Repair missing data to improve corporate credit risk prediction accuracy with multi-layer perceptron

Abstract

Data loss has become an inevitable phenomenon in corporate credit risk (CCR) prediction. To ensure the integrity of data information for subsequent analysis and prediction, it is essential to repair the missing data as accurately as possible. To solve the problem of missing data in credit classification, this study proposes a multi-layer perceptron ensemble (MLP-ESM) model that can perform data interpolation and prediction simultaneously to predict CCR. The model makes full use of non-missing information and interpolates more missing columns with fewer missing values. In this way, not only the data features needed for missing data interpolation are extracted, but also the structural relationship features between the predicted target and the existing data are extracted, which can achieve the effect of simultaneous interpolation and prediction. The results show that the MLP-ESM model can effectively interpolate and predict the missing dataset of CCR. The prediction accuracy is 83.11%, which is better than the traditional machine learning model. This fully shows that the dataset after interpolation can achieve a better prediction effect.

Keywords

Missing data, data interpolation, machine learning, Corporate credit risk, prediction

1 Introduction

In the credit environment of a developed market economy, credit risk prediction is very important for the development of companies Sivasankar et al. (2016); (Wang et al., 2011; Yuan et al., 2022). Among them, a company's credit rating, especially under the 2014 New Basel Accord , plays a decisive role in the capital market and can be used as an important parameter in the evaluation of the corporate credit risk (CCR) (Yu et al., 2022). However, in the prediction of the CCR, some data will inevitably be lost due to the restructuring of the database, incomplete company records, and the replacement of characteristic indexes (Florez-Lopez, 2010; Moscato et al., 2021; Soldatyuk et al., 2014). In the CCR prediction, missing data will have a great impact on the prediction results (Garcia et al., 2019; Lan et al., 2021; Twala, 2013). As early as 2006, Mählmann (2006) emphasized the importance of processing missing data for the CCR prediction in their CCR prediction.

For the missing data in CCR, even a small degree of data loss can lead to a reduction in the number of effective samples, deviation in the analysis results, and the weakening of the credibility of prediction results (Budagaga, 2020; Garcia et al., 2019). If the degree of data loss is serious, it is likely that the original CCR prediction cannot be effectively predicted due to the insufficient sample size (Li et al., 2020; Zhang et al., 2016). But for the missing data processing method, relatively simple is the direct deletion method. This method is simple to operate, but it is easy to produce estimation deviation. And simple deletion can lose a lot of information and affect the accuracy of prediction results (Hooke et al., 2021; Nakagawa et al., 2008; Qiu et al., 2008). Therefore, researchers usually do not use such a method when preprocessing missing data, except in cases where the missing proportion is very low. However, for the data with low missing degree and great predictive value, data repair through data interpolation is more easily accepted by researchers (Akin Arikan et al., 2018; Catellier et al., 2005; De Silva et al., 2016).

The idea of data repair is to use existing data to mine data with predictive value or analytical value from existing samples by advanced technology (Chiang et al., 2016; Gu et al., 2017; Wei et al., 2019). Many studies used generative adversarial networks (GANs) and autoencoders to repair the optimal parameters for missing data (Campanile et al., 2020; Chang et al., 2020b; D'Angelo et al., 2021). Chang et al. (2020b) proposed the dual-domain conditional GANs to interpolate seismic data to drive the generative network to learn optimal parameters. Meanwhile, compared with traditional methods that only identify outliers, Eduardo et al. (2020) proposed a robust variational autoencoder method that can detect and repair abnormal units in the dataset. In addition, the idea of data repair has also been found in practical applications. For example, in the text classification task of natural language processing, Wei and Zou (2019) proposed

synonym replacement, random insertion, random exchange, and random deletion to repair data. In addition, in the study of image recognition, Shorten et al. (2019) proposed horizontal flipping, clipping, rotation, translation, adding interference, convolution kernel filter, image mixing, and random deletion, and achieved good data repair effects. Therefore, for the existing CCR prediction dataset, the dataset can contain extremely valuable duration and frequency information. If the data with prediction potential can be mined from the existing samples, the data can be interpolated. It will be possible to improve the accuracy of the CCR forecasts through the repair of internal data (Florez-Lopez, 2010; Li et al., 2020).

In the actual prediction of the CCR, scholars mainly discuss manual interpolation, special value interpolation and mean means interpolation based on statistics (Florez-Lopez, 2010; Nijman et al., 2021). These methods mainly interpolate the data near the missing dataset so as to realize missing data repair (Chang et al., 2020a; Guo et al., 2009; Mählmann, 2006). In recent years, with the rise of artificial intelligence, machine learning algorithms are popular because of their significant advantages in processing complex data and improving model performance (D'Angelo et al., 2021; D'Angelo et al., 2019; Sarker, 2021). These algorithm have also been applied to missing data repair and proved to be better than statistical interpolation methods (Jerez et al., 2010; Moon et al., 2019). For example, Moon et al. (2019) used multi-layer perceptron (MLP) to accurately repair the missing temperature data, and improved data interpolation accuracy. Although machine learning algorithms have been proved to be capable of missing data repair and prediction, some algorithms only perform interpolation in the same row or column of data, and the data rows or columns that can be interpolated are also severely limited by data holding amount (Chang et al., 2020a). However, in the research field of CCR prediction, due to the limitations of data cost and privacy protection, it is increasingly difficult to find more predictive indicators from the vertical dimension (Shema et al., 2019; Yap et al., 2011; Yue et al., 2016). In addition, the existing interpolation methods of integration models can be limited by data samples, which means that the problem of data interpolation under the condition of limited data samples needs to be solved urgently.

In order to solve the above problems, this study focuses on the field of CCR prediction. By mining data with predictive potential from existing samples for interpolation, it aims to build a method that can take into account both missing data interpolation and prediction for missing datasets, thus further improve the CCR prediction accuracy. The rest of the study is structured as follows. Section 2 presents the method of this study, Section 3 describes in detail the dataset and data preprocessing steps used in this study, and Section 4 analyzes the results of this study method. Finally, the conclusion of this study and the way of future research are presented in Section 5.

2 Methods

For the topic of missing data in this study, this study intends to propose a model that can perform data interpolation and prediction simultaneously. The specific process is as follows.

2.1 Theoretical design of the model

In this study, the MLP model with multi-output and strong nonlinear fitting ability is introduced into the missing data interpolation prediction model (Belue et al., 1997; Gao et al., 2012; Khoygani et al., 2016). Multi-layer perceptron and interpolation model-specific process is as follows:

(1) The structure of the MLP

MLP, also known as a feedforward neural network, is extended from the perceptron model proposed by Rosenblatt. The perceptron model consists of two layers of neurons. The input layer receives the input signal and transmits it to the output layer to obtain the output result. Figure 1 shows a simple, functional neuron structure model. Each neuron accumulates one or more weighted input values to get the cumulative value, then performs a nonlinear transformation on it using activation function, and then transfers the value to the lower neuron.

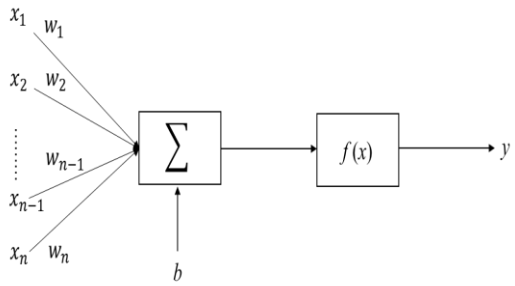

Figure 1 Functional neuron structure

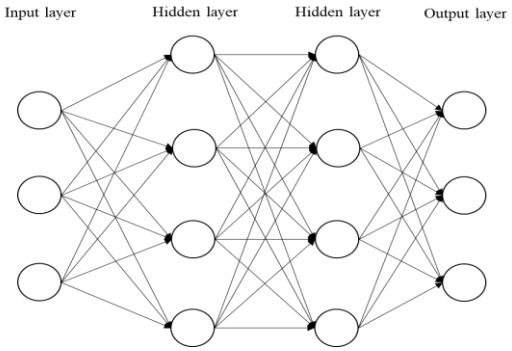

Figure 2 Multi-layer perceptron model

As the perceptron model has only one layer of functional neurons, its learning ability is very limited, and it can only complete simple linear separable problems. In order to enhance the fitting ability of the model and the ability to deal with nonlinear situations, the multi-layer perceptron model can be obtained by adding hidden layers to the perceptron model. It is mainly characterized by multiple layers, which are generally composed of one input layer, several hidden layers, and one output layer. There is a fully connected structure between the layers. Its model diagram is shown as follows:

Suppose that there is a multi-layer perceptron L layer, in which the first layer is the input layer. $\{x_1, x_2, \ldots\ldots, x_n\}$ as the input data, and the L layer to output layer.

$\{y_1, y_2, \ldots\ldots, y_m\}$ for the output data. For the l hidden layer, suppose n_l neuron. $[h_1^{(l)}, h_2^{(l)}, h_3^{(l)}, \ldots\ldots, h_{n_l}^{(l)}\}$ as the output data of the l layer. Set $w_{ij}^{(l)}$ for the jth l - 1 layer of neurons to the ith l layer neuron weights. $b_i^{(l)}$ is the bias of the ith neuron in layer L. $f(\cdot)$ is the activation function. Common activation function with sigmoid function tanh function, ReLU function, etc. The l layer of the ith neurons $o_i^{(l)}$ and an output layer of the ith neurons y_i output formula respectively as follows:

$$h_i^{(l-1)} = f(o_i^{(l-1)}\} \tag{1}$$

$$o_i^{(l)} = \sum_{j=1}^{s_{l-1}} w_{ij}^{(l)} h_j^{(l-1)} + b_i^{(l)} \tag{2}$$

$$y_i = f(\sum_{j=1}^{s_{L-1}} w_{ij}^{(L)} h_j^{(L-1)} + b_i^{(L)}) \tag{3}$$

Subsequently, the weight and threshold of each neuron at each layer can be learned by using the error inverse propagation model so as to make the output of the neural network as close to the real value as possible. The training data set is expressed as $\{(x_1, y_1), (x_2, y_3), \ldots\ldots, (x_n, y_n), y_k \in [-1,1]\}$, (x_k, y_k) is for the one training sample, multi-layer perceptron output for $\hat{y}_i$, error function of the data set is defined as follows:

$$E = \frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{y}_k)^2 \tag{4}$$

The weight W and bias B in the multi-layer perceptron can be updated iteratively according to the following formula, where α is the learning rate and its value range is (0,1).

$$W^{(l)} = W^{(l)} - \alpha\frac{\partial E}{\partial W^{(l)}} \tag{5}$$

$$B^{(l)} = B^{(l)} - \alpha\frac{\partial E}{\partial B^{(l)}} \tag{6}$$

MLP can be trained to complete some complex tasks by modifying network weights and biases.

(2) Design of interpolation process

For input dataset containing missing data, according to the data of this research suggests first arranged to organize data, and the columns of the missing parts for standardization, get missing values are arranged trapezoidal datasets $X = (x_1, x_2, \ldots, x_t, x_{t+1}, \ldots, x_{t+k})$, $x_1, x_2, \ldots, x_t$ is no missing data in the column. For the

existence of the missing data column $x_{t+1}, x_{t+2}, \dots, x_{t+k}$, k as the number of columns, which explain the index number of variables.

For any $i \le j$, number of x_i and x_j missing values is $n_i \le n_j$,.

The dataset X is divided into $k+1$ sub-datasets, respectively, $X_1 = (x_1, x_2, \dots, x_t), X_2 = x_{t+1}, \dots, X_{k+1} = x_{t+k}$

For the input X_1, a full connection layer F_1 with 8 neurons was constructed, and a full connection layer G_1 with a single neuron was constructed with the output of the full connection layer F_1. $G_1\big(F_1(X_1)\big)$ as the input to interpolate the sub-dataset X_2. And construct the loss function of interpolation

$$L_1 = \frac{1}{n_2}\sum_{i=1}^{m} g^*(x_{2i})\left(G_1\big(F_1(x_{1i})\big) - g(x_{2i})\right)^2 \quad (7)1$$

n_2 is the number of missing values in X_2, $G_1\big(F_1(x_{1i})\big)$ is the value of the dataset X_2 based on X_1 interpolation, and

$$g(x) = \begin{cases} x, & if\ x\ is\ numeral\ full \\ 0, & if\ x\ is\ umerals\ empty' \end{cases}$$

$$g^*(x) = \begin{cases} 1, & if\ x\ is\ numeral\ full \\ 0, & if\ x\ is\ umerals\ empty \end{cases}.$$

For input X_2, we construct the interpolation layer P_2, then

$$P_2(X_2) = G_1\big(F_1(X_1)\big) * \big(1 - g^*(X_2)\big) + X_2 * g^*(X_2) \quad (8)2$$

After nonlinear mapping F_2, vectors and splicing $F_1(X_1)$ are obtained, denoted as $F_1(X_1) \oplus F_2\big(P_2(X_2)\big)$, and a fully connected layer G_2 with single neurons is constructed $F_1(X_1) \oplus F_2\big(P_2(X_2)\big)$ as input so as to $G_2\left(F_1(X_1) \oplus F_2\big(P_2(X_2)\big)\right)$ interpolate sub-datasets X_3 and construct a similar loss function L_2.

In this way, the interpolation loss function $L_1, L_2, \dots, L_k$ constructed $X_2, X_3, \dots, X_{k+1}$ by pyramid interpolation can be obtained, at the same time to $F_1(X_1) \oplus F_2\big(P_2(X_2)\big) \oplus \dots \oplus F_{k+1}\big(P_{k+1}(X_{k+1})\big)$ and construct the full connection layer H_1 with 16 hidden layer neurons as the input and construct the full connection layer H_2 with the H_1 output as the input. With the output H_2 as the predicted value, the prediction model F is as follows:

$$F = H_2 \circ H_1 \circ (F_1 \oplus F_2 \circ P_2 \oplus \dots \oplus F_{k+1} \circ P_{k+1}) \quad (9)3$$

For the input vector of independent variables x, $F(x)$ is the prediction of the probability of corporate credit downgrade.

With cross-entropy as the loss function of the model, denoted as L_{pred}, then

$$L_{pred} = \frac{1}{n}\sum_{i=1}^{n} -\Big[y^i \log F\big(x_1^i, x_2^i, \dots, x_t^i\big) + \big(1-y^i\big)\log\Big(1-F\big(x_1^i, x_2^i, \dots, x_t^i\big)\Big)\Big] \quad (10)4$$

Then the loss function of this model is:

$$L = L_{pred} + \sum_{i=1}^{k} \lambda_k L_k \quad (11)5$$

λ_k is the weight of interpolation loss of each variable. The above formula indicates that the total loss of the model is the loss predicted L_{pred} by the model to the predicted target y in the sample, plus the loss L_k predicted by the missing independent variable data in the sample multiplied by the penalty weight. Therefore, the overall optimization objective of trapezoidal multi-layer perceptron is:

$$z = \min\left(L_{pred} + \sum_{i=1}^{k} \lambda_k L_k\right) \quad (12)6$$

In the subsequent experiments of this study, as it is difficult for general software packages to directly process the missing values, a large number is used to mark the missing values (after standardization, most of the values in the dataset are between -3 and 3. In this study, -10 is used to mark the missing values).

The interpolation process is shown as follows:

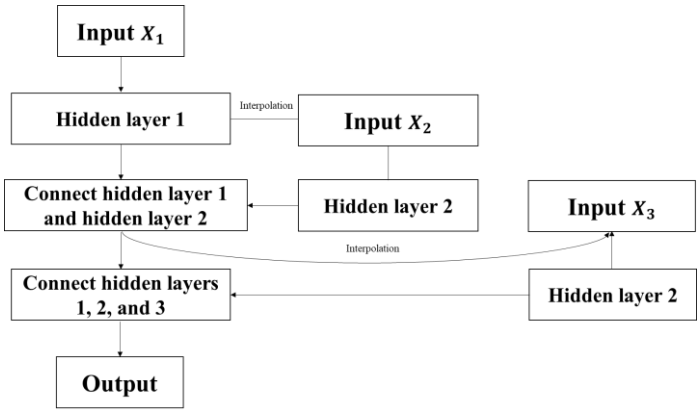

Figure 3 The interpolation process of the MLP

2.2 Model interpolation idea

The theoretical design ideas of the model proposed in this study are shown in Figure 4, which can be divided into four steps: trapezoidal arrangement of data, model building, model-based training, and model-based interpolation.

Table 1

Line item	Complete data					Missing data					
	Column 1	Column 2	Column 3	...	Column n	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
Line 1	Nf_{1-1}	Nf_{1-2}	Nf_{1-3}	...	Nf_{1-n}	Nf_{1-1}	Nf_{1-2}	Nf_{1-3}	Nf_{1-4}	Nf_{1-5}	Nf_{1-6}
Line 2	Nf_{2-1}	Nf_{2-2}	Nf_{2-3}	...	Nf_{2-n}	Nf_{2-1}	Nf_{2-2}	Nf_{2-3}	Nf_{2-4}	Nf_{2-5}	Ne_{2-6}
Line 3	Nf_{3-1}	Nf_{3-2}	Nf_{3-3}	...	Nf_{3-n}	Nf_{3-1}	Nf_{3-2}	Nf_{3-3}	Nf_{3-4}	Ne_{3-5}	Ne_{3-6}
Line 4	Nf_{4-1}	Nf_{4-2}	Nf_{4-3}	...	Nf_{4-n}	Nf_{4-1}	Nf_{4-2}	Nf_{4-2}	Ne_{4-4}	Ne_{4-5}	Ne_{4-6}
Line 5	Nf_{5-1}	Nf_{5-2}	Nf_{5-3}	...	Nf_{5-n}	Nf_{5-1}	Nf_{5-2}	Ne_{5-3}	Ne_{5-4}	Ne_{5-5}	Ne_{5-6}
Line 6	Nf_{6-1}	Nf_{6-2}	Nf_{6-3}	...	Nf_{6-n}	Nf_{6-1}	Ne_{6-2}	Ne_{6-3}	Ne_{6-4}	Ne_{6-5}	Ne_{6-6}
Line 7	Nf_{7-1}	Nf_{7-2}	Nf_{7-3}	...	Nf_{7-n}	Ne_{7-1}	Ne_{7-2}	Ne_{7-3}	Ne_{7-4}	Ne_{7-5}	Ne_{7-6}

Table 2

	Column 1	Column 2	Column 3	Column 4	Column 5
Line 1	Nf_{1-1}	Nf_{1-2}	Nf_{1-3}	Nf_{1-4}	Nf_{1-5}
Line 2	Nf_{2-1}	Nf_{2-2}	Nf_{2-3}	Nf_{2-4}	Ne_{2-5}
Line 3	Nf_{3-1}	Nf_{3-2}	Nf_{3-3}	Ne_{3-4}	Ne_{3-5}
Line 4	Nf_{4-1}	Nf_{4-2}	Ne_{4-3}	Ne_{4-4}	Ne_{4-5}
Line 5	Nf_{5-1}	Ne_{5-2}	Ne_{5-3}	Ne_{5-4}	Ne_{5-5}

Table 3

	Column 1	Column 2	Column 3	Column 4	Column 5
Line 1	Nf_{1-1}	Nf_{1-2}	Nf_{1-3}	Nf_{1-4}	Nf_{1-5}
Line 2	Nf_{2-1}	Nf_{2-2}	Nf_{2-3}	Nf_{2-4}	Nf_{2-5}
Line 3	Nf_{3-1}	Nf_{3-2}	Nf_{3-3}	Ne_{3-4}	Ne_{3-5}
Line 4	Nf_{4-1}	Nf_{4-2}	Ne_{4-3}	Ne_{4-4}	Ne_{4-5}
Line 5	Nf_{5-1}	Ne_{5-2}	Ne_{5-3}	Ne_{5-4}	Ne_{5-5}

Table 4

	Column 1	Column 2	Column 3	Column 4	Column 5
Line 1	Nf_{1-1}	Nf_{1-2}	Nf_{1-3}	Nf_{1-4}	Nf_{1-5}
Line 2	Nf_{2-1}	Nf_{2-2}	Nf_{2-3}	Nf_{2-4}	Nf_{2-5}
Line 3	Nf_{3-1}	Nf_{3-2}	Nf_{3-3}	Nf_{3-4}	Nf_{3-5}
Line 4	Nf_{4-1}	Nf_{4-2}	Ne_{4-3}	Ne_{4-4}	Ne_{4-5}
Line 5	Nf_{5-1}	Ne_{5-2}	Ne_{5-3}	Ne_{5-4}	Ne_{5-5}

Figure 4 Schematic diagram of interpolation ideas

As shown in the figure above, one is trapezoidal data. As shown in Table 1 of Figure 4, complete data and missing data in the dataset are arranged in trapezoidal order according to their completeness. On the left side of the figure is the distribution of complete data. The right side of the figure shows the distribution of missing data after clustering according to the number of missing data. Nf_{1-n} (the light-colored part) represents the attributes of the complete number in the dataset. Nf (N: Numerals full) represents the complete number in the dataset. The subscript number "1-n" represents the position of the n column in the first row; Ne_{2-1} in the figure (in blue) represents the

attributes of missing numbers. Where Ne (Ne: Numerals Empty) represents missing numbers; The subscript number "2-1" indicates the location in the data set and is located in row 2 and column 1 of the data set. As shown in Table 1, row 2 is missing to the right (column 6) is missing one data ($Ne_{2\text{-}6}$); The third line is missing the two right-most data ($Ne_{3\text{-}5}$ and $Ne_{3\text{-}6}$); And so on until line 7 is the row with all missing data.

The second is to build models. Based on the trapezoidal data in the first step, as shown in Table 2, a complete training model was established with the complete data $Nf_{1\text{-}1}$, $Nf_{1\text{-}2}$, $Nf_{1\text{-}3}$, $Nf_{1\text{-}4}$, and $Nf_{1\text{-}5}$ in the first row of Table 1.

The third is model-based training. As shown in Table 3, the complete data in the first left row of Table 2 were used to establish a training model for the complete data in the second row ($Nf_{2\text{-}1}$, $Nf_{2\text{-}2}$, $Nf_{2\text{-}3}$, and $Nf_{2\text{-}4}$).

Fourth, interpolation based on model training. Based on the training in step 3, interpolate $Ne_{2\text{-}5}$ in Table 3 in FIG. 4 and change $Ne_{2\text{-}5}$ into $Ne_{2\text{-}5}$, as shown in Table 4. After the interpolation of the second row in Table 3 is completed, the data of the second row is taken as the complete data of the first step, and the analogy continues until all the Spaces of the fifth row are interpolated. The third row of Table 4 also becomes a complete row data after interpolation. Just because the Spaces on the left side of Figure. 4 are trapezoidal after sorting, and the interpolation method is one by one from top to bottom until all Spaces are interpolated with data, this method is called the "pyramid interpolation method" in this study.

In summary, the multi-layer Perceptron integration model (MLP-ESM) based on the pyramid interpolation method makes full use of non-missing information and interpolates the columns with more missing values with fewer missing values in this study. In the process of coding, the model not only extracts the data features needed for interpolation and missing data but also extracts the structural relationship features between the predicted target and the existing data, which can achieve the effect of interpolation and prediction simultaneously.

3 Data

As for the absence of the CCR dataset used in this study, this section will specifically introduce the CCR dataset used in this study from data collection and data preprocessing. The details are as follows.

3.1 Data Collection

In this study, Compustat, Bloomberg, Google, and Facebook databases were used to form the missing dataset for the CCR predicti**on. The dataset s**pan from January 1, 2009, to December 31, 2019. In addition to corporate credit risk data and its historical

value, we also used data variables such as target company financial data, supply chain data, and non-financial data. We combine financial and non-financial data to include as much information as possible that affects our prediction results. In this study, the CCR dataset contained a total of 167,160 pieces of data from 441 S&P rating firms, with a data loss rate of 50%.

3.2 Data preprocessing

To analyze the value of all the data in this study, 118 original variables obtained were imported into the database. See Table 1 in the appendix. The table is divided into four columns, which are the serial number, name, number, and rate of missing data, respectively. A total of 167,160 databases have been built. It can be seen from Table 1 of the appendix that data loss has certain regularity. It can be summarized as follows :(1) the feature index with the least missing is Google trend (search number), with only 784 missing items, with a missing rate of 0.47%; The most missing data included 29 characteristic indexes, such as customer Wikipedia (clicks), customer risk and customer money, and 83,580 missing data, with the missing rate up to 50%. (2) Most of the characteristic indexes with low miss rate belong to the financial data reflecting the target company, such as return on assets, adjustment, total assets, capital turnover, and debt-equity ratio; However, the missing rate of customer-related data is high, such as customer-working capital ratio, customer-debt-equity ratio, customer-capital turnover, customer-fund debt ratio, and other characteristic indexes, the missing rate is as high as 50%. The missing rate is generally in the middle level of vendor-related indicators, such as vendor-leverage ratio, vendor-total assets, and vendor-cash, and the missing data rate is generally around 20%.

For the constructed database, this study uses Python 3.0 platform for missing value distribution presentation. The absence degree of the top 30 feature indicators in this study is shown in Figure 5. The more white lines in the figure, the more missing values; The darker the color, the fewer missing values. It can also be found from Figure 5 that the degree of data missing from different data sources is different. However, the distribution of missing data among homologous feature indicators is similar. For example, most of the features on the right of Figure 5 belong to customer-related feature indicators.

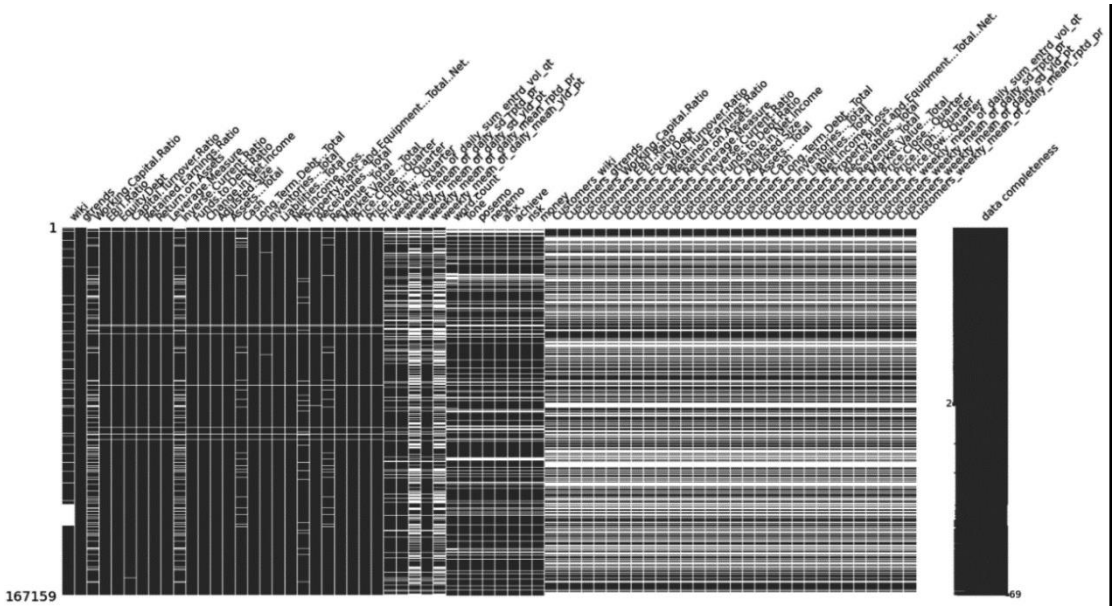

Figure 5 Top 30 indicators in terms of missing degree

In order to further show the distribution of missing feature indexes of the dataset. Bar of Python 3.0 analysis platform is used in this section to visually display the incompleteness of the dataset based on a single feature index. Figure 6 and 7 show the integrity of feature indexes whose data volume is in the top 30 and 30-60, respectively. Figure 6 and 7 are bar charts. The left bar (ranging from 0.0 to 1.0) represents the degree of data retention (the highest retention rate is 1=100%). The upper bar represents the actual number of data items.

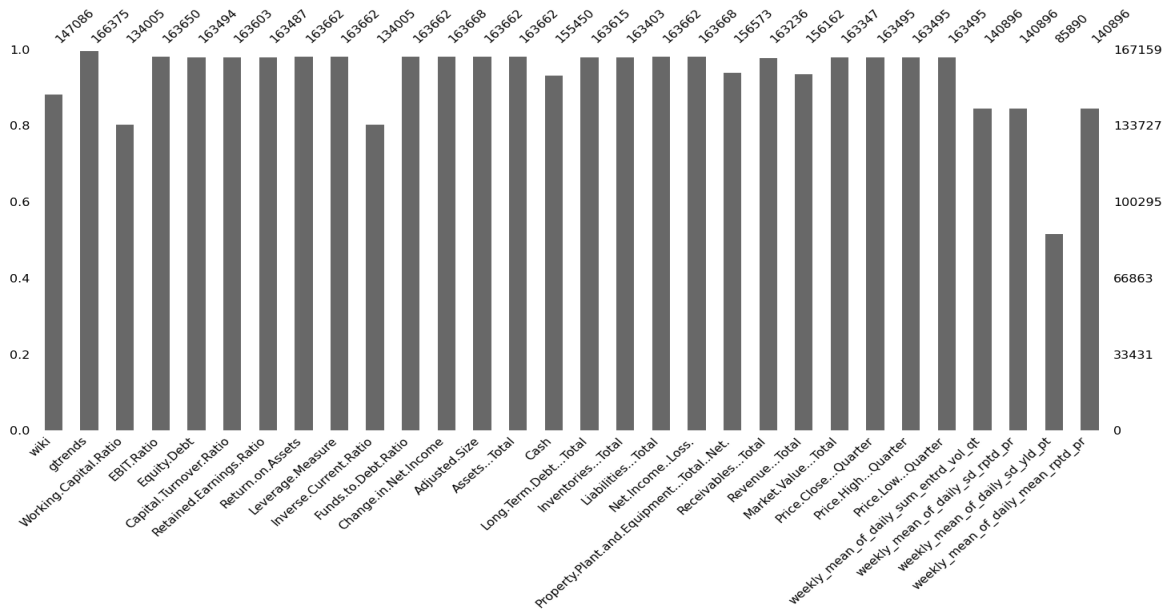

Figure 6 Top 30 indicators in terms of data volume

As you can see from Figure 6, the biggest misses are the standard deviation of the weekly average daily return of bonds (second bar from right), followed by the working capital ratio (third bar from left) and inverse cash ratio (10th bar from left).In addition, most of the top 30 feature indexes with data volume are concentrated in the target company's own financial data and network data, and most of the feature indexes are relatively complete. It can be seen from the data collection process in Chapter 3 of this study that this is mainly caused by the good degree of information recording and matching. It can also be seen from the distribution in Figure 7 that most of the feature indexes with data volume in the top 30-60 belong to the same source. It can be seen that, in the process of data restoration, the missing values in the dataset can be effectively interpolated by the correlation degree between the feature indexes in the dataset.

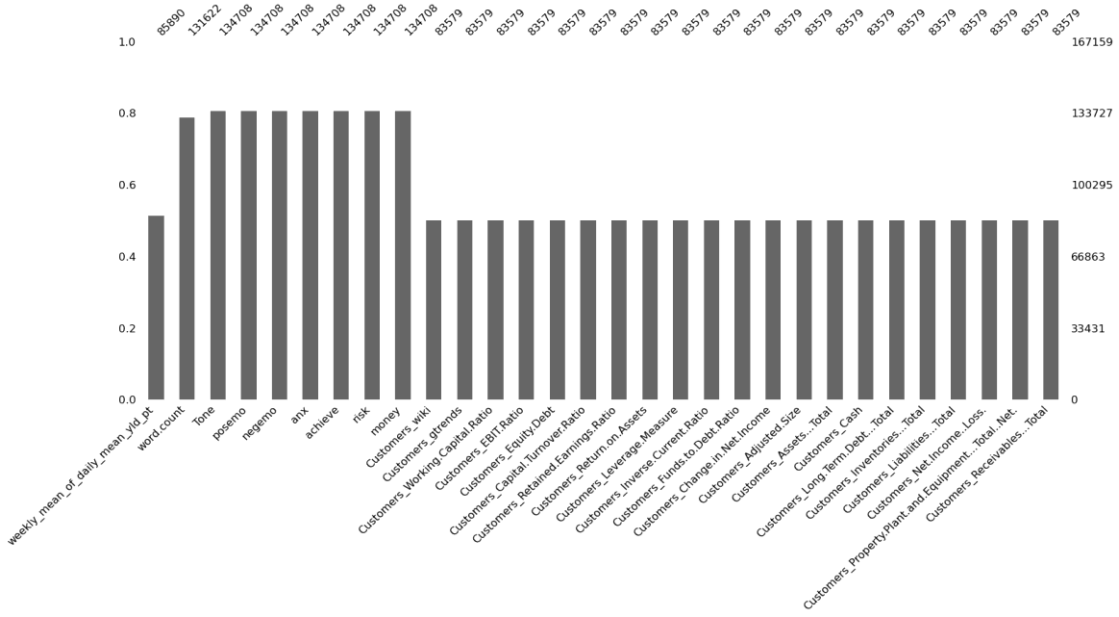

Figure 7 Top 30-60 indicators in terms of data volume

In order to explore the correlation between feature indexes of the dataset used in this study and to interpolate other missing data based on their correlation, 30 feature indexes were randomly selected from the dataset for correlation analysis. The drawn thermal map is shown in Figure 8.

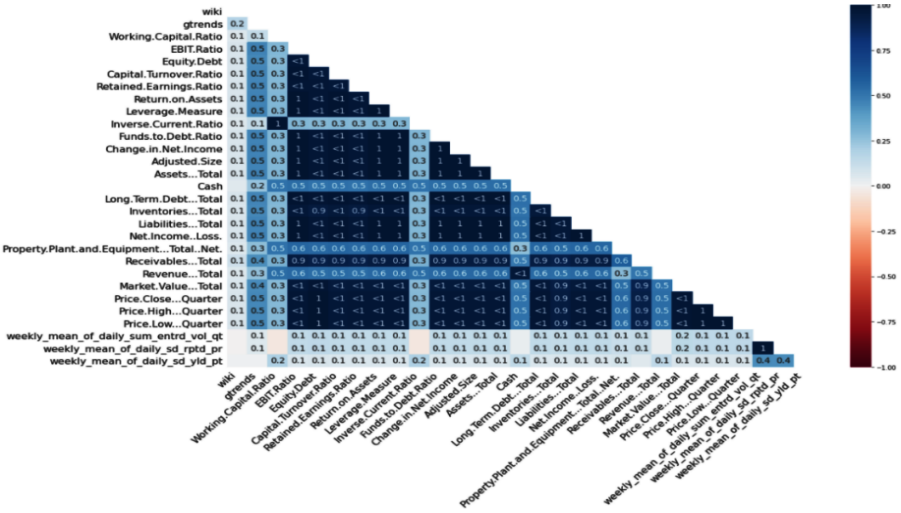

Figure 8 Ladder heat map based on 30 indicators

Figure 8 is a trapezoidal ladder diagram. The square "red on the top and bottom of the basket" on the right is the scale of thermal strength. The bluer the color up is, the stronger the positive correlation is. The redder the red down, the stronger the negative correlation. As can be seen from the blue depth of the left trapezoidal thermal ladder in Figure 8, the positive correlation intensity of the 30 randomly selected characteristic indexes is obvious. Figure 8 is a macroscopic demonstration of the correlation of feature indicators in the dataset. Therefore, this study also conducted hierarchical clustering to observe whether indicators with low miss rates in the dataset could be substituted with indicators with high miss rates to further explore the homology of feature indicators in the dataset of this study. The hierarchical clustering results are shown in Figure 9.

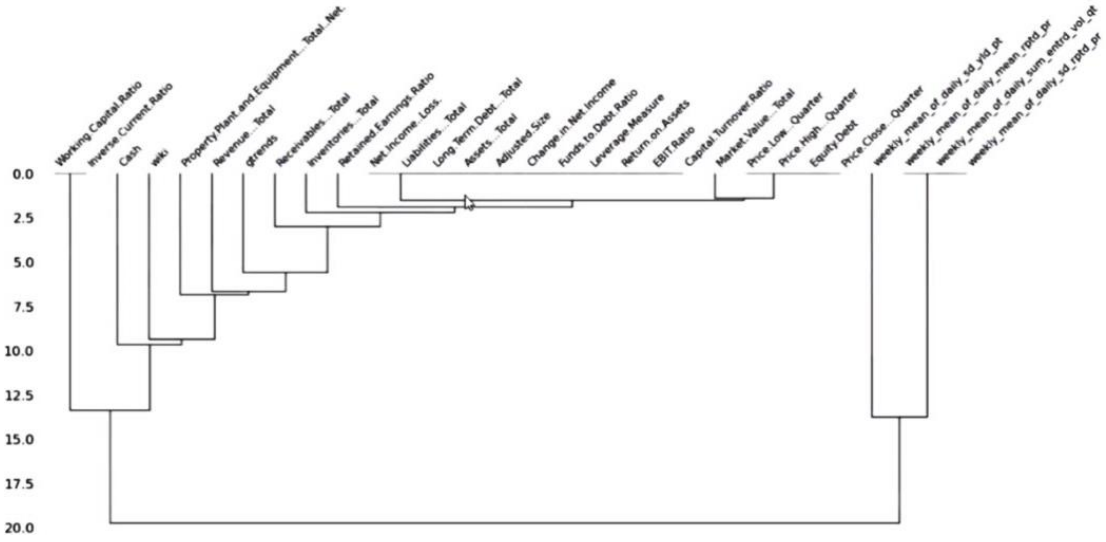

Figure 9 Hierarchical clustering plot

From the hierarchical clustering in Figure 9, it is possible to cluster data from the same source or different sources. For example, the standard deviation of the weekly average daily return of bonds, the total trading volume of bonds every Sunday, and the quoted standard deviation of the weekly total daily trading volume of bonds are directly clustered into one category, belonging to the same characteristic index. On the left, Wikipedia and Google Trends, which belong to different source metrics, also cluster into one category after two layers of clustering. Therefore, cross-column prediction is of practical significance when missing data is predicted.

4 Result

Based on the MLP-ESM model constructed in the previous section, this section verifies the model in the actual data set, and compares and analyzes the effects of this research model with other traditional machine learning prediction models to show the prediction results of the model presented in this study. The specific process is as follows.

4.1 Prediction effect of the model

This study is based on the MLP-ESM model constructed in Section 2. The missing data collected and processed in Section 3 is used as research data and analyzed on Python 3.0 analysis platform. The study allocates training samples and test samples in a ratio of 80% to 20%. The prediction performance of the model was evaluated by receiver operating characteristic curve (ROC) and area under curve (AUC). For the ROC curve, the closer the curve is to the upper left corner, the more accurate the model is. AUC is a performance index to measure the merits of the learner. The higher the AUC value, the higher the accuracy of the classifier. Through analysis, the AUC of the model in this study is 83.11%, and ROC is shown in Figure 10.

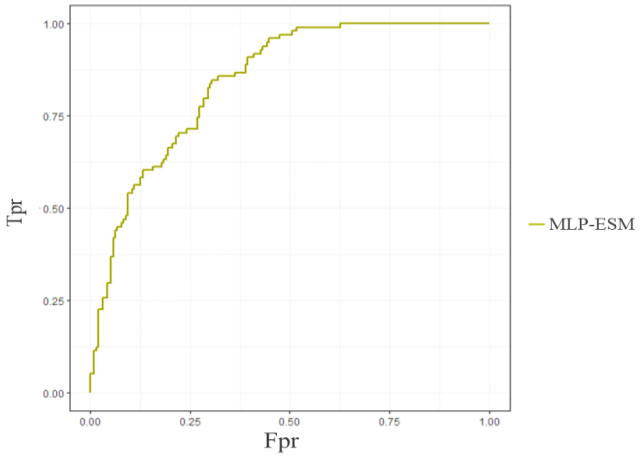

Figure 10 ROC curve of prediction with MLP-ESM model

As shown in Figure 10, the prediction accuracy of this research model reaches over 80%, but in the CCR prediction, it is worth noting what level 80% belongs to. Therefore, the prediction results of the model are discussed in this study with reference to previous literature. When evaluating ROC curves, Yang et al. (2017) pointed out that ROC curves and AUC judgment criteria are different in different application fields, such as face recognition, character recognition, and other areas, and AUC standards of 90% are acceptable. However, in the field of psychology (Ni et al., 2020), where the prediction requirement is relatively low, the AUC is acceptable at 70%. The prediction of CCR discussed in this study belongs to the prediction research in the field of economics. There are many factors influencing the prediction of CCR, the characteristics of datasets are different, and there is an unavoidable delay in the timeliness of data collection. Therefore, the acceptable accuracy of AUC value in the field of economics is lower than that in the field of psychology. Generally speaking, the classification of prediction grades based on AUC values in the field of economics is consistent despite certain controversies (Yang & Berdine, 2017). On Yang and Berdine (2017) criteria, 80% is an excellent prediction. And in this study, the prediction result of the MLP-ESM model proposed exceeds 80%, which can achieve a superb prediction effect.

4.2 Comparison of prediction effects

In order to further verify the prediction ability of the MLP-ESM model constructed in this study, four popular machine learning models, including support vector machine (SVM) model, XGboost (XGB) model, random forest (RF) model, and neural network (NN) model, are adopted in this section to compare their prediction effects. ROC curve and AUC values of the predicted results are shown in Figure 11 and Table5.

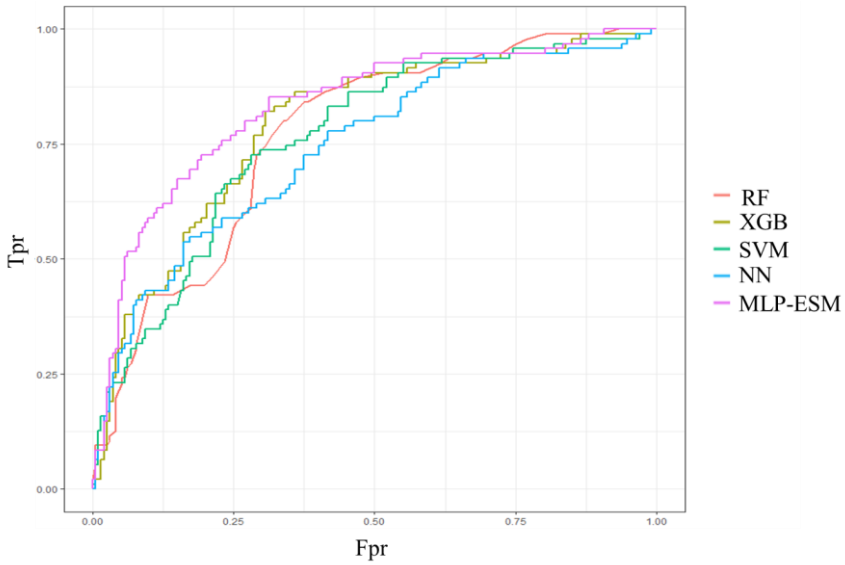

Figure 11 ROC plot based on five models

Table 5 AUC based on five models

Prediction method	MLP-ESM	Support vector machine	XGboost	Random forests	Neural network
AUC	83.11%	76.40%	79.01%	76.89%	74.31%

As can be seen from Table 5, the AUC values of the support vector machine, XGboost, random forest, neural network, and support vector machine model are all around 70% and do not exceed 80%. The prediction accuracy of the MLP-ESM model constructed in this study is 83.11%. This proves the validity of the model proposed in this study and shows that data restoration can improve the prediction results of the CCR dataset.

In conclusion, through comparative analysis, it is found that the corporate credit risk prediction model constructed in this study with missing data can achieve an excellent prediction level. At the same time, the interpolation method proposed in this study can repair the missing dataset well, and the interpolated dataset can achieve a better prediction effect compared with the prediction of non-missing data.

5 Conclusion

In order to advance the research on CCR prediction in the absence of data, this study proposes a MLP-ESM prediction model that can perform data interpolation and prediction simultaneously. The model makes full use of non-missing information and interpolates more missing columns with fewer missing values. In the process of self-coding, the hidden layer of the model contains variable information related to both the missing value and the predicted target, and the model can output the missing

interpolation value of the independent variable and the predicted target at the same time. In this way, not only the data features needed for missing data interpolation are extracted, but also the structural relationship features between the predicted target and the existing data are extracted, which can achieve the effect of simultaneous interpolation and prediction.

The results show that the MLP-ESM model proposed in this study can realize the simultaneous interpolation and prediction of missing data of CCR, and the AUC of the model is 83.11%. By comparison, the prediction effect of the MLP-ESM model constructed based on missing data in this study is better than that of the traditional machine learning model, which fully indicates that the dataset after interpolation can achieve a better prediction effect. This means that data restoration has application value to improve the accuracy of CCR prediction further and provides a new research idea for CCR prediction under the absence of data.

The current work is the basis for many future research directions. First of all, if there are too many values of continuous features in the model operation process, it can lead to sparse problems and low efficiency, which is worth noting. Secondly, how and to what extent data loss affects model performance still needs to be defined and quantified, which can provide more precise guidance for the model selection. These issues can be addressed further in the near future.

The appendix

Appended Table 1 List of missing values

The serial number	Feature Names	The missing number	Loss rate
1	wiki	20073	12.01%
2	gtrends	784	0.47%
3	Working.Capital.Ratio	33154	19.83%
4	EBIT.Ratio	3509	2.10%
5	Equity.Debt	3665	2.19%
6	Capital.Turnover.Ratio	3556	2.13%
7	Retained.Earnings.Ratio	3672	2.20%
8	Return.on.Assets	3497	2.09%
9	Leverage.Measure	3497	2.09%
10	Inverse.Current.Ratio	33154	19.83%
11	Funds.to.Debt.Ratio	3497	2.09%
12	Change.in.Net.Income	3491	2.09%
13	Adjusted. Size	3497	2.09%
14	Assets...Total	3497	2.09%
15	Cash	11709	7.00%
16	Long.Term.Debt...Total	3544	2.12%
17	Inventories...Total	3756	2.25%
18	Liabilities...Total	3497	2.09%
19	Net.Income..Loss.	3491	2.09%
20	Property.Plant.and.Equipment...Total..Net.	10586	6.33%
21	Receivables...Total	3923	2.35%
22	Revenue...Total	10997	6.58%
23	Market.Value...Total	3812	2.28%
24	Price.Close...Quarter	3664	2.19%
25	Price.High...Quarter	3664	2.19%
26	Price.Low...Quarter	3664	2.19%
27	weekly_mean_of_daily_sum_entrd_vol_qt	26263	15.71%
28	weekly_mean_of_daily_sd_rptd_pr	26263	15.71%
29	weekly_mean_of_daily_sd_yld_pt	81269	48.62%
30	weekly_mean_of_daily_mean_rptd_pr	26263	15.71%
31	weekly_mean_of_daily_mean_yld_pt	81269	48.62%
32	word.count	35537	21.26%
33	Tone	32451	19.41%
34	posemo	32451	19.41%
35	negemo	32451	19.41%
36	anx	32451	19.41%
37	achieve	32451	19.41%
38	risk	32451	19.41%
39	money	32451	19.41%

The serial number	Feature Names	The missing number	Loss rate
40	Customers_wiki	83580	50.00%
41	Customers_gtrends	83580	50.00%
42	Customers_Working.Capital.Ratio	83580	50.00%
43	Customers_EBIT.Ratio	83580	50.00%
44	Customers_Equity.Debt	83580	50.00%
45	Customers_Capital.Turnover.Ratio	83580	50.00%
46	Customers_Retained.Earnings.Ratio	83580	50.00%
47	Customers_Return.on.Assets	83580	50.00%
48	Customers_Leverage.Measure	83580	50.00%
49	Customers_Inverse.Current.Ratio	83580	50.00%
50	Customers_Funds.to.Debt.Ratio	83580	50.00%
51	Customers_Change.in.Net.Income	83580	50.00%
52	Customers_Adjusted.Size	83580	50.00%
53	Customers_Assets...Total	83580	50.00%
54	Customers_Cash	83580	50.00%
55	Customers_Long.Term.Debt...Total	83580	50.00%
56	Customers_Inventories...Total	83580	50.00%
57	Customers_Liabilities...Total	83580	50.00%
58	Customers_Net.Income..Loss.	83580	50.00%
59	Customers_Property.Plant.and.Equipment...Total..Net.	83580	50.00%
60	Customers_Receivables...Total	83580	50.00%
61	Customers_Revenue...Total	83580	50.00%
62	Customers_Market.Value...Total	83580	50.00%
63	Customers_Price.Close...Quarter	83580	50.00%
64	Customers_Price.High...Quarter	83580	50.00%
65	Customers_Price.Low...Quarter	83580	50.00%
66	Customers_weekly_mean_of_daily_sum_entrd_vol_qt	83580	50.00%
67	Customers_weekly_mean_of_daily_sd_rptd_pr	83580	50.00%
68	Customers_weekly_mean_of_daily_sd_yld_pt	83580	50.00%
69	Customers_weekly_mean_of_daily_mean_rptd_pr	83580	50.00%
70	Customers_weekly_mean_of_daily_mean_yld_pt	83580	50.00%
71	Customers_word.count	83580	50.00%
72	Customers_Tone	83580	50.00%
73	Customers_posemo	83580	50.00%
74	Customers_negemo	83580	50.00%
75	Customers_anx	83580	50.00%
76	Customers_achieve	83580	50.00%
77	Customers_risk	83580	50.00%
78	Customers_money	83580	50.00%
79	Suppliers_wiki	37191	22.25%
80	Suppliers_gtrends	37191	22.25%
81	Suppliers_Working.Capital.Ratio	37191	22.25%

The serial number	Feature Names	The missing number	Loss rate
82	Suppliers_EBIT.Ratio	37191	22.25%
83	Suppliers_Equity.Debt	37191	22.25%
84	Suppliers_Capital.Turnover.Ratio	37191	22.25%
85	Suppliers_Retained.Earnings.Ratio	37191	22.25%
86	Suppliers_Return.on.Assets	37191	22.25%
87	Suppliers_Leverage.Measure	37191	22.25%
88	Suppliers_Inverse.Current.Ratio	37191	22.25%
89	Suppliers_Funds.to.Debt.Ratio	37191	22.25%
90	Suppliers_Change.in.Net.Income	37191	22.25%
91	Suppliers_Adjusted.Size	37191	22.25%
92	Suppliers_Assets...Total	37191	22.25%
93	Suppliers_Cash	37191	22.25%
94	Suppliers_Long.Term.Debt...Total	37191	22.25%
95	Suppliers_Inventories...Total	37191	22.25%
96	Suppliers_Liabilities...Total	37191	22.25%
97	Suppliers_Net.Income..Loss.	37191	22.25%
98	Suppliers_Property.Plant.and.Equipment...Total..Net.	37191	22.25%
99	Suppliers_Receivables...Total	37191	22.25%
100	Suppliers_Revenue...Total	37191	22.25%
101	Suppliers_Market.Value...Total	37191	22.25%
102	Suppliers_Price.Close...Quarter	37191	22.25%
103	Suppliers_Price.High...Quarter	37191	22.25%
104	Suppliers_Price.Low...Quarter	37191	22.25%
105	Suppliers_weekly_mean_of_daily_sum_entrd_vol_qt	37191	22.25%
106	Suppliers_weekly_mean_of_daily_sd_rptd_pr	37191	22.25%
107	Suppliers_weekly_mean_of_daily_sd_yld_pt	37191	22.25%
108	Suppliers_weekly_mean_of_daily_mean_rptd_pr	37191	22.25%
109	Suppliers_weekly_mean_of_daily_mean_yld_pt	37191	22.25%
110	Suppliers_word.count	37191	22.25%
111	Suppliers_Tone	37191	22.25%
112	Suppliers_posemo	37191	22.25%
113	Suppliers_negemo	37191	22.25%
114	Suppliers_anx	37191	22.25%
115	Suppliers_achieve	37191	22.25%
116	Suppliers_risk	37191	22.25%
117	Suppliers_money	37191	22.25%
118	pred.weekly_mean_of_daily_mean_yld_pt	81450	48.73%

References

Akin Arikan, C., & Soysal, S. (2018). Investigation of Reliability Coefficients According to Missing Data Imputation Methods. *Hacettepe Universitesi Egitim Fakultesi Dergisi-Hacettepe University Journal of Education, 33*(2), 316-336.

Belue, L. M., Bauer, K. W., et al. (1997). Selecting optimal experiments for multiple output multilayer perceptrons. *Neural Computation, 9*(1), 161-183.

Budagaga, A. R. (2020). Determinants of banks' dividend payment decisions: evidence from MENA countries. *International Journal of Islamic and Middle Eastern Finance and Management, 13*(5), 847-871.

Campanile, L., Iacono, M., et al. (2020). *Towards the use of generative adversarial neural networks to attack online resources.* Paper presented at the Workshops of the International Conference on Advanced Information Networking and Applications.

Catellier, D. J., Hannan, P. J., et al. (2005). Imputation of missing data when measuring physical activity by accelerometry. *Medicine and Science in Sports and Exercise, 37*(11), S555-S562.

Chang, C., Deng, Y., et al. (2020a). Multiple imputation for analysis of incomplete data in distributed health data networks. *Nature communications, 11*(1), 1-11.

Chang, D., Yang, W., et al. (2020b). Seismic data interpolation using dual-domain conditional generative adversarial networks. *IEEE Geoscience and Remote Sensing Letters, 18*(10), 1856-1860.

Chiang, F., & Sitaramachandran, S. (2016). Unifying Data and Constraint Repairs. *Acm Journal of Data and Information Quality, 7*(3).

D'Angelo, G., & Palmieri, F. (2021). A stacked autoencoder-based convolutional and recurrent deep neural network for detecting cyberattacks in interconnected power control systems. *International Journal of Intelligent Systems, 36*(12), 7080-7102.

D'Angelo, G., Ficco, M., et al. (2021). Association rule-based malware classification using common subsequences of API calls. *Applied Soft Computing, 105*, 107234.

D'Angelo, G., Tipaldi, M., et al. (2019). A data-driven approximate dynamic programming approach based on association rule learning: Spacecraft autonomy as a case study. *Information Sciences, 504*, 501-519.

De Silva, H., Perera, A. S., et al. (2016, Sep 01-03). *Missing Data Imputation using Evolutionary k-Nearest Neighbor Algorithm for Gene Expression Data.* Paper presented at the 16th International Conference on Advances in ICT for Emerging Regions (ICTer), Negombo, SRI LANKA.

Eduardo, S., Nazabal, A., et al. (2020). *Robust Variational Autoencoders for Outlier Detection and Repair of Mixed-Type Data*. Paper presented at the Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics.

Florez-Lopez, R. (2010). Effects of missing data in credit risk scoring. A comparative analysis of methods to achieve robustness in the absence of sufficient data. *Journal of the Operational Research Society, 61*(3), 486-501.

Gao, D. Q., Yang, Z. P., et al. (2012). Performance evaluation of multilayer perceptrons for discriminating and quantifying multiple kinds of odors with an electronic nose. *Neural Networks, 33*, 204-215.

Garcia, V., Marques, 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.

Gu, B., Li, Z., et al. (2017). Web-ADARE: A web-aided data repairing system. *Neurocomputing, 253*, 201-214.

Guo, X., Jarrow, R. A., et al. (2009). Credit risk models with incomplete information. *Mathematics of Operations Research, 34*(2), 320-332.

Hooke, M., Mrozinski, J., et al. (2021, Mar 06-13). *Salvaging Data Records with Missing Data: Data Imputation using the Multivariate t Distribution.* Paper presented at the IEEE Aerospace Conference (AeroConf), Electr Network.

Jerez, J. M., Molina, I., et al. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine, 50*(2), 105-115.

Khoygani, M. R. R., & Ghasemi, R. (2016). Neural estimation using a stable discrete-time MLP observer for a class of discrete-time uncertain MIMO nonlinear systems. *Nonlinear Dynamics, 84*(4), 2517-2533.

Lan, Q. J., & Jiang, S. (2021). A method of credit evaluation modeling based on block-wise missing data. *Applied Intelligence, 51*(10), 6859-6880.

Li, W., Ding, S., et al. (2020). Heterogeneous ensemble learning with feature engineering for default prediction in peer-to-peer lending in China. *World Wide Web-Internet and Web Information Systems, 23*(1), 23-45.

Mählmann, T. (2006). Estimation of rating class transition probabilities with incomplete data. *Journal of Banking & Finance, 30*(11), 3235-3256.

Moon, T., Hong, S., et al. (2019). Interpolation of greenhouse environment data using multilayer perceptron. *Computers and Electronics in Agriculture, 166*, 105023.

Moscato, V., Picariello, A., et al. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications, 165*.

Nakagawa, S., & Freckleton, R. P. (2008). Missing inaction: the dangers of ignoring missing data. *Trends in Ecology & Evolution, 23*(11), 592-596.

Ni, C., & Jin, X. (2020). Could L2 Lexical Attrition Be Predicted in the Dimension of Valence, Arousal, and Dominance? *Frontiers in Psychology, 11*, 3464.

Nijman, S. W. J., Groenhof, T. K. J., et al. (2021). Real-time imputation of missing predictor values improved the application of prediction models in daily practice. *Journal of Clinical Epidemiology, 134*, 22-34.

Qiu, Z., Meng, M. R., et al. (2008, Jul 25-27). *Missing Value Treatment of the Data Mining Based on Bayesian Principle.* Paper presented at the 3rd International Conference on Computer Science and Education, Kaifeng, PEOPLES R CHINA.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*(3), 1-21.

Shema, A., & Assoc Comp, M. (2019, Jan 04-07). *Effective Credit Scoring Using Limited Mobile Phone Data.* Paper presented at the 10th International Conference on Information and Communication Technologies and Development (ICTD), Indian Inst Management, Ahmedabad, INDIA.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*(1), 1-48.

Sivasankar, E., Selvi, C., et al. (2016, Dec 10-11). *A Study of Dimensionality Reduction Techniques with Machine Learning Methods for Credit Risk Prediction.* Paper presented at the 3rd International Conference on Computational Intelligence in Data Mining (ICCIDM), Bhubaneswar, INDIA.

Soldatyuk, N., & Sopko, S. (2014, Sep 10-12). *Methods of solving missing data issues in credit risk*

scoring and comparison of its effectiveness. Paper presented at the 32nd International Conference on Mathematical Methods in Economics (MME), Olomouc, CZECH REPUBLIC.

Twala, B. (2013). Impact of noise on credit risk prediction: Does data quality really matter? *Intelligent Data Analysis, 17*(6), 1115-1134.

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.

Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Yang, S., & Berdine, G. (2017). The receiver operating characteristic (ROC) curve. *The Southwest Respiratory and Critical Care Chronicles, 5*(19), 34-36.

Yap, B. W., Ong, S. H., et al. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications, 38*(10), 13274-13283.

Yu, B. J., Li, C. M., et al. (2022). Forecasting credit ratings of decarbonized firms: Comparative assessment of machine learning models. *Technological Forecasting and Social Change, 174*.

Yuan, K. P., Chi, G. T., et al. (2022). A novel two-stage hybrid default prediction model with k-means clustering and support vector domain description. *Research in International Business and Finance, 59*.

Yue, Y. M., Tian, J. W., et al. (2016, Sep 21-23). *Applications of block chain technology in credit rating.* Paper presented at the 13th International Conference on Industrial Management (ICIM 2016), Hiroshima, JAPAN.

Zhang, S. L., Wang, P., et al. (2016). Missing value data processing based on statistical correlation. *Statistics and Decision, (12) : 13-16*.