ISSN: 2710-0588

Comparing the Performance of Four Machine Learning Algorithms in Prediction

Bi He*

School of Civil Engineering, Shandong Jiaotong University, Jinan 250357, China

*Corresponding Author

Abstract

To investigate and compare the performance of different machine learning algorithms in forecasting, this paper presents a case study on predicting industrial land use. We conduct a benchmarking study employing four representative algorithms: Linear Regression, Random Forest, Naïve Bayes, and Artificial Neural Networks (ANN). The historical data of Shandong Province from 2001 to 2022 was established as the dataset for this study, in which GDP, FAI(Fixed Asset Investment), and IOVAS(Industrial Output Value Above Scale) were set as independent variables, while ILV(Industrial Land Volume)was used as the dependent variable, and four machine learning algorithms, utilized the dataset to complete the training and testing of the models, and finally, this study provides the accuracy of these four machine learning algorithms in industrial land use prediction.

Keywords

Machine Learning, Algorithms, Prediction.

1. Introduction

Machine learning is a core branch of artificial intelligence. Its fundamental premise is to endow computer systems with the ability to "learn" from data and make decisions or predictions without relying on explicit, pre-defined programming instructions. In contrast to traditional programming, where humans input rules and data to solve problems, machine learning involves feeding algorithms data and expected answers (or allowing the algorithms to discover patterns autonomously). The algorithm then automatically constructs a "model" for prediction or decision-making[1]. This model is, in essence, a complex mathematical function capable of capturing the underlying patterns and relationships hidden within the data. This study employs four representative machine learning algorithms to predict industrial land use in Shandong Province, China, with the aim of comparing their performance.

2. Data and Method

2.1. Data

This study took the Shandong province (located in Northen China) as example subject, from the official statistical yearbook released by the Shandong Provincial Bureau of Statistics, we obtained the datasets, mainly includes GDP, FAI, IOVAS and ILV from year 2001 to 2022. It is organized and shown in Table 1.

ISSN: 2710-0588

Table 1. Datasets of study

Year	ILV (km^2)	GDP (Billon CNY)	FAI (Billon CNY)	IOVAS (Billon CNY)
2001	363.5	943.83	280.78	937.74
2002	413.5	1007.65	350.93	1103.85
2003	472.6	1090.32	532.84	1493.22
2004	557.1	1330.81	762.90	2105.51
2005	603.0	1594.75	1054.19	3002.39
2006	643.7	1896.78	1113.61	3811.61
2007	690.7	2271.81	1253.70	4918.62
2008	733.6	2710.62	1543.59	6203.42
2009	734.4	2954.08	1903.10	7082.61
2010	775.7	3392.25	2327.67	8366.30
2011	801.1	3906.49	2592.71	9976.62
2012	819.5	4295.73	3031.98	11808.69
2013	807.9	4734.43	3587.59	13213.03
2014	934.6	5077.48	4159.91	14314.03
2015	1044.1	5528.88	4738.15	14562.89
2016	997.1	5876.25	5236.45	15064.12
2017	1025.0	6301.21	5423.60	14085.68
2018	1069.3	6664.89	5789.18	14658.24
2019	1113.4	7054.05	5981.65	16186.16
2020	1094.5	7279.82	6309.00	17379.05
2021	1137.5	8287.52	6737.01	18957.13
2022	1232.2	8743.51	7142.97	21030.09

The correlation between dependent variable and independent variable should be test when construct the prediction model., this indicator is used to prove that the independent variable can significantly affect the dependent variable, and the prediction model based on this has practical significance. Pearson's correlation coefficient was utilized to test the relationship between the independent and dependent variables in this study [2][3], and the results are shown Table 2:

Table 2. r_value between each variable of Pearson's correlation coefficient

	GDP	FAI	IOVAS
ILV	r=0.974242	r=0.970758806	r=0.971471739

According to the r_value between each independent variable and dependent variable, it can be judged there is significantly affect between them. This result ensures that this dataset can be utilized to build predictive models.

2.2. Method

2.2.1. Liner Regression

Linear regression is one of the most widely used machine learning algorithms and it has the advantage of being easy to understand. Linear regression expresses the relationship between one or more independent variables and the dependent variable by establishing an equation for the relationship between them, which can be expressed in Equation 1:

ISSN: 2710-0588

$$y = \beta_0 + \beta_i * x_i \quad (1)$$

In this model, y is the dependent variable, x_i are the independent variable, β_i are the coefficient of the independent variable, and β_0 is a constant term. By training the model with training samples, using methods such as least squares, β_i and β_0 can be derived, and when they are derived, the model can be used to predict the value of the dependent variable y corresponding to different independent variables [4].

2.2.2. Random Forrest

Random forest is an algorithm that integrates multiple decision trees using the idea of ensemble learning [5]. When a prediction is needed, the results of multiple estimators are combined through an integrator as the final output. Principles of random forest algorithm can be expressed by Figure 1.

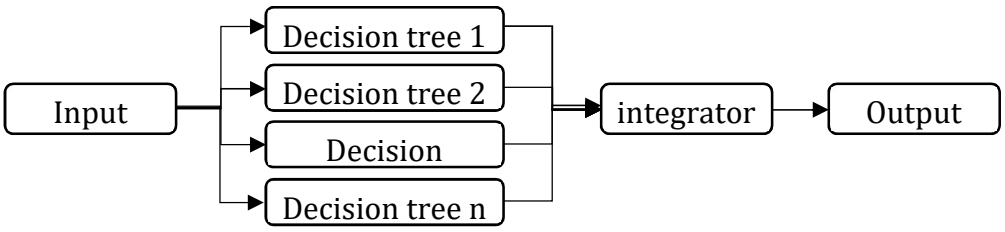

Figure 1. Principles of random forest algorithm

2.2.3. Bayesian

Bayesian regression is a type of conditional modeling in which the mean of one variable is described by a linear combination of other variables, with the goal of obtaining the posterior probability of the regression coefficients (as well as other parameters describing the distribution of the regression) and ultimately allowing the out-of-sample prediction of the regress and conditional on observed values of the regressors [6]. The simplest and most widely used version of this model is the normal linear model, in which dependent variable given independent variables is distributed Gaussian. In this model, and under a particular choice of prior probabilities for the parameters, the posterior can be found analytically. With more arbitrarily chosen priors, the posteriors generally have to be approximated.

2.2.4. Artificial NEURAL NETWORK

Artificial neural network is an algorithmic model that mimics the human brain's thinking, it can be applied in both classification and regression problems, and it is capable of handling both linear and nonlinear problems. Structurally, artificial neural networks are mainly composed of input layer, hidden layers and output layer [7]. By training samples, the weight matrix of the hidden layers is constantly adjusted to approximate the sample values, and the trained model can be used for prediction.

2.2.5. Accuracy Evaluation

In this study, the MSE (Mean Square Error) is used to rate the accuracy of each prediction model, and the MSE can be represented by Equation 2:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \quad (2)$$

Where Y_i denotes real value, $\widehat{Y}_i$ represents predict value.

ISSN: 2710-0588

3. Results and Discusses

Python was selected as the data process and calculate tool in this study, and some libraries like NumPy, PANDAS, Ski-learn were imported to enhance it. After processed by the Python, the real values and predict values of Shandong province industrial land use by four models sorted and expressed by Figure 2.

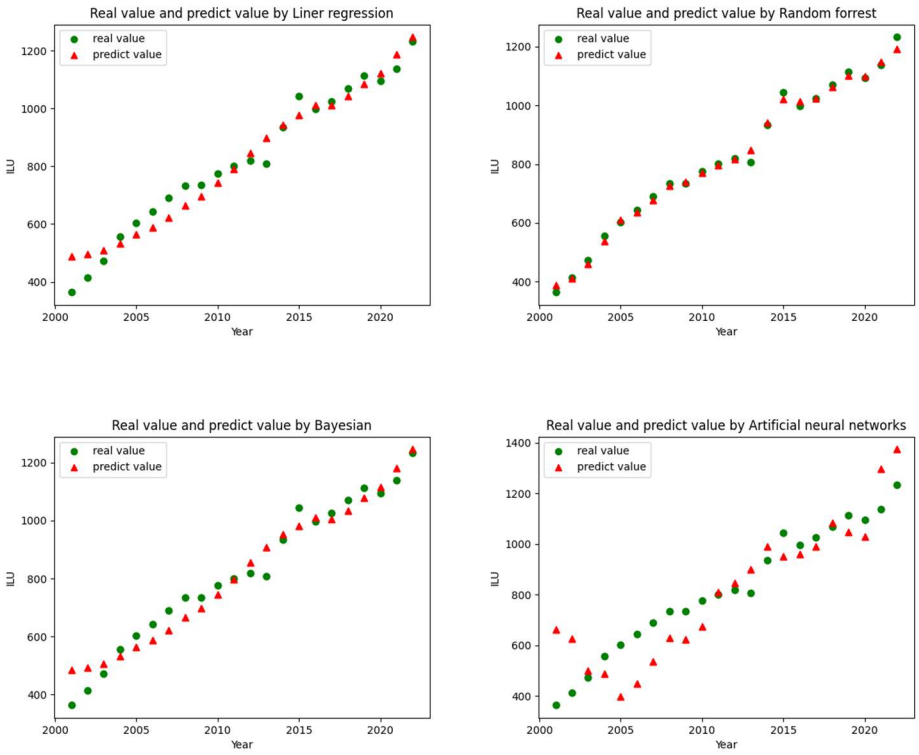

Figure 2. Scatters of real values and predict values of each model

After trained and predicted each model, the MSE of each model can be calculated also with Python, its results were sorted and expressed in Table 3.

Table 3. MSE of each model

Linear regression	Random forest	Bayesian	Artificial neural networks
2727.68	293.11	2755.01	18856.81

According MSE of all models shown in Table 3, there is lowest or 293.11 MSE of Random Forest, it implied in this prediction, accuracy of Random Forest is the highest level.

4. Conclusion

The purpose of this study is to investigate and compare the performance of different machine learning algorithms in forecasting. And we got the following finding:

When set GDP, FAI, and IOVAS as independent variables, and ILV as the dependent variable to build the regression model, the correlation coefficient between the independent and dependent variables was able to reach more than 0.97, which indicates that it is reasonable to use these three variables to predict the amount of industrial land.

ISSN: 2710-0588

All of linear regression, random forest, Bayesian and artificial neural network can be used to build the prediction model for industrial land use. Particularly, in this study, according to the MSE as evaluation indicator, we found the accuracy level of random forest is highest.

In the future study, there are two issues supposed to be addressed, one is to find more and better variables to make predictions, and the other is to further investigate the reasons for the varying accuracy of these algorithms.

References

[1] Yuan Meng, Feng-Rong Zhang, Ping-Li An, et al., Industrial land-use efficiency and planning in Shunyi, Beijing, Landscape and Urban Planning, Volume 85, Issue 1, 2008, Pages 40-48,

[2] Egghe, L., & Leydesdorff, L. (2009). The relation between Pearson's correlation coefficient r and Salton's cosine measure. Journal of the American Society for information Science and Technology, 60(5), 1027-1036.

[3] Afyouni, S., Smith, S. M., & Nichols, T. E. (2019). Effective degrees of freedom of the Pearson's correlation coefficient under autocorrelation. NeuroImage, 199, 609-625.

[4] Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends, 1(4), 140-147.

[5] Cai, J., Xu, K., Zhu, Y., Hu, F., & Li, L. (2020). Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. Applied energy, 262, 114566.

[6] Izmailov, P., Vikram, S., Hoffman, M. D., & Wilson, A. G. G. (2021, July). What is Bayesian neural network posteriors really like? In International conference on machine learning (pp. 4629-4640). PMLR.

[7] Chen, Y., Song, L., Liu, Y., Yang, L., & Li, D. (2020). A review of the artificial neural network models for water quality prediction. Applied Sciences, 10(17), 5776.