Intelligent Environments 2019
A. Muñoz et al. (Eds.)

doi:10.3233/AISE190020

Predicting the Interbank Capital Adequacy Level Based on Financial Data Analysis

Yaojun DU [a], Fangjun WANG [a] and Liping SHEN [a]
[a] *School of Electronic Information and Electrical Engineering Shanghai Jiao Tong University, Shanghai, China*

Abstract. The adequacy of interbank capital is affected by many factors, including the tightness of the entire financial market and the ups-and-downs of interest rates. It has great significance for commercial banks and other non-bank institutions to diagnose the disturbance factors of interbank capital, and predict future capital adequacy level in advance to deploy appropriate countermeasures accordingly. This paper attempts to analyze the relevant factors affecting the interbank capital and to predict the adequacy level of interbank capital based on structured and unstructured financial data. For unstructured data, we crawl the texts from Sina Financial News and then make pre-processing, including word segmentation, emotional word extraction and word-to-vector transformation. For structured data the preprocessing includes padding missing value, data normalization, feature selection and data dimensionality reduction. The prediction models we tried include GBDT, XGBoost, LSTM, SVM, and Perceptron. Experiments show that two-category (loose and tight) average accuracy of the overall adequacy level of interbank capital can achieve more than 94.5%.

Keywords. Interbank Capital Adequacy Level, Prediction, Natural Language Processing (NLP), Financial Data Analysis

1. Introduction

Commercial banks have a special and important role in the entire financial system and even in the national economy. The capital adequacy level is a measure of a bank's available capital to protect depositors and promote the stability and efficiency of financial systems around the world. It usually indicates money supply and the ability of the market regulation policy to support financial products. The indicators include broad money supply (M2), stamp duty, central bank interest rates, etc. The prediction of the capital adequacy of commercial banks can provide decision-making support for the assets allocation, risk control and interbank lending, and enhance the liquidity of the inter-bank money market, and very important, plays a positive role in the early warning of financial risks.

With the rapid development of information technology, big data analysis and artificial intelligence, the financial industry is also actively trying to use new technologies to solve traditional problems. At present, academic researches on financial market mostly focus on the analysis of structured data, and few research studies the prediction of capi-

tal adequacy level from the perspective of banks. The state-of-art technology of big data and artificial intelligence can automatically train interesting models from a large amount of structured and unstructured data, which breaks through the limitation of traditional predicting methods and greatly improve the prediction accuracy.

In this paper, we explore several machine learning methods on traditional structured financial data and text-based financial news, to predict the capital adequacy level of banks. Section 2 introduces the workflow and the related work. Section 3 introduces the prediction architecture and processing of the structured and unstructured data. Section 4 explains the experiment process and five kinds of machine learning algorithms explored in our study. Section 5 analyzes and compares the experimental results.

2. Related Work

Researches on the prediction of adequacy of capital and liquidity of commercial banks has been widely undertaken in recent decades. Diamond, Dybvig [1] (1983) believed that commercial banks were based on liquidity conversion, providing capital liquidity to the market while facing a liquidity crisis caused by tight capital and liquidity gaps. The paper of Matz, Neu [2] (2007) put forward a pressure test model based on bank balance sheet, pointed out that banks should set the pressure scenario index of balance sheet based on historical data in the pressure test, and finally estimated the expected cash flow that banks can afford under different pressure scenarios. Since January 4th, 2007, China has officially operated the Shanghai Interbank Offered Rate (Shibor) to promote the rapid development of the money market. Shibor is a barometer of the adequacy of bank capital, with Shibor upward representing the tight capital market and the reverse is the loose capital market.

With the rapid development of artificial intelligence technology, many researches began to use artificial intelligence technology to study financial problems, especially using text-based data. As one of the classic scenes in the field of natural language processing (NLP), text categorization has accumulated a large number of technical implementation methods. The implementation approach can be roughly divided into two categories: text classification based on traditional machine learning and text classification based on deep learning. R. Batra, S. M. Daudpota [3] (2018) focus on techniques involving sentiment analysis in predicting stock trends. Fuli Feng et al. [4] predict the stock movement with a new machine learning solution. Xiao Ding et al. [5] (2014) applied the deep learning method to predict stock. They proposed a new method of event extraction, which extracted events from the news as input to the neural network. Sundermeyer et al. [6] (2012) used the LSTM (Long Short-Term Memory) unit to construct a language model, and discovered its potential in the language model.

3. Prediction Architecture

This paper is based on both the traditional and deep learning text classification. Fig.1 shows the workflow of the prediction, which is composed of Data Processing, Model Training & Predicting and Result Evaluation & Analysis. The Data Processing is composed of Data Preprocessing, Feature & Keywords Extraction and Data Dimension Re-

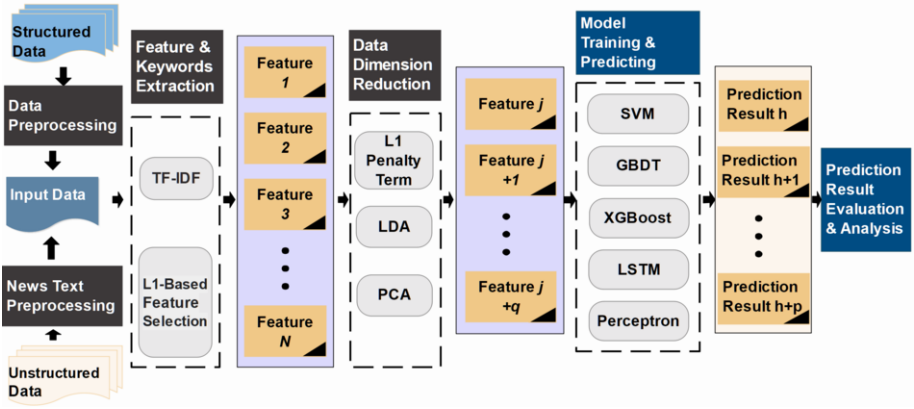

Figure 1. Workflow of Prediction.

duction. The input is structured and unstructured data. First, features are extracted from original unstructured data. The feature extraction methods include frequency method, TF-IDF, mutual information method N-gram, word embedding and empirical manual design. Then, the features extracted from text, together with structured data, are fed into different machine learning models to do prediction. The output is capital adequacy levels.

Table 1. The Adequacy Indicators of Bank Capital.

Warning Color	Adequacy Level	Adequacy Indicators
Green	Normal	1. Major counterparties have abundant capital 2.The supply of overnight capital is sufficient and overnight price financed by non-bank institutions is not high 3. Ample money supplied by policy backs to X-REPO
Yellow	Tight	1. Major counterparties have 1/3 shrinkage of money supply or stop lending 2. The overnight capital limited supply or financed overnight capital with high prices for non-bank institutions 3. Policy banks reduce the overnight capital supply on X-REPO
Orange	Very Tight	1. Major counterparties have 2/3 shrinkage of money supply or stop lending 2. The overnight capital supply is scarce or financed long term capital with high prices for non-bank institutions 3. Policy banks do not supply the overnight capital on X-REPO
Red	Extremely Tight	1. Major counterparties stop lending 2. Banks do not lend money to non-bank institutions

4. Data Processing

4.1. Data Description

In view of the measurement of the adequacy level of commercial banks capital, we give the early warning level of capital and the definition indicators as shown in Table 1. The capital is divided into four levels: Normal, Tight, Very Tight and Extremely Tight. Due to the liquidity risk has become the most fundamental and fatal risk since 2007, while many financial institutions and commercial banks went bankrupt or closed down. As Basel Committee issued Basel III in 2010 and China Banking Regulatory Commission issued The Management Measures on Commercial Bank Liquidity Risk in 2011, liquidity risk regulation of banking industry is strengthened. Since that, in the study of this paper, the focus is on the situation of tight capital. And due to the uneven distribution of sample categories, the red (29 days) and yellow (41 days) samples are too few, resulting in low predicting accuracy. Therefore, the four-classification problem is downsized into two classifications, where the normal adequacy level and tight adequacy level are grouped into loose category, and very tight adequacy level and the extremely tight adequacy level are grouped into tight category. The capital adequacy warning level data used in this paper is the natural trading days data between August 1, 2014 and May 11, 2018 a total of 1087 days. The dataset of the news text (unstructured data) used in this experiment is crawled from four sections of Sina Financial news website[1] namely macro, data, central bank and market segment, as shown in Table 2.

The dataset of the structured data used in the experiment is provided by Wind Financial Database[2] which is a powerful tool for financial information services .In the experiment, the data frequency includes daily, weekly, monthly, seasonal, semi-annual and annual. The following Table 3 lists some of the factors.

Table 2. News Text Examples.

Date	News Text
2018/3/15	US dollar against the Canadian dollar rose above 1.3044, the highest in the last eight months.
2018/3/16	Offshore Renminbi (CNH) was quoted at 6.3293 yuan against a US dollar at 04:59 Beijing time, compared with late Wednesday in New York fell 223 points.

Table 3. The Adequacy Indicators of Bank Capital.

NO.	Indicators	Data Frequency
1	Spot Exchange Rate: RMB/USD	Daily
2	Banking Industry Climate Index	Seasonal
3	RMB Required Reserve Ratio	Monthly

[1] http://finance.sina.com.cn/7x24/

[2] http://www.wind.com.cn

4.2. Unstructured Data Preprocessing

4.2.1. Sentence Segmentation

In order to judge whether there are corresponding words in the emotional dictionary in the sentence, we need to cut the sentence accurately into words, namely the automatic segmentation of the sentence. After comparing the existing Word segmentation tools, considering the accuracy and the ease of use on the Python platform, we finally chose the Jieba Chinese word segmentation [7] as our word segmentation tool. The results of word segmentation examples are shown in Table 4.

4.2.2. Word Vectorization

After sentence segmentation, Word2Vec [8] is used to produce high-dimensional vectors (Word Embedding) to represent the words, and converts the samples into word sequence vectors. In the experiment, we do the word vectorization via calling the function gensim.models.word2vec, which takes the news text as input and produces the word vectors as output. And a sentence vector is the average of all the words it contains. In this procedure, the whole text corpus was mapped into a 300-dimensional vector space, where similar words are nearer than others.

Table 4. Sentence Segmentation Examples.

Date	Segmentation Example
2018/3/15	US dollar against the Canadian dollar rose above 1.3044 the highest in the last eight months
2018/3/16	Offshore Renminbi (CNH) was quoted at 6.3293 yuan against a US dollar at 04:59 Beijing time

4.3. Structured Data Preprocessing

4.3.1. Imputation

Since there are some missing values in the structured dataset of this paper, it is a common practice to delete all the relevant rows and columns of the data if there is a missing value, resulting in the consequence that important features lose easily. Therefore, imputation occupies a significant place in the preprocessing stage. This paper fills in the missing values with the pandas.DataFrame.fillna function, using *pad* (padding the missing value with the previous non-missing value) and *bfill* (filling the missing value with the next non-missing value) modes.

4.3.2. Data Normalization

Normalization is the standardized processing of all structured data to eliminate the dimensional impact between various indicators. The purpose of this procedure is to make the original data of the indicators in the same order of magnitude under comprehensive comparative evaluation. In this paper, the Min-Max normalization method is used to linearly transform the original data so that all the values are mapped between [0-1]. This

paper implements the normalization method by Sklearn.preprocessing.MinMaxScaler class:

$$x^{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

where x_{min} is the minimum value of the sample data, and x_{max} is the maximum value of the sample data.

4.4. Feature and Keywords Extraction

4.4.1. Unstructured Data Feature Extraction

The selection of feature entries and their weights is called the feature extraction of target samples, and the advantages and disadvantages of feature extraction will directly affect the operation effect of the model. Except the word vectors, we extract two other kinds of features: word frequency and keywords.

Here the TF-IDF (Term Frequency-Inverse Document Frequency) [9] algorithm is used for word frequency analysis to evaluate the importance of a term in the news text corpus. The importance of a word increases proportionally with the number of times it appears in a text, but at the same time decreases inversely with the frequency it appears in the corpus. The top 30 words are selected as the input of recognition models. Table 5 lists the top 4 words.

Table 5. TF-IDF Word Frequency Analysis.

No.	TOP Words	Weights
1	Year-on-year	0.075403657
2	Increase	0.064093650
3	Trillion	0.055898826
4	Interest rate	0.050473410

Sentiment analysis can classify the polarity of the news text and determine whether the expressed opinion is positive, negative or neutral. In this paper, we define four keywords groups, which are key_words, pos_words, neg_words and non_words. The key_words are the keywords extracted in a half-automatic way: first the keywords are selected by matching the key nouns in the news text with financial dictionary; then they are checked and filtered by students from financial major; the pos_words and neg_words are polar verbs and adjectives, with the words of positively or negatively affecting the capital adequacy level. These words are extracted manually with professional knowledge in the financial field and a large amount of reading on Sina news text; non_words are the privative words, such as no and none. Some of the keywords examples as shown in Table 6. If a news text contains keywords and positive words, it was thought to be a positive news, and was labeled as 1. Likewise, negative and neutral news were labeled -1 and 0 respectively. News would be thought to be neutral if it contains none of these words.

Table 6. Keywords Examples.

Name	Words
key_words	'Capital', 'Liquidity', 'Cash Flow', 'Funds', 'Cash', 'Monetary Policy'
pos_words	'Suffient', 'Putting Currency', 'Quantitative Easing', 'Easy Monetary Policy'
neg_words	'Tight Monetary Policy', 'Tight', 'Tight Fiscal Policy', 'Tighten', 'Tight money'
non_words	'Not Tight', 'Not Loose', 'Temporary', 'Unintentional', 'No', 'Neutral'

4.4.2. Structured Data Feature Extraction

Feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. In this paper, the feature is selected by embedded method (Embedded). Firstly, various of machine learning models are trained to obtain weight coefficients of each features. And then the features are selected according to the coefficient from large to small. Then the features are selected by using the base model with the penalty term, which is implemented by combining the SelectFromModel class of the Sklearn.feature_selection library with the logistic regression model and L1 penalty term.

4.5. Data Dimension Reduction

After feature extraction, the model can be trained directly, but it may be necessary to reduce the feature matrix dimension because the feature matrix is too large, which leads to the problem of complicated calculation and long training time.

We compared three methods of the dimensionality reduction, namely LDA (Linear Discriminant Analysis), PCA (Principal Component Analysis) and L1 penalty term. LDA is a supervised learning method, which considers the classification label information and seeks the direction with best classification performance. In this paper, since the sample size is small and the feature dimension is large, resulting in the inability to obtain the optimal projection direction. PCA is an unsupervised learning method, which performs a linear mapping of the data to a lower-dimensional space, while does not utilize any internal classification information when mapping, making classification more difficult. The dimensionality reduction method adopted in this paper is the model based on the L1 penalty term mentioned above. The principle of L1 penalty term reduction is to retain one of a plurality of features that have same relevance to the target value so that the dimensionality is reduced.

5. Model Selection and Algorithm Analysis

In this paper, five methods are used to train and predict news text (unstructured data) and structured data corpus. They are SVM (support vector machine), GBDT (Gradient Boosting Decision Tree), XGBoost (eXtreme Gradient Boosting), LSTM (long short-term memory) and Perceptron. The whole data has a total of 973 days, which were divided into a training set of 773 days and a test set of 200 days. The backtracking time window is the time period which is used for the data training and testing, while the prediction time window is the prediction time period after time stamp. The larger backtracking time window, which used for training, the wider time period for selecting data. The

larger prediction time window, which used for testing, the more difficult prediction and the lower accuracy. The length of the time window is generally chosen in conjunction with experience and actual computing needs. In this paper, backtracking time window is set to 28 days and the prediction time window is 7 days. That is to say, using known information of 28 days before to predict target value of 7 days after.

In the experiment, SVM combines text data through word vectors with structured data to predict the target. GBDT algorithm can capture the context of the word to some extent, for example, to identify whether news text will lead to the tight capital level, the text appears ”Cash Flow” along with words like ”abundant” and ”released” leading to the reduction of the probability of tight capital. XGBoost is equivalent to a logistic regression with L1 and L2 regularization terms, which improves the accuracy of the model. LSTM networks is suitable for processing and predicting important events with very long intervals and delays in the time series, and the number of nodes per hidden layer is set to 10, and the number of layers is set to 10, and iteration time is 5000. Perceptron is an algorithm for supervised learning of binary classifiers. A binary classifier is a function which can decide whether or not an input, represented by a vector of numbers, belongs to some specific class. In this experiment, the model is simpler and consists of two full-connected layers [10]. An array of keywords is also defined to strengthen the model. It is divided into two steps. The first one is that the model only uses structured data for training and test, and the other is to add the news text data on the basis of structured data, by defining the Keyword Array.

6. Result Analysis

In the case of imbalance distribution of samples (there are very few red and yellow samples), the error rate resulted from model over-fitting is particularly large. The accuracy of green is very high, while the accuracy of other three categories is very low. The experimental results are shown in Table 7. As can be seen from the predicted results, because there are too few red and yellow samples, their predicting accuracy is 0.0, while the orange accuracy rate is only about 0.044.

Table 7. Predicting Results of Four-Classification.

NO.	Adequacy Level	Accuracy	NO.	Adequacy Level	Accuracy
1	Green	0.94615338	3	Orange	0.04444444
2	Yellow	0.00000000	4	Red	0.00000000

In the two-class capital adequacy level predict, this paper trained and tested five models of SVM, GBDT, XGBoost, LSTM and Perceptron, and uses ten cross-validation analysis, and then used the mean value as the model accuracy rate. The experimental results are shown in Table 8.

By comparing and analysis of the predicting accuracy of the above five models, we can find that the results predicted by simple perceptron are much better than those of other more complex models. The main reason is that the complexity of the research scene in this paper is too high, and the amount of data is not much, which leads to over-fitting of the complex models such as SVM, GBDT, XGBoost and LSTM, and the generalization

Table 8. Prediction Accuracy Comparison of Five Models.

NO.	Model	Accuracy	NO.	Model	Accuracy
1	SVM	0.650	4	LSTM	0.625
2	GBDT	0.683	5	Perceptron	0.945
3	XGBoost	0.740			

accuracy is not high. At the same time, it can also be concluded that when choosing a wide variety of complex machine learning, the specific choice of what method depends on the size of the dataset and the complexity of the problem itself. In the experiment of this paper, the effect of simple perceptron is the most effective.

The following is a detailed introduction to the experimental situation of the simple perceptron, and the test results of the perceptron are shown in Table 9.

The 'before' column indicates that the model only uses structured data for training and testing, and its testing accuracy is 92.5%; the 'after' column represents the combination of structured data, news text data and defined Keyword Groups (Table 6), and then training and testing with an accuracy of 94.5%. It can be seen that adding unstructured text data will be helpful in improving accuracy.

Table 9. Predicting Results of Perceptron.

NO.	before	after	NO.	before	after
1	0.92792793	0.95495495	6	0.927927928	0.954954955
2	0.92792792	0.936936937	7	0.90990991	0.927927929
3	0.90990991	0.963963965	8	0.918918919	0.945945948
4	0.927927928	0.936936937	9	0.927927928	0.945945947
5	0.936936937	0.936936938	10	0.936936937	0.945945946
Average	0.925225225	0.945045045			

The following is a determination of the pros and cons of the model through the ROC (Receiver Operating Characteristic) curve, which is often used to evaluate the merits of a Binary Classifier [11]. As can be seen from Fig.2, the ROC curve (before) of one classifier is completely 'below' the curve (after) of another [12], then it can be asserted that the performance of the latter is better than the former, that is, by adding the news text can effectively improve the accuracy of the predict.

7. Conclusion

In this paper, in order to improve the accuracy of predicting the level of capital adequacy of commercial banks, works are summarized as follows:

- The combination of structured and unstructured data.
- The method of extracting keyword features.
- Multiple model comparison and selection.

The average accuracy rate of predicting the level of capital adequacy of the commercial bank in the next seven days is more than 94.5%, which proves the validity of the model.

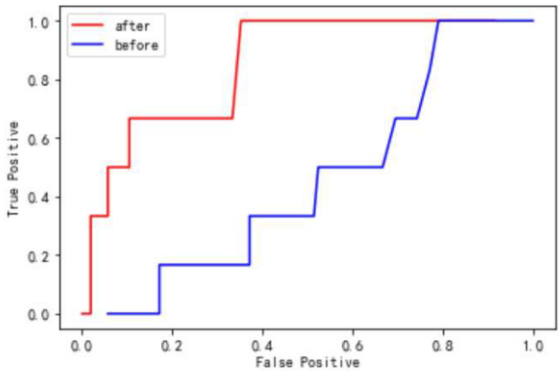

Figure 2. ROC Curve Comparison.

In the future, we consider strengthening the quality of data, such as considering the cyclicality of the economy, so that the data covers at least one full economic cycle, and collecting more unstructured training data to make more complex deep learning models applicable. Also we would consider more economics implications, such as distinguishing between the different effects of monetary and fiscal policies on the adequacy of capital.

References

[1] Diamond D W, Dybvig P H. Bank runs, deposit insurance, and liquidity[J]. Journal of political economy, 1983, 91(3): 401-419.

[2] Liquidity risk measurement and management: a practitioner's guide to global best practices[M]. John Wiley & Sons, 2006.

[3] Batra R, Daudpota S M. Integrating StockTwits with sentiment analysis for better prediction of stock price movement[C] 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET). IEEE, 2018: 1-5.

[4] Feng F, Chen H, He X, et al. Improving Stock Movement Prediction with Adversarial Training[J]. arXiv preprint arXiv:1810.09936, 2018.

[5] Ding X, Zhang Y, Liu T, et al. Using structured events to predict stock price movement: An empirical investigation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1415-1425.

[6] Sundermeyer M, Schluter R, Ney H. LSTM neural networks for language modeling[C]//Thirteenth annual conference of the international speech communication association. 2012.

[7] Tiantian Wang, "Research of feature selection and feature extraction methods in internet news classification", university of science and technology of china, 2016.

[8] Goldberg Y, Levy O. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method[J]. arXiv preprint arXiv:1402.3722, 2014.

[9] Soucy P, Mineau G W. Beyond TFIDF weighting for text categorization in the vector space model[C]//IJCAI. 2005, 5: 1130-1135.

[10] Tang J, Deng C, Huang G B. Extreme learning machine for multilayer perceptron[J]. IEEE transactions on neural networks and learning systems, 2016, 27(4): 809-821.

[11] Metz C E, Herman B A, Roe C A. Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets[J]. Medical Decision Making, 1998, 18(1): 110-121.

[12] Zhou Z H. Machine Learning[M]. Beijing: Tsinghua University Press, 2016: 23-51.