., et al. (2017). A review of financial distress prediction models: logistic regression and multivariate discriminant analysis. *Indian-Pacific Journal of Accounting and Finance, 1*(3), 13-23.
Wang, G., Gunasekaran, A., et al. (2016). Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International journal of production economics, 176*, 98-110.
Wang, G., & Ma, J. (2011). Study of corporate credit risk prediction based on integrating boosting and random subspace. *Expert Systems with Applications, 38*(11), 13871-13878.
Wang, L., Chen, Y., et al. (2020). Imbalanced credit risk evaluation based on multiple sampling, multiple kernel fuzzy self-organizing map and local accuracy ensemble. *Applied Soft Computing, 91*, 106262.
Wang, L., Jia, F., et al. (2022). Forecasting SMEs' credit risk in supply chain finance with a sampling strategy based on machine learning techniques. *Annals of Operations Research*, 1-33.
Wang, M., & Ku, H. (2021). Utilizing historical data for corporate credit rating assessment. *Expert Systems with Applications, 165*, 113925.
Wu, H.-C., Hu, Y.-H., et al. (2014). Two-stage credit rating prediction using machine learning techniques. *Kybernetes, 43*(7), 1098-1113.
Wu, J., Zhang, Z., et al. (2022). Credit rating prediction through supply chains: A machine learning approach. *Production and Operations Management, 31*(4), 1613-1629.
Wu, L., & Brynjolfsson, E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales. In *Economic analysis of the digital economy* (pp. 89-118): University of Chicago Press.
Yang, R. J., Yu, L., et al. (2020). Big data analytics for financial Market volatility forecast based on support vector machine. *International Journal of Information Management, 50*, 452-462.
Yang, Y., Gu, J., et al. (2016). Credit risk evaluation based on social media. *Environmental Research, 148*, 582-585.

Yu, B., Li, C., et al. (2022). Forecasting credit ratings of decarbonized firms: Comparative assessment of machine learning models. *Technological Forecasting and Social Change, 174*, 121255.
Zamore, S., Ohene Djan, K., et al. (2018). Credit risk research: Review and agenda. *Emerging Markets Finance and Trade, 54*(4), 811-835.
Zhang, J., & Ibrahim, M. (2005). A simulation study on SPSS ridge regression and ordinary least squares regression procedures for multicollinearity data. *Journal of Applied Statistics, 32*(6), 571-588.
Zhang, T., Zhang, W., et al. (2018). Multiple instance learning for credit risk assessment with transaction data. *Knowledge-Based Systems, 161*, 65-77.
Zhang, W., Yan, S., et al. (2022). Credit risk prediction of SMEs in supply chain finance by fusing demographic and behavioral data. *Transportation Research Part E: Logistics and Transportation Review, 158*, 102611.
Zhang, Z., Gao, G., et al. (2014). Credit risk evaluation using multi-criteria optimization classifier with kernel, fuzzification and penalty factors. *European Journal of Operational Research, 237*(1), 335-348.
Zhang, Z., He, J., et al. (2019). Sparse multi-criteria optimization classifier for credit risk evaluation. *Soft Computing, 23*(9), 3053-3066.
Zhu, Y., Xie, C., et al. (2016). Predicting China's SME Credit Risk in Supply Chain Finance Based on Machine Learning Methods. *Entropy, 18*(5), 195.
Zhuang, M., Zhu, W., et al. (2021). Research of influence mechanism of corporate social responsibility for smart cities on consumers' purchasing intention. *Library Hi Tech*(ahead-of-print).