JOSEIT JOURNAL OF SYSTEMS ENGINEERING AND INFORMATION TECHNOLOGY
http://jurnal.iaii.or.id/index.php/JOSEIT

Predicting ICO Prices Using Artificial Neural Network and Ridge Regression Algorithm

Trầhn Kim Toại[1*], Võ Thị Xuân Hạn[2], Võ Min Huân[3]
[123]Ho Chi Minh City University of Technical Education, Ho Chi Minh City, Vietnam
Email: [1*]toait@hcmute.edu.vn

Abstract

An Initial Coin Offering (ICO) is a method of raising funds for digital currency projects. Investors purchase these coins at a very low initial price before they are released. These coins are then listed on the trading platform, and their prices may increase rapidly if the currency performs well. After six months of release, ICO evaluation is the expected time for investors to profit. A dataset consisting of 109 ICOs was constructed from reputable websites after data preprocessing. Correlation analysis of 12 inputs revealed issues of multicollinearity, leading to biased regression model results. Overfitting occurred when using the regression model. To address these limitations, the Ridge regression method resolved the issues with the ICO data. An artificial neural network model addressed the complex nonlinear relationships between inputs and ICO prices. By adjusting parameters to achieve the best performance according to the Root Mean Square Error, R-squares, and Mean Absolute Error metrics, the results showed that the Ridge regression algorithm with a test set of three ICOs achieved accuracy ranging from 63% to 92% of ICO prices, while the artificial neural network model predicted with 98% accuracy depending on the metric used.

Keywords: ICO prediction; ridge regression; linear regression; ICOs; artificial neural network

1. Introduction

A company seeking to raise capital may issue its own digital currency by creating a certain amount of tokens, which act as a type of share, to sell in the market and attract investors during the initial offering[1]. The introduction of ICOs aims to address the issue of startups facing a lack of capital to implement their innovative business or technological ideas[2]. Therefore, they turn to investors for funding. However, ensuring profitability of an ICO is not absolute. An ICO may fail or succeed after its release. There are ICO projects that are not good and fraudulent that still exist in the community, or profitable investment projects that do not meet expectations[3]. Investors typically wait from 25 to 30 weeks to decide to exit the investment channel if the profit is not as expected[4].

Once a potential ICO is released, and the token is widely accepted, its price will increase exponentially and generate profits for investors compared to the purchase price at the time of issuance. At this point, investors will sell these tokens to make a profit, which is the gain[5]. Factors affecting the uncertainty of ICO success include the quality of the issuing team, information about the ICO, product ideas, media, social networks, and opinions of experts in the cryptocurrency field. The quality of the issuing team includes experienced and well-disciplined teams that will also produce good quality products[6].

Information about the ICO should include the start and end date of token sales, how to trade, price, total supply, market capitalization, etc. The product idea should present the company's idea through video presentations, technology, platforms, services used, and milestones that the company wants to achieve when launching the product. Media, social networks for a prominent ICO project are when it is mentioned a lot on social networks such as Facebook, Twitter[4].

The ICO price prediction algorithm is a tool that supports the interests of investors and advisors. Machine learning-based prediction algorithms bring high efficiency and accuracy, making them the trend of forecasting applications[7]. Therefore, researchers have proposed many machine learning algorithms applied in forecasting, such as the multiple linear regression model, Ridge regression model, artificial neural network, support vector machine, etc. Predicting the price of ICOs to be launched in the near future is crucial. When companies, investors, or advisors have specific predictions with high accuracy, they will take appropriate steps based on the predicted results[8].

Various machine learning methods have been applied to predict the success of ICO projects. Sentiment analysis has been regarded as a useful tool for evaluating the attractiveness of ICOs[9]. Based on user comments on Twitter, the authors constructed user sentiment data to evaluate the success of an ICO and the amount of successfully raised investment. Machine learning models such as logistic regression and random forest were used to predict the success of ICOs due to their high accuracy in analyzing user sentiment[10].

Accepted by editor: 30-03-2023 | Final revision: 31-03-2023 | Online publication: 31-03-2023

In addition, analysis based on ICO whitepapers has also been studied to predict the success of ICOs. A natural language processing model analyzed the language commonly used in successful ICO whitepapers to evaluate a different ICO. The study concluded that successful ICOs have complete whitepapers with terms considered as input variables of the model. In this study, we predict the price of ICOs six months after release to evaluate the success of ICOs, considering many factors that can affect the ICO price, which can result in profit or loss. After six months of investment, investors expect a return on investment that satisfies the profit rate[11].

Based on input data such as prediction model parameters, this study analyzes the correlation between inputs and outputs to evaluate the impact of parameters on the ICO price after six months. The dataset includes 12 fields, including the key factors that influence the price of an ICO. These factors include the US dollar price, bitcoin price, total supply, market capitalization, available supply, amount received, Ethereum price at the time of ICO opening, bitcoin price at the time of ICO opening, month of ICO opening, ICO opening date, country of ICO opening, and ICO release duration. Six months after the ICO release is the expected period for investors to evaluate the success of ICOs[12]. Two methods, ridge regression and artificial neural network, are proposed and compared for their predictive accuracy on the dataset constructed from reliable cryptocurrency ICO websites by simulating the process using Python programming language. The tables and graphs presented are results extracted from the simulation process using this Python programming tool.

The application of machine learning methods to predict ICO prices based on various factors affecting the ICO price has not been fully explored. Previous research on predicting the success of ICOs has mainly focused on analyzing user sentiment factors that influence the success of ICOs or relying on whitepaper factors to make predictions. However, ICO prices are also influenced by many other factors that have not been studied for their potential impact on the success of ICO prices six months after the initial offering[13], [14].

This paper describes the theoretical foundation for building a Ridge regression model based on linear regression and artificial neural network architectures in Section 2. Section 3 presents the data preprocessing and model building process. Section 4 shows the results of the analysis and evaluation of the model's ability to predict ICO prices six months after the offering. Section 5 provides conclusions and contributions to the research.

2. Method

2.1 Multiple Linear Regression Problem

Equation (1) shows the relationship of the multiple linear regression model. The equation represents the relationship between the output and input variables of a linear function:

$$\hat{y} = h\theta(x) = \theta 0 + \theta 1 x1 + \theta 2 x2 + \cdots + \theta n x n = \theta T X \ (1)$$

Where ŷ is the predicted value. The matrix $\theta T = [\theta 0\ \theta 1\ \cdots\ \theta n]$ describes the model parameters and is the transpose of the θ matrix. The variables $[\theta 0\ \theta 1\ \cdots\ \theta n]$ are the parameters of the regression model. The predicted value ŷ and the parameters represent a linear relationship. The X matrix is a matrix containing the variables of the model. $X = [x1\ x2\ \cdots\ xn]$. Linear regression has limitations, such as being very sensitive to noise. When the system encounters noise, it can cause a non-linear relationship between the input and output. This will affect the prediction results when using the linear regression method. Additionally, linear regression cannot represent complex models with many inputs. When the model becomes complex, it can lead to overfitting[15], [16].

2.2 Ridge Regression Problem

The Ridge regression differs from linear regression by adding a regularization term to the loss function. Essentially, Ridge regression optimizes two components simultaneously, including the sum of squared residuals and the regularization term. The equation for the loss function of Ridge regression is presented in equation (2).

$$J(\theta) = \frac{1}{2N}\left(\sum_{i=1}^{N} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{i=1}^{N} \theta_i^2\right) \qquad (2)$$

Here, $x\ i$ represents the i-th input feature of the model. N is the sample size. $y\ i$ represents the actual value. λ is called the complexity adjustment parameter of the model. This parameter is used to control the magnitude of the adjustment component affecting the loss function. Ridge regression applies this adjustment component control technique to help address the issues of multicollinearity and overfitting in data. In the case of a very large λ, almost all model parameters decrease to 0 and this is called underfitting.

When λ is very small, Ridge regression becomes a regular linear regression, leading to overfitting. In the case of a small λ, the role of the adjustment component becomes less important, and the control of the overfitting phenomenon becomes less effective. It is essential that the λ parameter is appropriately designed. We experimented with different values of λ during the training process. The best trained model is the one that minimizes overfitting, and the optimal λ value can be determined from formula (2) as shown in formula (3).

$$J(\theta) = \frac{1}{2N}(|\theta^T X - y|^2 + \lambda\theta^2) \tag{3}$$

Similar to the problem posed by the multiple regression problem, the model parameter J(θ) needs to be designed with a minimum value.

$$\theta = (X\ TX + \lambda I)\ -1\ X\ T \tag{4}$$

In which, the matrix X contains the input variables and the displacement matrix is $X\ T$. The unit matrix I is designed with suitable dimensions. It is easy to see that the Ridge regression solution involves adding a different amount of λI than the linear regression.

2.3 Algorithm of Artificial Neural Network (ANN)

Artificial Neural Networks (ANN) are a predictive technique based on a model that simulates the operation of the human brain. ANN represents the nonlinear relationship between input and output variables. The ANN algorithm consists of two stages: forward propagation and backpropagation. The forward propagation algorithm computes the sum of the product of inputs and weights. The result is passed through the hidden layer. The values in the hidden layer and the output layer are computed using the Tanh function. The backpropagation algorithm involves updating the weights using the Gradient Descent algorithm to reduce the loss. The weight updates are performed continuously until the model achieves an acceptable loss value[16].

Figure 1 illustrates the neural network architecture for the ICO price prediction system after six months. To determine the number of hidden layers and the number of nodes in each hidden layer, numerous experiments were conducted to identify the parameters that produce the most accurate results. The designed neural network model includes an input layer with 128 data fields, three hidden layers, each with 100 nodes, and one output neuron. The Tanh function is used as the activation function in the hidden layers[17].

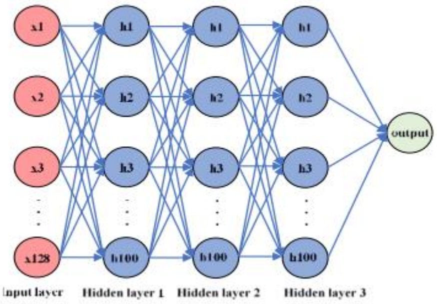


Figure 1. Neural network model for ICO price prediction

2.4 ICO Price Forecasting System Design

This ICO price prediction system consists of two components: an algorithm and a dataset. The dataset is constructed from multiple websites related to ICO prices. The algorithm employs the Ridge regression method to perform on this dataset. The algorithm's output is the price of an ICO after six months of its issuance.

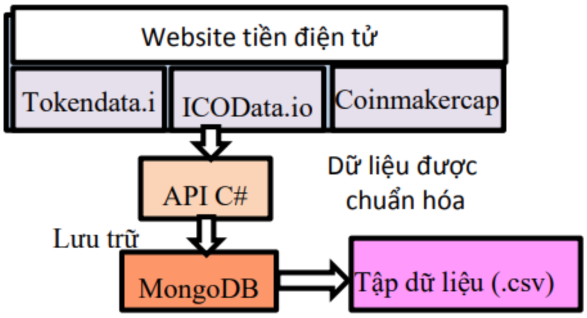


Figure 2. Flowchart of Data Collection Steps

The dataset was constructed from reliable Internet sources of ICO data in the field of cryptocurrency such as tokendata.io, icodata.io, and coinmartketcap.com using an API interface programmed in C# (Maria & Omar, 2020). Then, a data preprocessing step was performed to filter and process invalid or missing data. The .csv file format was used for storage. This data preparation step aims to clean and transform raw unformatted data into useful data for analysis. Invalid data is defined as data that contains start_date and end_date fields with invalid values. The complete dataset consists of 109 ICOs, including 109 rows and 12 columns.

Table 1. The Dataset Divided Into 3 Parts

Cross data		Test holdout data	Test data
Training data	**Validation data**	21 ICO	3 ICO
68 ICO	17 ICO		

To improve the accuracy of the machine learning algorithm's predictions, One Hot encoding is used to convert textual fields into binary data. One hot encoding is applied to classify the fields of data. However, this technique increases the number of input features. One hot encoding is applied to classify the last four fields, including ICO duration day, Date ICO was launched, Month ICO was launched, and the Country an ICO was launched from. These fields need to be converted from textual to numerical format. Therefore, the dataset will no longer have 12 columns, but up to 128 columns, while the number of rows remains at 109.

The dataset consists of 109 ICOs, which are divided into 106 ICOs for model training and 3 ICOs for testing (testData) and evaluating the proposed algorithm's accuracy. The 106 ICOs are further divided into two smaller sets: the cross-validation set with XCross and yCross coordinates and the testHoldout set with XTestHoldout and yTestHoldout coordinates. The dataset is split into 80% for the cross-validation set (85 ICOs) and 20% for the testHoldout set (21 ICOs). The cross-validation set (85 ICOs) is used to train the Ridge regression algorithm. From this cross-validation set, two subsets are extracted: the training set with XTrain and yTrain coordinates and the validation set with XVal and yVal coordinates.

The training set is used to train the Ridge regression model, while the validation set (20% of the cross-validation set) is used to evaluate the model and avoid overfitting. The best hyperparameters are determined based on the validation set's performance, as shown in Table 1. The trained model's accuracy is evaluated using the validation set (XVal, yVal) with performance metrics, including RMSE, MAE, and R2, to measure the prediction error between the predicted values and the actual values.

The cross dataset is used to train the algorithm and save the model in .sav file format. The training results will generate three models with three RMSE, MAE, and R2 values in the first iteration. The testHoldoutData set is used to monitor the accuracy by the iteration to adjust the parameters and find the optimal training model based on the performance evaluation metrics of the model. The model performs optimization based on a loop with a different sample set selected for each iteration. The result is different RMSE, MAE, and R2 values after each iteration. The performance metrics are evaluated and compared with each other after each iteration. The optimized model selects the best RMSE, MAE, and R2 values.

3. Result and Discussion

3.1 Data Collection Process

Based on the input-output relationship, this study analyzes the strong and weak correlation between these factors. A scatter plot with correlation coefficients between input and output is presented in Figure 4. Figure 4 describes the correlation between corresponding input and output, and between input variables themselves. Strong correlations are represented by large correlation coefficients and are bolded for emphasis. As described in Figure 4, the main diagonal represents the names of the inputs and output.

a. USD Price: represented in the input variable price_usd. This is the price of a token converted in USD. The correlation coefficient is 0.88 between the two variables price_usd and output. This correlation represents a strong positive relationship, indicating that if price_usd increases, the price of the ICO output also increases.
b. Bitcoin Price: represented by the input variable price_btc. This is the price of a token converted in bitcoin. The correlation coefficient is 0.88 between the two variables price_btc and output. This correlation is also strong between these two variables.
c. Total Supply: using the variable total_supply, which is the total number of tokens that can be supplied to the market. For example, the 0x token supplies the market with one billion tokens. Aeron token supplies the market with 20 million tokens. These two variables have almost no correlation with each other because the correlation coefficient is -0.05.
d. Market Capitalization: represented by the input variable market_cap_usd and output is 0.24. This correlation value means that the two variables have a weak correlation.
e. Available Supply: represents the number of tokens currently circulating. Figure 4 shows a correlation value of -0.05 between the input variable and the output variable, indicating that they have no correlation with each other.
f. Money Raised: is the total amount of money raised from the time the company issues the token until the end of the sale, measured in dollars. The correlation coefficient is 0.24 between the two variables usd_raised and output based on Figure 4. The relationship between these two variables is weak.
g. Ethereum Price at Launch: the price of Ethereum at the time of issuance does not have a significant impact on the output. The correlation coefficient is -0.366 between the two variables eth_price_launch and output. These two variables have a weak relationship.

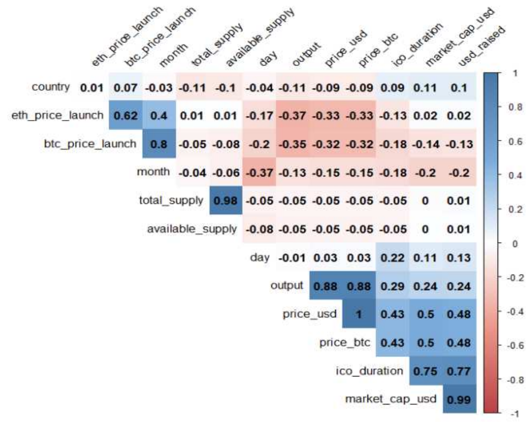


Figure 4. Analysis Chart of The Correlation Between Input Variables with Output and Between Inputs After Correlation Simulation

The price of Bitcoin at the time of the ICO launch has a minor influence on the output, similar to Ethereum. The correlation coefficient between the two variables, btc_price_launch and output, is -0.36565, indicating a weak relationship between the two. The month in which the ICO is launched, described by the variable "month," has little effect on the output. The correlation coefficient between "month" and "output" is -0.37, indicating a weak relationship between the two variables. The variable "day," which describes the specific day of the ICO launch, has a small impact on the output, with a correlation coefficient of 0.03 as presented in Figure 4. Similarly, the country in which the ICO is launched, described by the variable "country," has a weak correlation with the output, with a correlation coefficient of -0.11.

The duration of the ICO launch, which represents the number of days for ICO issuance, has almost no correlation with the market capitalization and available supply variables. The correlation coefficient between the two variables is approximately 0.2, indicating that they are nearly independent of each other. The analysis above highlights a strong correlation between the two input variables, Price_usd and Price_btc, and the output variable. However, it is not possible to build a model that only considers these two variables while ignoring all other input variables. Conversely, while the input variables have a weak correlation with the output variable, they still have some degree of correlation with each other. In this study, the model considers all 12 input variables for predicting ICO prices.

Based on the correlation analysis depicted in Figure 4, two input variables, market_cap_usd and available_supply, are selected as they have a correlation coefficient close to 0 after rounding (correlation coefficient of the two variables is 0.00066). This suggests that the two variables are independent of each other. Assuming that the regression coefficient of market_cap_usd is X1, and the regression coefficient of available_supply is X2, Se(X1) represents the standard error of the regression coefficient of X1, while Se(X2) represents the standard error of the regression coefficient of X2. The standard error of regression is a measure used to assess the accuracy of the estimated regression coefficients. The results in Table 2 indicate that the regression coefficients, standard error of regression, and sum of squares practically remain unchanged. Therefore, it is concluded that for independent input variables, the values of regression coefficients, standard error of regression, and sum of squares do not change significantly.

Likewise, Figure 4 describes the correlation between the variables, and two quantitative variables, usd_raised and ico_duration, are chosen due to their strong correlation coefficient of approximately 0.77. Assuming that the regression coefficient of market_cap_usd is X1, and the regression coefficient of available_supply is X2, Se(X1) represents the standard error of the regression coefficient of X1, while Se(X2) represents the standard error of the regression coefficient of X2. The regression coefficient significantly changes when using two quantitative variables with a strong correlation coefficient as presented in Table 3.

Table 2. Correlation of Variables market_cap_usd and available_supply

Mô hình	**X_1**	**Se(X_1)**	**X_2**	**Se(X_2)**	**Sum of squares**
market_cap_usd	6.7e-10	2.6e-10	NA	NA	73.9
available_supply	NA	NA	-2.3e-12	4.2e-12	3.4
Model gồm 2 biến trên	6.7e-10	2.6e-10	-2.3e-12	4.1e-12	73.9/3.4

Table 3. Correlation Analysis of Variables ico_duration and usd_raised

Mô hình	**X_1**	**Se(X_1)**	**X_2**	**Se(X_2)**	**Sum of squares**
usd_raised	2.9e-09	1.0e-09	NA	NA	70
ico_duration	NA	NA	2.4e-02	7.7e-3	104.4
Model gồm 2 biến trên	3.8e-10	1.6e-09	2.2e-02	1.2e-02	70/34.9

DOI: https://doi.org/10.29207/joseit.v2i1.5022

In contrast to two variables that are almost independent, when two variables have a high correlation and are included in the model, the returned regression coefficient is significantly different from when using a univariate regression model. Specifically, the regression coefficient of usd_raised and ico_duration decreased when included in the model. The regression coefficient is influenced by the values of the independent variables. The linear regression equation is a function of the model's variables and the regression coefficients. If the values of the model's variables are large, the regression coefficient will be small to achieve a suitable result for the variables and other regression coefficients. Based on the analysis of the regression coefficients and the standard errors of the regression coefficients, the data in the regression model may have multicollinearity. To overcome this issue, the study proposes using a Ridge regression model that can handle multicollinearity.

The change in the standard error of the regression coefficient can affect its accuracy. For example, when using a univariate regression model with only the variable usd_raised, the standard error is 1.0e-9. However, when included in the multiple regression model, it increases to 1.6e-09. Similarly, the standard error of the parameter ico_duration also increases, from 7.7e-3 to 1.2e-02. This phenomenon is called multicollinearity.

Based on the two-error metrics for training and testing, we conclude that overfitting is present, causing the prediction model to become biased. Table 4 shows that the training error at this point is 0.0027, indicating that the prediction model performs very well on the training set. This result is achieved by fitting the dataset to a linear regression model with coefficients that minimize the difference between the actual and predicted values. This error metric demonstrates excellent prediction performance on the training set. However, the predicted values do not match the true values on the testing set. The predicted values have changed considerably compared to the true values. Based on the results in Table 4 for training and testing errors, we conclude that overfitting is present. To avoid this overfitting, adding a regularization term to the loss function will solve this problem.

3.2 Simulation results

To evaluate the accuracy of the algorithmic model, performance metrics are utilized, including the Root Mean Square Error (RMSE) as presented in equation (5), R^2 as presented in equation (6), and Mean Absolute Error (MAE) as presented in equation (7) (Jigar, Sahil, Priyank, & Kotecha, 2015). In this context, ȳ is defined as the predicted value of y, and y is defined as the actual value, while ȳ is considered as the mean value.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum\nolimits_{i=1}^{N}(y-\hat{y})^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y-\hat{y})^2}{\sum_{i=1}^{N}(y-\bar{y})^2}$$

$$MAE = \frac{\sum_{i=1}^{N} abs(y-\hat{y})}{N}$$

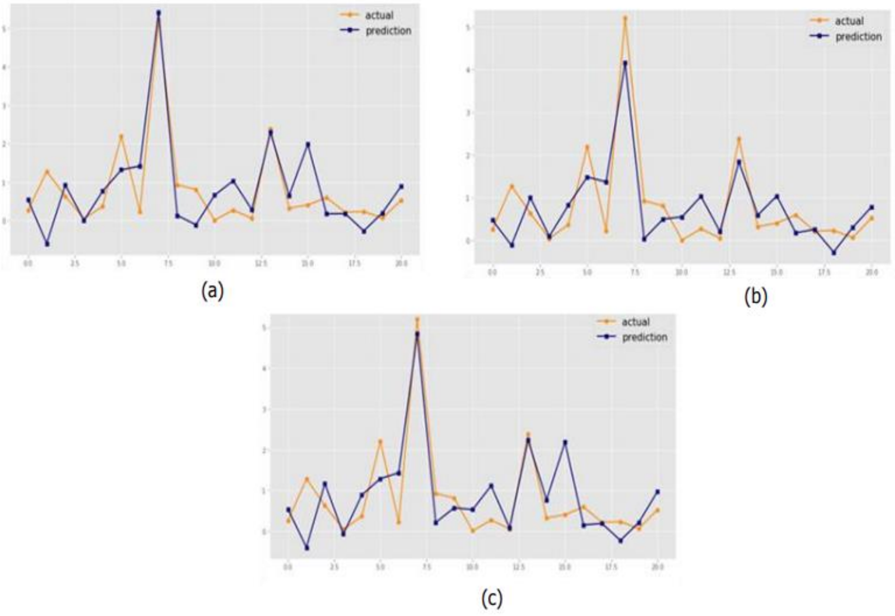


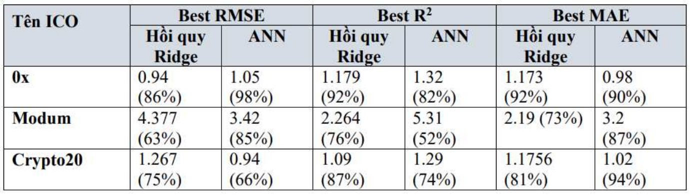
Figure 5. Evaluation of comparison results between yTestHoldout & yPredHoldout values using Ridge regression model (a) RMSE = 0.75 (b) Rsquared = 0.71 (c) MAE = 0.56

In this study, which discusses the occurrence of bullying in the YouTube streamer column using the Naïve Bayes method with Gain Ratio weighting has tested 3 times. From these tests, using the Naïve Bayes method obtained an accuracy result of 80%, then the Naïve Bayes method with Gain Ratio weighting obtained a result of 84%, there was an increase in accuracy of 4%. With this research that Gain Ratio weighting is an effective method to be able to optimize accuracy results with a combination of the Naïve Bayes method.

A random dataset was used to train three best-performing models. One model achieved the best performance in terms of the RMSE evaluation metric, another model achieved the best performance in terms of the R2 evaluation metric, and the third model achieved the best performance in terms of the MAE evaluation metric. To evaluate the Ridge regression algorithm, the training process was executed with 10,000 iterations to generate the model. The evaluation of the training results returned by each model was performed according to the RMSE, R2, and MAE evaluation metrics.

After the training process, the XTestHoldout data was fed into each of the best-trained models to perform the testing step. The result of the testing process returned the predicted data yPredHoldout. Then, it was combined with the yTestHoldout data to calculate the performance metrics of the model. Figures 5 and 6 (a), (b), (c) respectively show the evaluation results of the three best-performing Ridge regression and ANN models.

Table 5. Evaluation of three ICO coins in the unseen dataset to predict price prediction accuracy using Ridge and ANN regression models



Tên ICO	Best RMSE		Best R^2		Best MAE	
	Hồi quy Ridge	ANN	Hồi quy Ridge	ANN	Hồi quy Ridge	ANN
0x	0.94 (86%)	1.05 (98%)	1.179 (92%)	1.32 (82%)	1.173 (92%)	0.98 (90%)
Modum	4.377 (63%)	3.42 (85%)	2.264 (76%)	5.31 (52%)	2.19 (73%)	3.2 (87%)
Crypto20	1.267 (75%)	0.94 (66%)	1.09 (87%)	1.29 (74%)	1.1756 (81%)	1.02 (94%)

Table 4 presents the prediction results on the test set (testData) consisting of three ICOs selected from the dataset of 109 ICOs. This dataset was collected as the ICO price after 6 months of release, which was used in the application of two forecasting methods. Based on this dataset, a subset of 85 ICOs after 6 months was used as a training set to find the best model. The evaluation set consisted of 21 ICOs selected from this dataset. Three randomly selected ICOs that were issued at a price after 6 months were used as the test set to evaluate the accuracy of the forecasting methods. If the forecasted results have an accuracy close to the price after 6 months of release, it means that the predicted ICO has high accuracy.

From this, investors can predict which ICO to invest in for profit, meeting their expectations. The test set was used to evaluate the performance of the three best models using three performance metrics: RMSE, R^2, and MAE. All three algorithmic models yielded relatively accurate results. The simulation results using the Python programming language show that the Ridge regression algorithm with the R^2 performance metric optimized the forecasted results with an accuracy of up to 92% of the true value of the ICO for the case of the 0x ICO in the test set.

In the case of applying the Artificial Neural Network (ANN) algorithm, simulation results show that the model with the optimal MAE achieves the most accurate prediction up to 98% of the true value of Crypto20 ICO. The ANN algorithm produces higher accuracy in predictions than Ridge regression, but requires longer training time. Comparing these two forecasting methods is useful information for investors to choose the appropriate prediction method between the Ridge regression model and the ANN algorithm. When investors need to consider choosing a method for predicting ICO prices, they need to consider the hardware resources available to choose the appropriate prediction method. Ridge regression predicts with an accuracy of 63% to 92%, which is lower than the ANN algorithm model with an accuracy ranges from 52% to 98%. However, the Ridge regression model is simpler and requires less hardware resources for training compared to the ANN algorithm model.

4. Conclusion

Investors expect a return on investment when investing in ICOs. However, the price of ICOs depends on many unique factors as they are often invested in before being released. Therefore, managing risks and analyzing them is essential for successful ICO investments. Analyzing ICO prices with correlation analysis of twelve factors that impact ICO prices shows that the price_usd and price_btc factors have a high correlation with the output, while the remaining variables have weaker correlations. However, among the input factors, they are correlated with each other, so it is necessary to consider all these input factors as the main factors that affect ICO prices. Using these twelve factors as a basis for evaluating and comparing two forecasting methods, it is clear that these factors play an important role in predictive analysis.

Analyzing the correlation between the factors that affect ICO prices leads to the conclusion of multicollinearity phenomenon in the linear regression model. This phenomenon leads to bias in the results of the regression model. Overfitting when using the multiple regression model requires regularization techniques, which help to reduce overfitting by adding a regularization term to the regression model's loss function. The Ridge regression algorithm is a nonlinear regression method that can overcome the challenges of the current data problem that multiple regression cannot solve. Ridge regression uses the lambda coefficient to adjust the regression coefficient.

The dataset of 109 ICOs was collected and preprocessed as the dataset for two prediction models. Of these, 85 ICOs were used for training and evaluation, 21 ICOs were randomly selected to estimate the model's performance, and three ICOs were randomly selected for testing. After the training process, the best RMSE, R-square, and MAE criteria were used to find the optimal prediction model. The simulation results show that the accuracy of predicting ICO prices after six months was 92% of the actual value using the Ridge regression model with the test dataset in the case of the 0x ICO. The simulation results of using the ANN algorithm for prediction showed that the accuracy of the prediction reached 98% of the actual value with the test dataset in the case of the 0x ICO.

Therefore, comparing the two forecasting methods is useful information for investors to choose the appropriate forecasting method between the Ridge regression model and the ANN algorithm. When investors consider choosing the ICO price forecasting method, they need to consider the hardware resources they have to choose the appropriate forecasting method. Although the ANN algorithm has higher accuracy, it requires longer training time than the Ridge regression model, which is simpler and requires less hardware resources to train.

References

[1] R. J. Balvers and B. McDonald, "Designing a global digital currency," *J Int Money Finance*, vol. 111, 2021, doi: 10.1016/j.jimonfin.2020.102317.

[2] T. Zhang and Z. Huang, "Blockchain and central bank digital currency," *ICT Express*, vol. 8, no. 2, 2022, doi: 10.1016/j.icte.2021.09.014.

[3] J. Campino, A. Brochado, and Á. Rosa, "Initial coin offerings (ICOs): Why do they succeed?," *Financial Innovation*, vol. 8, no. 1, 2022, doi: 10.1186/s40854-021-00317-2.

[4] J. Campino, A. Brochado, and Á. Rosa, "Success Factors of Initial Coin Offering (ICO) projects," *Economics Bulletin*, vol. 41, no. 2, 2021.

[5] M. Hashemi Joo, Y. Nishikawa, and K. Dandapani, "ICOs, the next generation of IPOs," *Managerial Finance*, vol. 46, no. 6. 2020. doi: 10.1108/MF-10-2018-0472.

[6] N. Ayarci and A. O. Birkan, "Determinants of ICO investment decision: An exploratory factor analysis," *International Journal of Financial Research*, vol. 11, no. 5, 2020, doi: 10.5430/IJFR.V11N5P69.

[7] S. Jin, R. Ali, and A. V Vlasov, "Cryptoeconomics: Data Application for Token Sales Analysis," *Proceedings of the 38th International Conference on Information Systems (ICIS) 2017 Special Interest Group on Big Data, Seoul, South Korea*, 2017.

[8] Y. E. Shao, C. J. Lu, and C. D. Hou, "Hybrid soft computing schemes for the prediction of import demand of crude oil in Taiwan," *Math Probl Eng*, vol. 2014, 2014, doi: 10.1155/2014/257947.

[9] Y. Liu, J. Sheng, and W. Wang, "Technology and Cryptocurrency Valuation: Evidence from Machine Learning," *SSRN Electronic Journal*, 2021.

[10] R. S. Domingo, J. Piñeiro-Chousa, and M. Ángeles López-Cabarcos, "What factors drive returns on initial coin offerings?," *Technol Forecast Soc Change*, vol. 153, 2020, doi: 10.1016/j.techfore.2020.119915.

[11] A. Dürr, M. Griebel, G. Welsch, and F. Thiesse, "Predicting fraudulent initial coin offerings using information extracted from whitepapers," *European Conference on Information Systems*, no. May, 2020.

[12] M. Butterworth, "The ICO and artificial intelligence: The role of fairness in the GDPR framework," *Computer Law and Security Review*, vol. 34, no. 2, 2018, doi: 10.1016/j.clsr.2018.01.004.

[13] ICO, "Big Data, artificial intelligence, machine learning and data protection," *Data Protection Act and General Data Protection Regulation*, 2017.

[14] J. Witton, "The ICO guidance on big data, AI and machine learning - what humans can learn," *PwC blog*, no. July 10, 2017.

[15] H. Itakura, "A solution to multiple linear regression problems with ordered attributes," *Computers and Mathematics with Applications*, vol. 25, no. 2, 1993, doi: 10.1016/0898-1221(93)90222-H.

[16] M. S. Khrisat and Z. A. Alqadi, "Solving multiple linear regression problem using artificial neural network," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 1, 2022, doi: 10.11591/ijece.v12i1.pp770-775.

[17] I. G. P. Christyaditama, I. M. Candiasa, and I. G. A. Gunadi, "Optimization of artificial neural networks to improve accuracy of vocational competence selection of vocational school students using nguyen-widrow," *J Phys Conf Ser*, vol. 1516, p. 012052, Apr. 2020, doi: 10.1088/1742-6596/1516/1/012052.