AUTOENCODER BASED GENERATOR FOR CREDIT INFORMATION RECOVERY OF RURAL BANKS

Gujun Yan[1, *]

[1]Institute of Finance
Zhejiang University
Hangzhou, China
[*]Corresponding author's e-mail: yjts@zju.edu.cn

By using machine learning algorithms, banks and other lending institutions can construct intelligent risk control models for loan businesses, which helps to overcome the disadvantages of traditional evaluation methods, such as low efficiency and excessive reliance on the subjective judgment of auditors. However, in the practical evaluation process, it is inevitable to encounter data with missing credit characteristics. Therefore, filling in the missing characteristics is crucial for the training process of those machine learning algorithms, especially when applied to rural banks with little credit data. In this work, we proposed an autoencoder-based algorithm that can use the correlation between data to restore the missing data items in the features. Also, we selected several open-source datasets (German Credit Data, Give Me Some Credit on the Kaggle platform, etc.) as the training and test dataset to verify the algorithm. The comparison results show that our model outperforms the others, although the performance of the autoencoder-based feature restorer decreases significantly when the feature missing ratio exceeds 70%.

Keywords: Autoencoder; Missing Feature; Restored Features; Dot Product Operation; Banks.

(Received on November 3, 2022; Accepted on March 20, 2023)

1. INTRODUCTION

With the development of the financial industry, credit risk management has become more complex. To cope with these challenges, the realization of credit evaluation through machine learning algorithms has become a rapidly developing direction (Turkson *et al.,* 2016; Munkhdalai *et al.,* 2019; Khandani *et al.,* 2010). In the capital market, privacy violations remain a problem. Given the sensitivity of their information, capital markets are a top victim of cybercrime. Risk and safety reduction have been top priorities due to the rise in digital interactions and communications. However, for small-scale banks such as township banks and rural banks, due to the small number of credit transactions, a single business point has a small amount of data. However, due to the differences in the economic development and target customers of the regions where the banks are located, there are significant differences in the characteristics of users, and the data of bank outlets in different regions cannot be directly merged (Mandala *et al.,* 2012). At the same time, due to incomplete information collection, user data often contain many missing values. Therefore, to retain data for model training, it is particularly important to retain and restore data with missing features. Recently, with the emergence of deep learning, autoencoder has been widely researched. Autoencoder is a type of neural network used for unsupervised learning. The purpose of an autoencoder is to learn a compressed representation of the input data, which can then be used for tasks such as data compression, dimensionality reduction, and feature extraction.

A lender's risk-adjusted rate of profitability is optimized through the financial sector, which keeps credit risk sensitivity within reasonable bounds. Lenders must control both the overall portfolio's underlying creditworthiness and the risk associated with specific loans or operations. The probability of failure, as well as loss intensity in the case of default, is the two main elements of credit risk. The average outcome is the result of the two factors. One of the important features of modern credit risk management is just about the accurate measurement of credit risk, which is not only a means to effectively identify risks, but also a prerequisite for the use of a series of risk control methods. With the advent of the era of big data, more and more banks have realized the significance of comprehensive measurement of credit risk based on data, and the evaluation system has evolved from the past one-dimensional system to the multi-dimensional system that considers the probability of default, loss given default and so on.

Colleges, as well as other borrowing organizations, may create sophisticated contingency planning systems for their loan companies by employing machine learning techniques. However, the training process must be completed with the necessary attributes filled in, particularly whenever applicable to financial institutions with limited credit information.

DOI: 10.23055/ijietap.2023.30.2.8697 ISSN 1943-670X

Thus, in order to recover the erroneous collected data in the characteristics by taking advantage of the connection among the information, we proposed an autoencoder-based approach. This method uses the correlation between data to restore the missing data items in the features so that the credit risk management prediction can have more prediction data. Also, an anomaly feature classifier is discussed to avoid incorrect permutations. To validate the method, we also used several open-domain databases as the learning and validation datasets. The comparative findings demonstrate that our approach surpasses the competition, although whenever the featured lacking percentage approaches 70%, the effectiveness of the autoencoder-based featured restorative drastically degrades.

The organization of the paper is as follows: Section 1 shows the introduction; Section 2 depicts the background and related works; Section 3 illustrates methods; Section 4 describes experimental evaluation; finally concludes Section 5.

2. BACKGROUND AND RELATED WORKS

In recent years, machine learning and deep learning-based methods have received extensive attention in the field of economics and finance (Saraswathi *et al.*, 2022; Delgoshaei *et al.*, 2021; Jiang *et al.*, 2022). Among all the active researches, using machine learning algorithms to predict the repayment ability of customers is one of the most important applications. Most financial firms are obligated to establish a rational as well as perpetuity decision regarding whether you are capable of repaying the loan under the ability-to-repay criterion. Typically, the law requires banks to learn about, take into account, and record a borrower's earnings, property, profession, payment history, and expenditure. The weekly spending power (or excess, depending on how much total salary fewer average expenses is), as well as other criteria, including the joint income, possessions, obligations, and consistency of revenue, are used to determine the capability to return the loan. Based on the borrower's information, lending institutions can not only build an intelligent risk control model and expand the business boundary for the bank but also bring convenience to borrowers. However, existing research is mainly focused on the evaluation and prediction of the risk level and creditworthiness of the borrowers. Malekipirbazari and Aksakalli (2015) proposed a random forest (RF) based classification method for predicting borrower status and found RF-based method outperforms the FICO credit scores in the identification of good borrowers. The ultimate forecast is made by the randomized forest classification using proportional representation. The brand has been selected as the forecast by the overwhelming of the selection forests. Tao (2020) introduced an improved random forest model to predict individual credit defaults, trying to solve the problem that the poor classification effect of some decision trees may affect the prediction effect of the entire random forest model. Boughaci and Alkhawaldeh (2020) evaluated eleven techniques to distinguish between bad and good applicants on seven datasets to measure their performance and concluded that Bayes Net, Random Forest, AdaBoost, and LogitBoost machine learning classifiers produce efficient models for credit scoring. Creditors could more correctly determine a borrower's hazard with the use of AI. This may be accomplished by looking at information that isn't considered in typical creditworthiness, such as if the lender uses their money on needs or pleasures. Luo *et al.* (2017) compared the performance of the deep belief network (DBN) model with that of the traditional model in the classification of corporate credit scores by using the credit default swap data of the 2007-2008 U.S. financial crisis and found that the deep belief network performs best (Xu *et al.*, 2022). Yu *et al.* (2018) proposed a deep belief network-based resampling support vector machine (SVM) ensemble learning paradigm and believe the paradigm can be used to deal with imbalanced data in credit risk classification. Ma *et al.* (2018) used the new machine learning algorithms to predict the default risk and found the LightGBM algorithm-based result is the best.

Chen *et al.* (2019) proposed a deep learning framework that combines neural networks and GBDT for credit assessment, and they believe that DeepGBM deep learning framework achieves good results in the classification learning task of the credit assessment. The factors that affect credit availability in banks are a high ratio of credit use, lack of a credit mix, remaining debt, careless payment practices, and so on. Carta *et al.* (2020) introduced a method that calculates the entropy from the input features and does further classification tasks. In the paper, they believe this method can classify a new instance without the knowledge of past non-reliable instances. Alasbahi and Zheng (2022) used a transfer learning-based method that enables using missing feature values to facilitate the learning of credit scores and believe their proposed method solves the feature irregularities, class imbalance, and concept drift issues in binary classification problems. Zhang *et al.* (2022) proposed an iForest model that combines Isolation Forest and EasyEnsemble to detect fake data in credit evaluation work, and they believe their results are significantly better than the vanilla model. Lan and Jiang (2021) introduced a credit evaluation model based on a multi-task feature selection approach. This method divides the dataset into several nonoverlapping subsets based on missing patterns and integrates the multi-task feature selection approach using logistic regression to perform joint feature learning on all subsets. Comparable to single-task image segmentation, the multi-task classification algorithm seeks to choose a common sample of characteristics that are crucial for all associated activities. In the proposed method, they trained several sub-models and chose the optimal one from them. They claim that the framework can effectively process block-wise missing data for credit evaluation.

However, in the actual evaluation of bank loans, it is inevitable to encounter data with some missing credit characteristics. At the same time, those machine learning methods require data with complete features for model training and further prediction. Abandoning such data may cause overfitting for those institutions with little credit data and violates the principle of maximizing the use of existing data (Wang *et al.,* 2022; Wu *et al.,* 2021). Therefore, commonly used models often adopt methods such as filling median and mean values to complete the missing features and then carry out the next step of prediction work. But those ways of supplementing the missing features did not fully utilize the correlations between features, which affects the fairness of the prediction results.

With the development of deep-learning methods, there are also several works on imputation using the autoencoder-based method. Yu *et al.* (2021) proposed a two-stage network for missing value imputation that uses an autoencoder to estimate the missing value and a multi-layer perceptron to refine the estimation. A source, outcome, and one or more convolutional units with several cells layered together—make up a deep network. The synapses in a Multilayer Perceptron can employ any random input layer, in contrast to transistors in a Perceptron, which require an input vector that enforces a cutoff, such as ReLU or exponential. A densely integrated kind of feedforward artificial neural network is called a multilayered deep neural (MLP). The word MLP is utilized unclearly; it can apply to every convolutional ANN in some contexts or specifically to systems comprised of multiple levels of activation functions in others. In the refining stage of this work, it tries to decrease the difference between the original value and the predicted value.

Aidos and Tomás (2021) used the k-nearest neighbors algorithm for initial missing data estimation, and then they put the initial value into the autoencoder network to recover the missing value. Several works integrate the missing data imputation and later tasks. Lai *et al.* (2020) used an autoencoder to fill in the missing data. Then, they do regression and classification based on the recovery result. Zhu *et al.* (2021) used an encoder structure for latent feature extraction. After that, it uses the latent feature to recover the original input and further regression work. Those works are based on an autoencoder for missing value imputation. However, for the network that was trained on a little dataset, the training dataset might just cover a very little portion of the whole data. Thus, it might cause some failed predictions when the input is different from the training distribution. It will be detrimental to those machine learning methods that are sensitive to false input.

Currently, there are few studies on the recovery of missing features of bank customers' credit information for small dataset problems. Because of the above phenomenon, it is very meaningful to develop appropriate data analysis technology to accurately realize the supplement and recovery of missing features, to effectively help banking institutions to lend fairly and efficiently to individuals and enterprises.

3. METHODS

Due to the serious problem of missing features in the credit data of township banks, at the same time, due to the relatively small amount of data itself, the data with missing features cannot be directly discarded. At the same time, since the input features have a certain redundancy, that is, there is a certain correlation between the features, the existing features can be used to complete the missing feature items. Therefore, this paper proposes an autoencoder-based missing feature completion network.

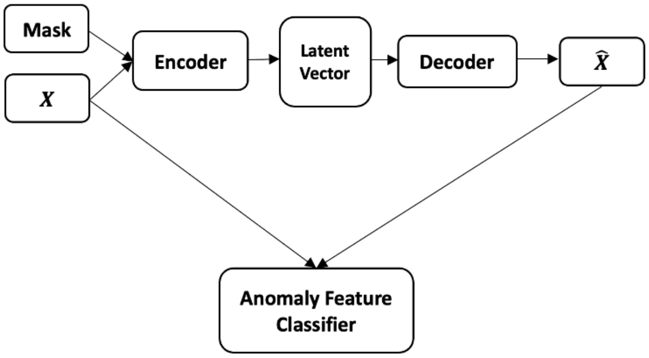


Figure 1. Structure of the Whole Model

As shown in Figure 1, during the training, the ground truth input feature is multiplied by a mask vector which will produce the feature with missing items. Then, our proposed Encoder encodes this incomplete feature into a latent vector that represents the characteristics of the input feature. Based on this latent vector, the proposed Decoder tries to predict an output feature that is the same as the ground truth input feature. During the prediction process, the parameters of the autoencoder are fixed, the missing features of the abnormal data are filled with 0, and the corresponding mask is generated and then input into the model. Since the model performs data completion based on the redundant relationship between the features, when the features are severely missing, the obtained features may not be recovered correctly. Since the model is trained based on a small amount of data which might not cover all the distribution of the features, when the input data are not aligned with the training data, the autoencoder will not be robust enough to get the correct output.

For that failed imputation, since the input data are not well-learned, this paper assumes that the recovered output features should be quite different from the original features in all the channels. Based on that assumption, this paper proposed an anomaly feature classifier in the recovery feature verification stage which is shown in Figure 1.

3.1 Autoencoder-based Missing Data Imputation

Autoencoder is a popular unsupervised neural network model whose overall structure consists of two parts: encoder and decoder. Clusters are created throughout the establishment of ANNs via unsupervised learning by combining model parameters of the same kind. The neural network responds with a proportional gain after applying a unique input sequence, identifying the category to which the training data belongs. Unsupervised learning's primary goal is to recognize latent as well as intriguing connections in large datasets. Unsupervised learning techniques, in contrast to reinforcement methods, are unable to solve a prediction or classifying issue immediately since it is unknown what the extracted features will be. Image compression, knowledge discovery, fault diagnosis, computer graphics, therapeutics, attractiveness forecasting, language processing, and object tracking are just a few examples. The encoder is the procedure used in computing to convert a collection of data (alphabets, numerals, punctuation, and specific signs) into a form that is specifically designed for effective data transfer. The method of converting an encrypted form back into the underlying string of symbols is described as a decoder. In the encoder, it obtains its corresponding latent features by learning the input features, as shown in equation 1; in the decoder part, the model uses the learned latent features to reconstruct the original input data, as equation 2 shows (Wang *et al.*, 2016). Latent characteristics are obscured traits to set them apart from seen aspects at the cost of over-implications. Matrix factorization is used to calculate latent characteristics from visual attributes. Analyzing text documents is one instance. Characteristics are "items" taken directly from the texts. The auto-encoder can use the redundancy between the input features to encode the input features through the multi-layer perceptron to obtain the corresponding hidden features and input the decoder to restore the original input features according to the hidden features as Equation 3.

$$\phi\colon \mathrm{X} \to F \tag{1}$$

$$\varphi\colon F \to \hat{X} \tag{2}$$

$$\varphi, \phi = \underset{\varphi,\phi}{\arg\min}\left\|X - \hat{X}\right\|^2 = \underset{\varphi,\phi}{\arg\min}\|X - (\varphi \circ \phi)X\|^2 \tag{3}$$

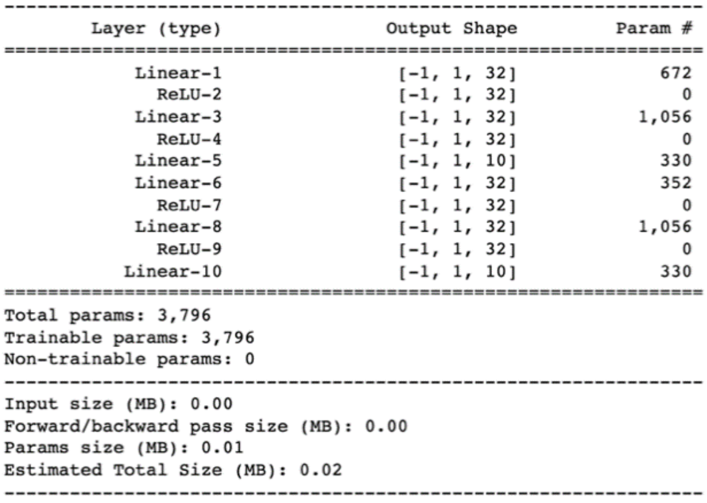
X is the input vector and $\hat{X}$ is the output vector, they both belong to feature space $X, \hat{X} \in R^n$. F is the latent vector that belongs to space $F \in R^k$ which refers to the underlying characteristics or patterns within the input data that are learned by the model. These features are not directly observable in the input data but are inferred by the model through the process of reconstruction.

The encoder is composed of a multi-layer fully connected network, which realizes the mapping from the original feature space to the latent feature space. The MLP works by applying a series of non-linear transformations to the input features through a sequence of hidden layers. Each hidden layer applies a linear transformation to the output of the previous layer, followed by a non-linear activation function that introduces non-linearity into the model. The final layer of the MLP produces the encoded representation of the input features, which is the latent feature that is needed. In contrast, the decoder performs a mapping from the latent feature space to the original feature space, as equations 4 and 5 show. Through the training of the autoencoder, the original features can be mapped to the corresponding hidden features, and then the original input features can be recovered from the hidden features, and the features restored by the encoded hidden features can be as close to the original input features as possible (Lopez-Martin *et al.*, 2017). Methods for minimizing the number of source parameters in the learning algorithm are referred to as wavelet transform. It is sometimes beneficial to decrease the dimensions whenever working with highly large datasets by expressing the information to a lower-level domain that retains the core of the information. Among them, the dimension of the hidden feature is smaller than the original feature space, so through the mapping of the encoder, the redundant information existing in the input feature can be removed, the valid information in it can be retained, and the original feature with missing information can be recovered.

$$Hidden_{vec} = f(W \cdot X) \tag{4}$$

$$X^* = f(W \cdot Hidden_{vec}) \tag{5}$$

In this model, due to the lack of features in the input features, the auto-encoder is first used to restore the missing features. Different from ordinary autoencoders, to solve the problem of missing features, this paper improves the autoencoders. The basic structure of the model is shown in Figure 2 and Figure 3.



```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Linear-1                [-1, 1, 32]             672
              ReLU-2                [-1, 1, 32]               0
            Linear-3                [-1, 1, 32]           1,056
              ReLU-4                [-1, 1, 32]               0
            Linear-5                [-1, 1, 10]             330
            Linear-6                [-1, 1, 32]             352
              ReLU-7                [-1, 1, 32]               0
            Linear-8                [-1, 1, 32]           1,056
              ReLU-9                [-1, 1, 32]               0
           Linear-10                [-1, 1, 10]             330
================================================================
Total params: 3,796
Trainable params: 3,796
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 0.00
Forward/backward pass size (MB): 0.00
Params size (MB): 0.01
Estimated Total Size (MB): 0.02
----------------------------------------------------------------
```

Figure 2. Parameters of the Network

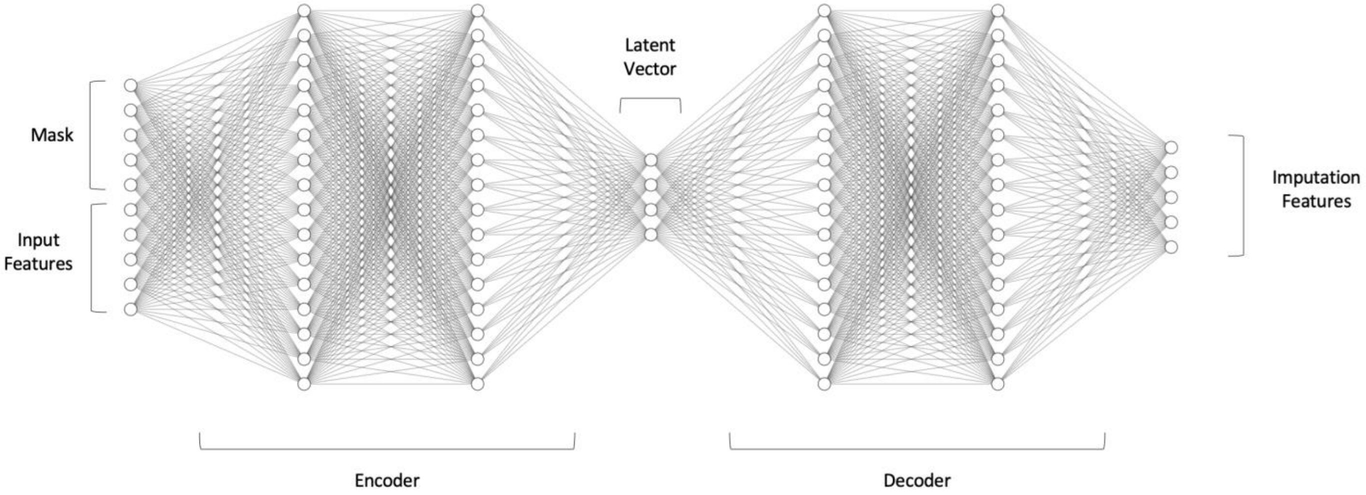


Figure 3. Structure of Autoencoder Network

During the training process, for the input feature data X, a vector mask is randomly generated to represent the missing value and the corresponding position in equation 6, where 0 means that there is a missing feature at the position and 1 means that there is no missing feature. In Equation 7, product the original features with the mask to obtain the input features with missing features.

$$mask = random(length(X)) \tag{6}$$

$$\hat{X} = mask \cdot X \tag{7}$$

The missing features and masks are combined to obtain a new input, and the hidden features are mapped and restored by the autoencoder. The difference between the original features and the output features is used as the loss function of the network for training. Whenever learning neural network modeling, the two primary gradient descent categories to employ are cross-entropy as well as mean squared error. Thus, an autoencoder that can perform data completion based on missing data is obtained.

3.2 Anomaly Feature Classifier

For those rural banks, the above autoencoder network usually has to be trained on a relatively small dataset. As a result, the training data is likely to cover only a limited portion of the overall distribution. Thus, when the distribution of the test data is different from the distribution of the train data, the autoencoder may be unrobust and make fake predictions. Identifying data trends that do not match appropriate results is described as neural networks. In many uses, these defected tendencies are frequently alluded to as abnormalities, deviations, conflicting findings, exclusions, excesses, unexpected oddities, or contamination. At the same time, that failed prediction will have a significant influence on the later prediction or regression task. Since an anomaly is not expressly modeled in the systems, the detection system is not a supervised classifier. Rather, kids are taught to identify exactly what is usual. In reality, if we had a large number of irregularities of various types to deal with, we might apply ternary categorization.

At the same time, when the input feature is from a distribution that is unaligned with the training distribution, the Autoencoder has a large chance fail to recover the feature. In those cases, both the encoder network and decoder network need to deal with unseen data and will not be likely to work well. Therefore, the prediction will not be well-recovered on all channels. In order to detect those failed cases, this method decides if the imputation is correct by classifying the existing feature channels. Thus, this framework used a multi-layer perceptron-based method for fake prediction classification.

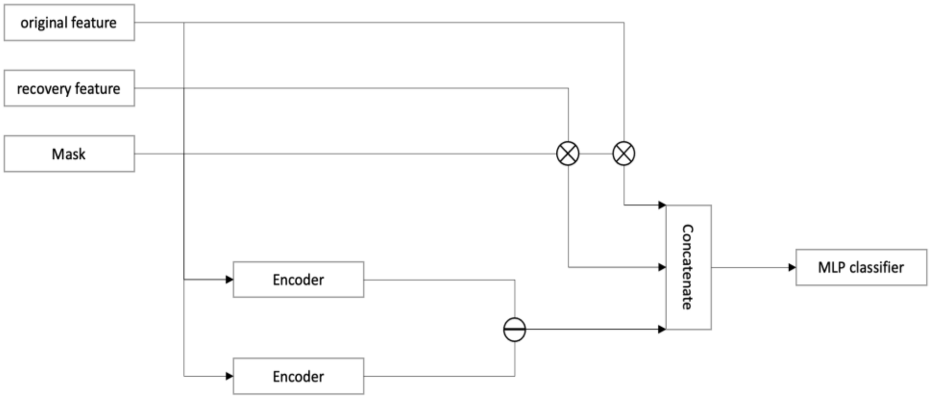


Figure 4. Structure of Anomaly Feature Classifier

As shown in Figure 4, the classifier does the dot product operation on the input feature and output feature with the Mask, respectively, to get the remaining features after removing the missing features. Then, it concatenates the original feature with the mask and feeds it to the encoder obtained from the autoencoder to get the encoded latent vector. At the same time, it does the same operation as the recovery features. By doing the minus operation between those two latent vectors, we can get the difference between the hidden features. Two SELECT expressions are utilized in conjunction with the SQL Minus Operator. The outcome collection acquired using the initial selected statement is subtracted from the dataset received by the subsequent SELECT statement using the MINUS operation. After that, the classifier concatenates the differential latent vector with the output of the above dot product result. Finally, we use the concatenate features as the input of the multi-layer perceptron to get the binary classification output. By calculating the distance between the input feature and the output feature as the similarity between the two, it is judged whether the autoencoder can correctly restore the input feature. If the output of the classification is true, the input features are considered to be normal data; otherwise, the input features are considered abnormal data.

During the training process, the proposed classifier has to generate the training label first. In the current stage, we threshold the Mahalanobis distance to decide if the output features should be regarded as the incorrect imputation. The multidimensional extension of measuring the distance between a point as well as the multidimensional distribution's average is recognized as the Mahalanobis distance. The Mahalanobis distance, as opposed to the Euclidean distance, takes into consideration the degree of correlation between the factors. For instance, you may have observed a strong correlation between fuel economy and capacity. As a result, the Euclidean distance computation contains a lot of duplicate data.

4. EXPERIMENTAL EVALUATION

Since the bank's credit data involves the private information of customers, in the process of verifying the algorithm, we use several open-source datasets to train and test the algorithm.

4.1 Datasets

German Credit Data is a public dataset from the University of California Irvine. The dataset contains personal credit loan information, which can be divided into basic personal information and loan information. The dataset comprises 1000 samples and 20 features (7 continuous and 13 categorical). The label of the dataset comes with two values {1,0}, where 1 (positive class) indicates that a customer has good credit, and 0 (negative class) indicates that a customer has poor credit (Zhang *et al.*, 2022). For those categorical values, we use one-hot encoding to represent their value. Since the dataset is relatively small, we choose 800 samples as training data and 200 samples as testing data. The function of one-hot encoding is given below:

$$f(j) = \begin{bmatrix} a_1^j \\ \vdots \\ a_N^j \end{bmatrix} \tag{8}$$

$$a_i^j = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases} \tag{9}$$

Home credit default risk. This dataset is from Kaggle.com; the original target of this dataset is aimed to predict their clients' repayment abilities. To verify the result of our proposed algorithm, we use the training data from Home credit default risk. The training data contains 307511 samples with 104 features. The training data have a label column which also with binary values indicating the credit risk of a certain customer. We randomly choose 5000 samples as training data and 200 samples as testing data.

Give Me Some Credit Data. This dataset is public and available from Kaggle.com; the initial purpose of this dataset is for banks to determine whether or not a loan should be granted. The data set contains 10 features used to evaluate customer credit conditions. After removing the items containing abnormal data, there are a total of 120,269 pieces of data. For verification purposes, we only use the training data from this dataset. This dataset comes with the probability that somebody will experience financial distress in the next two years. To simulate the small dataset case of the rural bank, we randomly choose 5000 samples as training data and 200 samples as testing data.

4.2 Evaluation Metric

During the experiment, to compare the performance of the model, it is necessary to remove some features manually. In the implementation process, a mask with the same dimension as the feature is generated with a certain probability, and the missing feature data is generated by adding a mask on top of the original features. The original features are used as the label for the prediction. The variance's computations are identical to those for the mean squared error. Taking the standard deviation, consider taking the anticipated value out, then square that disparity to get the MSE. That should be done for each report. Divide the total sum of these squared integers by the set of measurements. An indicator of how closely a device monitor is to datasets is the Mean Squared Error (MSE). Then square the quantity for each dataset by multiplying it by the vertical distance between the item as well as the matching absolute values on the appropriate amounts.

The mean square error (MSE) is used to judge, and its calculation formula is Equation 10.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{X_\iota})^2, \tag{10}$$

where the X_i indicates the ground truth features, while $\widehat{X_\iota}$ represents the reconstructed features. By measuring MSE, we are summing the difference between the original input features and the features we recovered from all the samples. Since all features are normalized, the mean square error can be used to measure the distance between the restored features and the original features.

4.3 Results

In the experiment, we first preprocess the training dataset by eliminating the samples with abnormal values and converting the category value into a one-hot encoding value. Information can be converted using one-hot encoding as a means of getting an improved forecast and preparing the information for an algorithm. With one-hot, we create a unique category column for every classified item as well as give it a binary integer of 1 or 0. A binary vector is used to describe every decimal number. Since the previous research on the problem of missing features mainly focused on the method of fill-in in the missing data using mean or the median, we compared our method with them. The mean-based method calculates the arithmetic mean of the dataset by adding up all the values and dividing by the total number of values. The recovered feature is filled out using this mean. This method is sensitive to outliers, as a single extreme value can significantly affect the overall mean. In contrast, the median-based method calculates the median of the dataset by sorting the values and selecting the middle value. Meanwhile, to find out the impact of the missing ratio of the input features, we also compared the result under the different missing ratios. The comparison results are as Table 1 shows.

Table 1. Methods Comparison

German Credit Data		
Methods	Mask Ratio = 0.95	Mask Ratio = 0.80
AE-based recovery (ours)	0.051	0.152
Mean based recovery	0.110	0.193
Medium based recovery	0.146	0.231
Home credit default risk		
Methods	Mask Ratio = 0.95	Mask Ratio = 0.80
AE-based recovery (ours)	0.231	0.426
Mean based recovery	0.549	0.978
Medium based recovery	0.793	1.164
Give Me Some Credit Data		
Methods	Mask Ratio = 0.95	Mask Ratio = 0.80
AE-based recovery (ours)	0.032	0.092
Mean based recovery	0.080	0.163
Medium based recovery	0.134	0.212

Note: The table shows the mean square error of different methods. Mask ratio means the ratio of the remaining features, which equals (1 - missing ratio).

From the result, we can find out that our method is significantly better than mean-based and medium-based methods. An information set's mean (average) is calculated by summing all of the integers in the collection, then splitting by the maximum population of variables in the sequence. Whenever a collection of information is ranked from lowest to largest, the middle is the midpoint. However, for the home credit default risk dataset, the feature dimension is much larger than the rest dataset. Thus, the MSE is much bigger than the rest of the two datasets. Moreover, since the ability of the model to perform feature interpolation is affected by the degree of a feature missing, to evaluate the impact of a different feature missing conditions on the model's recovery ability, for the AE model, the experimental MSE of recovering features under different feature missing ratios is shown in Figure 5.

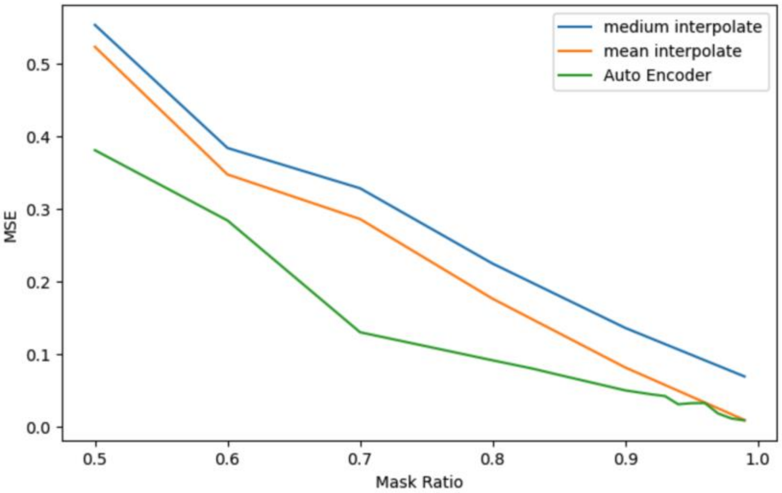


Figure 5. MSE under Different Mask Ratios for Give Me Some Credit Data

It can be found that when the feature missing ratio exceeds 70%, the performance of the Auto Encoder-based feature restorer decreases significantly. Due to the lack of many features and insufficient redundant information between features, the complete original features cannot be restored correctly.

5. CONCLUSION

Using machine learning algorithms to predict the repayment ability of customers can help lending institutions build intelligent risk control models and bring convenience to borrowers. But in small rural banks, there exist serious problems of missing features in the credit data and a relatively small amount of data itself. Because of this, this paper proposes an autoencoder-based algorithm that can use the correlation between data to restore the missing data items in the features and check the correctness of the imputation. Several open-source datasets are selected as the training and test dataset to verify the algorithm, and the comparison results show that our model is significantly better than mean-based and medium-based methods. Also, the proposed method achieves 70% efficiency by using MSE as compared with the traditional techniques. For future research, it is meaningful to address the fairness of the data to build a more robust model and make in-deep discussions about the practical applications in small banks.

REFERENCES

Aidos, H. and Tomás, P. (2021). Neighborhood-aware Autoencoder for Missing Value Imputation. *2020 28th European Signal Processing Conference (EUSIPCO)*, Amsterdam, Netherlands.

Alasbahi, R. and Zheng, X. (2022). An Online Transfer Learning Framework with Extreme Learning Machine for Automated Credit Scoring. *IEEE Access*, 10:46697-46716.

Boughaci, D. and Alkhawaldeh, A.A.K. (2020). Appropriate Machine Learning Techniques for Credit Scoring and Bankruptcy Prediction in Banking and Finance: A Comparative Study. *Risk and Decision Analysis*, 8(1): 15-24.

Carta, S., Ferreira, A., Recupero, D.R., Saia, M., and Saia, R. (2020). A Combined Entropy-Based Approach for a Proactive Credit Scoring. *Engineering Applications of Artificial Intelligence*, 87: 103292.

Chen, X., Liu, Z., Zhong, M., Liu, X., and Song, P. (2019). A deep learning approach using DeepGBM for credit assessment. *Proceedings of the 2019 International Conference on Robotics, Intelligent Control and Artificial Intelligence*, Shanghai, China.

Delgoshaei, A., Mohammad Azari, M., Hanjani, S. E., Fard, F., Beigizadeh, R., and Aram, A. K. (2021). A Fuzzy Logic-Based Algorithm for Supply Chain Management Considering Different Cases. *International Journal of Industrial Engineering: Theory, Applications and Practice*, 27(6): 5883.

Jiang, Y., Huang, Y., Liu, J., Li, D. P., Li, S. Y., Nie, W. J., and Chung, I.-H. (2022). Automatic Volume Calculation and Mapping of Construction and Demolition Debris Using Drones, Deep Learning, and GIS. *Drones*, 6(10): 279.

Khandani, A. E., Kim, A. J., and Lo, A. W. (2010). Consumer Credit-risk Models via Machine-learning Algorithms. *Journal of Banking & Finance*, 34(11): 2767-2787.

Lai, X. C., Wu, X., and Zhang, L.Y. (2020). Autoencoder-based Multi-task Learning for Imputation and Classification of Incomplete Data. *Applied Soft Computing*, 98: 106838.

Lan, Q. J. and Jiang, S. (2021). A Method of Credit Evaluation Modeling Based on Block-wise Missing Data. *Applied Intelligence*, 51(4): 1-22.

Lopez-Martin, M., Carro, B., Sanchez-Esguevillas, A., and Lloret, J. (2017). Conditional Variational Autoencoder for Prediction and Feature Recovery Applied to Intrusion Detection in IoT. *Sensors*, 17(9): 1967.

Luo, C. C., Wu, D. S., and Wu, D.X. (2017). A Deep Learning Approach for Credit Scoring Using Credit Default Swaps. *Engineering Applications of Artificial Intelligence*, 65: 465-470.

Ma, X. J., Sha, J. L., Wang, D. H., and Yu, Y.B. (2018). Study on a Prediction of P2P Network Loan Default Based on the Machine Learning LightGBM and XGboost Algorithms According to Different High Dimensional Data Cleaning. *Electronic Commerce Research, and Applications*, 31: 24-39.

Malekipirbazari, M. and Aksakalli, V. (2015). Risk Assessment in Social Lending via Random Forests. *Expert Systems with Applications*, 42(10): 4621-4631.

Mandala, I. G. N. N., Nawangpalupi, C. B., and Praktikto, F.R. (2012). Assessing Credit Risk: An Application of Data Mining in a Rural Bank. *Procedia Economics and Finance*, 4: 406-412.

Munkhdalai, L., Munkhdalai, T., Namsrai, O. E., Lee, J. Y., and Ryu, K. H. (2019). An Empirical Comparison of Machine-learning Methods on Bank Client Credit Assessments. *Sustainability*, 11(3): 699.

Saraswathi, S., Deepa, G., Vennila, G., Parthasarathy, S., and Ramadoss, B. (2022). A Survey on Big Data: Infrastructure, Analytics, Visualization and Applications. *International Journal of Industrial Engineering: Theory, Applications and Practice*, 29(5): 7643.

Tao, Y. L. (2020). Research on Prediction of Personal Credit Default by Improved Random Forest Model. Master's Thesis. Hebei University of Economics and Business.

Turkson, R. E., Baagyere, E. Y., and Wenya, G. E. (2016). A machine learning approach for predicting bank credit worthiness, *2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR)*, Lodz, Poland.

Wang, H., Gao, Q., Li, H., Wang, H., Yan, L. P., and Liu, G.H. (2022). A Structural Evolution-Based Anomaly Detection Method for Generalized Evolving Social Networks. *The Computer Journal*, 65(5): 1189-1199.

Wang, Y., Yao, H., and Zhao, S. (2016). Auto-encoder Based Dimensionality Reduction. *Neurocomputing*, 184: 232-242.

Wu, X., Zheng, W., Xia, X., and Lo, D. (2021). Data Quality Matters: A Case Study on Data Label Correctness for Security Bug Report Prediction. *IEEE Transactions on Software Engineering*, 48(7): 2541 - 2556.

Xu, L., Chen, W. J., Wang, S. L., Mohammed, B. S., and Kumar, R. (2022). Analysis on Risk Awareness Model and Economic Growth of Finance Industry. *Annals of Operations Research*. DOI: https: //doi.org/10.1007/s10479-021-04516-z.

Yu, J., He, Y., and Huang, J. Z. (2021). A two-stage missing value imputation method based on autoencoder neural network, *2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA.

Yu, L., Zhou, R. T., Tang, L., and Chen, R. D. (2018). A DBN-based Resampling SVM Ensemble Learning Paradigm for Credit Classification with Imbalanced Data. *Applied Soft Computing*, 69: 192-202.

Zhang, X. D., Yao, Y., Lv, C. D., and Wang, T. (2022). Anomaly Credit Data Detection Based on Enhanced Isolation Forest. *The International Journal of Advanced Manufacturing Technology*, 122: 185-192.

Zhu, H., Ren, Y., Tian, Y., and Hu, J. (2021). A Winner-Take-All Autoencoder Based Piecewise Linear Model for Nonlinear Regression with Missing Data. *IEEJ Transactions on Electrical and Electronic Engineering*, 16(12): 1618-1627.