Asian Journal of Social Science Studies; Vol. 4, No. 4; 2019
ISSN 2424-8517 E-ISSN 2424-9041
Published by July Press

Regression Model for Bike-Sharing Service by Using Machine Learning

Zhifeng Wang[1]

[1] Department of Engineering and Computer Science, Australian National University Canberra, Australia

Correspondence: Zhifeng Wang, Department of Engineering and Computer Science, Australian National University Canberra, ACT 2617, Australia.

Received: October 11, 2019 Accepted: November 4, 2019 Online Published: November 6, 2019

doi:10.20849/ajsss.v4i4.666 URL: https://doi.org/10.20849/ajsss.v4i4.666

Abstract

The bike sharing system has brought wide convenience to residents in the city and serves as important tools to transport from one place to another place. For the bike sharing companies, they need to know the total users of bike, so they can release suitable number of bikes into the market. This paper uses visualization technology to visualize data and figure out the possible factors which can impact the total number of users. After completing the data analyzing, this paper figures out the season, weather sit, feeling temperature, humanity and wind speed are the main factors which can have impacts on the total number of users. In the second stages, this paper uses regression model, NN model, ELM model and DELM model to predict the possible number of bike users. The input factors are season, weather sit, feeling temperature, humanity and wind speed. By analyzing regression model results, the ELM model has the best prediction, which can be used for real practice.

Keywords: bike sharing, neural network, extreme learning machines, regression model

1. Introduction

Bicycle Sharing Systems are very green, healthy and cheap way to navigate from one place to another place. Now with the new methods of electronic sharing and registration, the whole process of bicycle sharing, from the rental to returning back has become much more automatic and convenient. Through the bicycle sharing system, the users can easily rent a bicycle from one place and return it in another place. Bicycle sharing companies such as Airbike, OBike and Mobike has become much more popular in the world in the past few years, due to the pro-health and environment-friendly mode of transport. The bike sharing systems can bring a lot of data, but the characteristics of the data have not been fully understood. Opposed to other data which is generated by transport devices such as subway and bus. The arrival and departure position are clearly recorded in the bus or subway systems. The bike sharing system is a virtual sensor network since the bike can be used to detect the most important events in the city by monitor these data.

Although monitoring the number of bikes in the city and figure out the factors which influence the number of bike users is not easy, many scholars provide some algorithms to predict the number of bike users. Short-term traffic forecasting is an important field to predict the number of sharing bikes. Wang use RNNs (LSTM and GRU) and Random Forest methods to predict short-term available number of bikes (Wang & Kim, 2018). Wang (2016) found that the tree-based and NN-based models can achieve higher accuracy in bike rental prediction. Yan (2019) used the deep LSTM model with two layers to predict the number of bike renting and returning and use the long short term memory and the ability to deal with the sequence data of recurrent neural network. Campbell (2016) pointed out that the bike demand is impacted by temperature, air quality, trip distance, and precipitation. Zhang et al. (2016) developed DeepST methods to predict the NYC bike usage.

2. Methodology

2.1 Data Collection

The dataset in this study case is chosen from Capital Bikeshare System, Washington D.C., USA which includes two years dataset from 2011 to 2012. The dataset has 15 attributes including weather situation, date, weekday/public holiday, the count of number of bikes rented on that day and temperature (Fanaee-T, Hadi, Gama & Joao, 2013).

2.2 Data Description

The bike sharing data is recorded on daily basis. The records are about 731 days. All the data information can be shown in Table 1.

Table 1. Bike sharing dataset

bike-share

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
4	2011-01-04	1	0	1	0	2	1	1	0.2	0.212122	0.590435	0.160296	108	1454	1562
5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.22927	0.436957	0.1869	82	1518	1600
6	2011-01-06	1	0	1	0	4	1	1	0.204348	0.233209	0.518261	0.0895652	88	1518	1606
7	2011-01-07	1	0	1	0	5	1	2	0.196522	0.208839	0.498696	0.168726	148	1362	1510
8	2011-01-08	1	0	1	0	6	0	2	0.165	0.162254	0.535833	0.266804	68	891	959
9	2011-01-09	1	0	1	0	0	0	1	0.138333	0.116175	0.434167	0.36195	54	768	822
10	2011-01-10	1	0	1	0	1	1	1	0.150833	0.150888	0.482917	0.223267	41	1280	1321

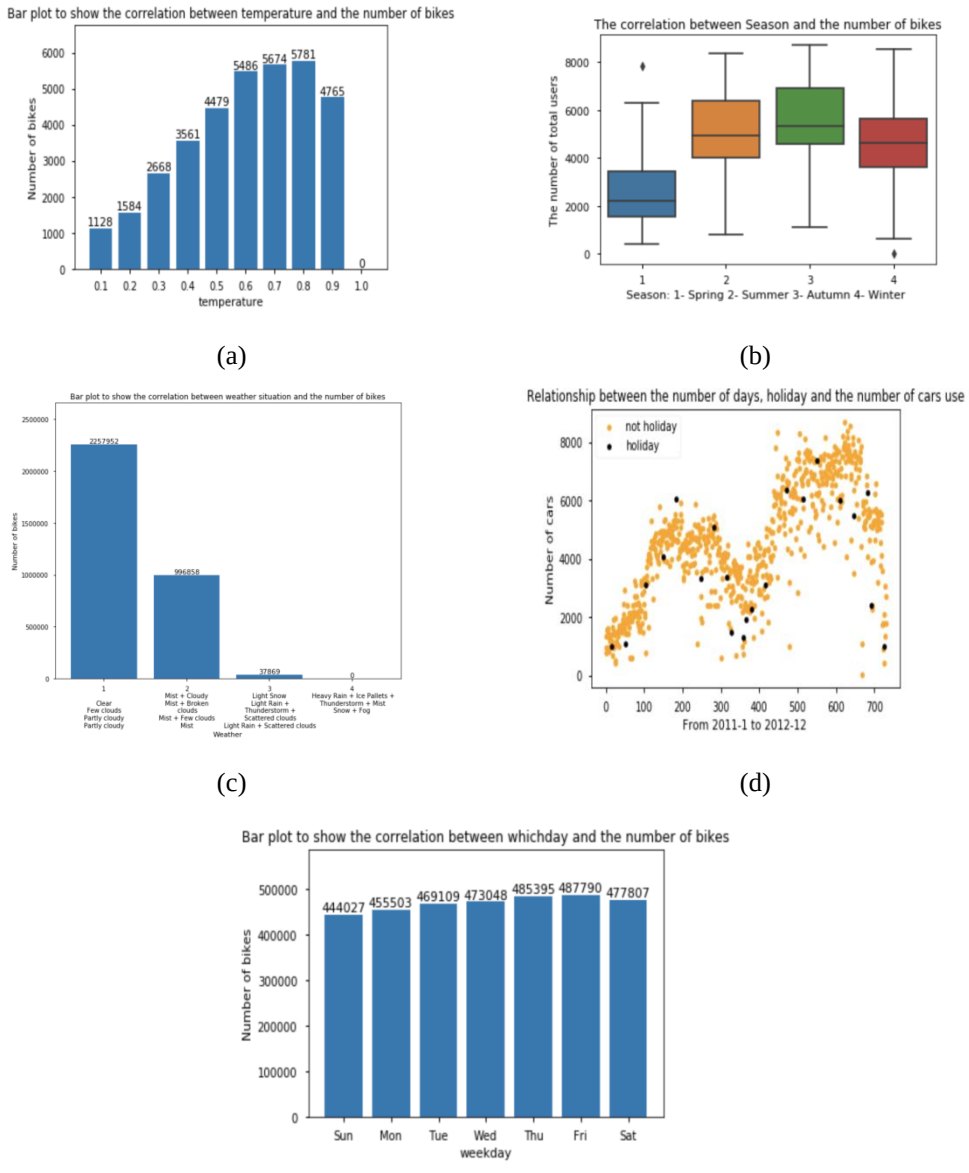
In the bike sharing dataset, the dataset has following fields:

1. Instant: record index
2. dteday: date
3. Season: season (1: spring, 2: summer, 3: fall, 4: winter)
4. yr: year (0: 2011, 1:2012)
5. mnth: month (1 to 12)
6. holiday: weather day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)
7. weekday: day of the week
8. workingday: if day is neither weekend nor holiday is 1, otherwise is 0.
9. weathersit: -1: Clear, Few clouds, Partly cloudy, Partly cloudy
   -2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
   -3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
   -4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
10. temp: Normalized temperature in Celsius. The values are divided to 41 (max)
11. atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
12. hum: Normalized humidity. The values are divided to 100 (max)
13. windspeed: Normalized wind speed. The values are divided to 67 (max)
14. casual: count of casual users
15. registered: count of registered users
16. cnt: count of total rental bikes including both casual and registered

2.3 Data Analysis

In order to figure out which factors have impact on the number of bikes. The bar plot, box plot and scatter plot will be drawn to show the influence on the number of bikes. Figure 1.a shows that there is a relationship between temperature and number of users, the temperature is 0.8, which has the highest number. Figure 1.b shows that there is a correlation between Season and the number of users. Summer and Autumn has the most users. Figure 1.c: The weather and number of bikes bar plot shows that when the weather sit is in 1 means that the weather is clear, few clouds mist or broken cloud, the total number of bikes is around 2,250,000. When the weather sit is in 2 means that the weather is mist and cloudy, mist and broken or mist, the total number of bikes is around 996,000. When the weather sit is in 3 means that the weather is light snow, light rain and thunderstorm or light rain and scattered clouds, the total number of bikes is around 37,000. No users use bikes in weather sit 4 which are heavy rain and ice pallets and thunderstorm and mist or snow and fog. Figure 1.d: As we can see that the

yellow dot is for the number of users which is not in holiday, the number of users change depending on the time. As we can see in 2011-01-01 which respond to 0 in the x-axis, which is in the winter. The number of users is very small. With the time is going on, the number of users increase and at the summer, the figures reach peak. Then the figures decrease from autumn to winter. And the 2012 has the similar rule. But in 2012, the largest number of users is larger than the number of 2011, and the total number of users is also larger than the number of 2011. The black dot is for the number of users in holiday. As we can see that the number of users in holiday also has the similar relationship with the time. The season factor will impact the user's number. The everyday data can be seen in the following graph-Figure 1.e



(a) (b)

(c) (d)

(e)

Figure 1. Correlation between different factors and total users

3. Models

3.1 Linear Regression Model

In order to predict the number of bikes in the future, this paper will build a multiple linear regression model. The multiple linear regression model is to model the correlation between two or more variable and a response

variable by fitting a linear equation to observed data.

A multiple linear regression model with k predictor variable x1,x2,…,xp and a response Y, can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \quad (1)$$

The ϵ is the error term which can be used to estimate the accuracy of the linear regression model. When giving a data set $\{y_i, x_{i1}, \ldots x_{ip}\}_{i=1}^{n}$ of the n statistical units, a linear regression model assume that the relationship between the dependent variable y and the given x is linear. The ϵ is the error variable. Thus the model can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \quad i = 1, \ldots, n \quad (2)$$

The formula (2) can be denoted as matrix notation as

$$y = X\beta + \epsilon \quad (3)$$

where

$$y = \begin{Bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{Bmatrix}, \quad X = \begin{Bmatrix} x_1{}^T \\ x_2{}^T \\ . \\ . \\ . \\ x_n{}^T \end{Bmatrix} = \begin{Bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots\vdots & \vdots \\ . & . & \ldots & . \\ 1 & x_{n1} & \ldots & x_{np} \end{Bmatrix}, \quad \beta = \begin{Bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_p \end{Bmatrix}, \quad \epsilon = \begin{Bmatrix} \epsilon_0 \\ \epsilon_1 \\ . \\ . \\ . \\ \epsilon_p \end{Bmatrix}$$

3.2 Variable Selection in Linear Regression Model by Using the AIC

A variable selection method should to select the best variable for a special purpose such as prediction. The best should find the balance between the goodness of fit and the number of variables. Earlier selection standard uses the residual sums of squares which have overfit problems (Miller, 2002). Akaike found a Akaike Information Criterion method which can be used for model or variable selection via Kullback-Leibler divergence (Gutierrez & Heming, 2018). The AIC can be given as:

$$AIC_P = nln(\sigma_{mle}^2) + 2p \quad (4)$$

Where σ_{mle}^2 is the maximum likelihood estimate of σ^2. By using this selection standard, the model with the smallest AIC can be deemed as the best (Akaike, 1974).

3.3 Neural Network

The Neural Network include input layer, hidden nodes and output layer. The Neural Network can be calculated as following formula. In the formula, the x_i^{l-1} is the output from previous layer, x_j^l is the output of the ith channel of the jth layer. f is the activation function. k_{ij}^l is a kernel and b_j^l is the corresponding offset (Bouvrie, 2006).

$$x_j^l = f\left(\sum_{i \in M_j} x_i^{l-1} * k_{ij}^l + b_j^l\right) \quad (5)$$

3.4 ELM (Extreme Learning Machine)

Extreme learning machine are feedforward neural networks which can be used for regression, clustering and classification with a single layer of hidden nodes. The ELM training model can be treated as following model. In the single hidden layer sigmoid neural networks, the W_1 is the matrix of input to hidden layer weights, σ is the activation function, and W_2 is the matrix of hidden to output layer weights. The algorithm mainly includes two steps. The first step is to fill W_1 with random values, the second step is to estimates W_2. Calculating W_2 by using pseudoinverse. The algorithm can be written as formulas (Huang, Ding & Zhou, 2010):

$$\check{Y} = W_2 \sigma(W_1 X) \quad (6)$$

$$W_2 = \sigma(W_1 X)^+ Y \quad (7)$$

4. Results

4.1 Linear Regression Model Result

In order to calculate the number of bikes in the future, the linear regression model can help us to predict the number of users in the future. The regression parameters

$\beta = \{\beta_0, \beta_1, \beta_2, \ldots, \beta_p\}$ will be calculated by using the training set while the AIC values will be implemented for selecting the variables in the bike sharing model. The Residual Mean Squared Error (RMSE) and R^2 are calculated. The Residual Mean Squared Error (RMSE) can be written as:

$$RMSE = \sqrt{\sum_{i=1}^{k}\left(Y_i - \check{Y}_i\right)^2 / k} \tag{8}$$

Where k is the length of the test set. By using Python Statsmodels to get linear regression model statistics information, the summary of the linear regression statistics information can be shown as Table 2.

Table 2. Linear regression model statistics information

	coef	std err	t	P>|t|
const	3297.2475	342.853	9.617	0.000
Season	406.3726	49.073	8.281	0.000
weather	-449.1705	120.199	-3.737	0.000
Temp	2539.1628	2165.114	1.173	0.241
Feeltemp	3558.6258	2451.185	1.452	0.147
Humidity	-2419.8385	477.214	-5.071	0.000
windspeed	-3187.6293	704.153	-4.527	0.000

In the linear regression model, the predictor variable is season, weather, temp, feeltemp, humidity and windspeed. The response variable is totaluser. In order to reduce the standard error of the variables in the linear regression model, this paper used the AIC method to remove some variables in the linear regression model to get the best model. The AIC results can be shown as Table 3.

Table 3. Calculate the value of AIC for different variables

Combination	AIC
season', 'weather', 'feeltemp', 'humidity', 'windspeed	12617.0083
'season', 'weather', 'temp', 'feeltemp', 'humidity'	12636.0244
'season', 'weather', 'temp'	12650.3454
'feeltemp', 'humidity', 'windspeed'	12692.5960
'weather', 'temp', 'feeltemp', 'humidity'	12709.2905

By calculating AIC value, the variable combination [season', 'weather', 'feeltemp', 'humidity', 'windspeed] has the smallest AIC, so this combination is the best. After removing variable temp in the previous linear regression model, this paper updates the linear regression model and calculate Linear regression model statistics information, the new Linear regression model statistics information can be shown as Table 4.

Table 4. New Linear regression model statistics information

	coef	std err	Improve	t	P>|t|
const	3194.4497	331.545	3.41%	9.635	0.000
Season	406.0427	49.085	-0.02%	8.272	0.000
weather	-443.7828	120.143	0.05%	-3.694	0.000
Feeltemp	6406.0994	336.371	628.71%	19.045	0.000
Humidity	-2455.1804	476.385	0.17%	-5.154	0.000
windspeed	-3054.0407	695.059	1.31%	-4.394	0.000

For the standard error, the standard error for const improve about 3.41%, the standard error for weather improve

about 0.05%, the standard error for Feeltemp improve about 628.71%, the standard error for Feeltemp improve about 0.17%, the standard error for windspeed improve about 1.31%. Therefore, the AIC method can effectively reduce the linear regression model standard error. The final linear regression prediction model can be shown as Figure 2.

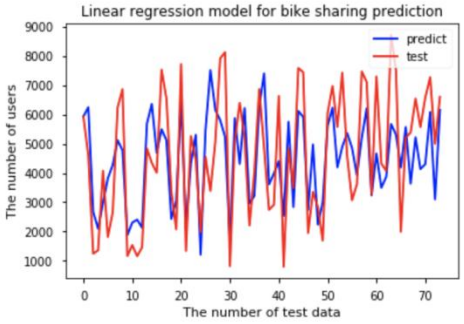


Figure 2. Linear regression model for bike sharing prediction

4.2 Results for ELM model

In the Neural network and ELM model, the input parameter, output parameter, number of hidden layers, number of nodes of hidden layers, activation function and number of data can be descripted as Table 5(a). The RMSEs for Linear regression model, NN, ELM and DELM model can be shown as Table 5(b). In the Table 5(b), we can see that the RMSEs for ELM is about 1214, which has the smallest number, which means that it has the best performance in all the regression models. Figure 3 shows the regression model prediction for bike sharing total users by using different hidden nodes. By analyzing the Figure 3, the ELM model has good performance for the total users and researchers can use the ELM model to predict the possible number of bike users in the future. The number of hidden nodes can have important impacts on the accuracy of models.

Table 5(a). Experimental parameters

Parameter	**value**
Input parameter	4
Output parameter	1
Number of hidden layers	1
Number of nodes of hidden layers	20,30,40
Activation function	sigmoid
Number of data	732

Table 5(b). Average RMSEs for each method

Method	Training set	Test set
Linear regression model	-	1522
NN	1173	1256
ELM	1182	1214
DELM	1183	1293

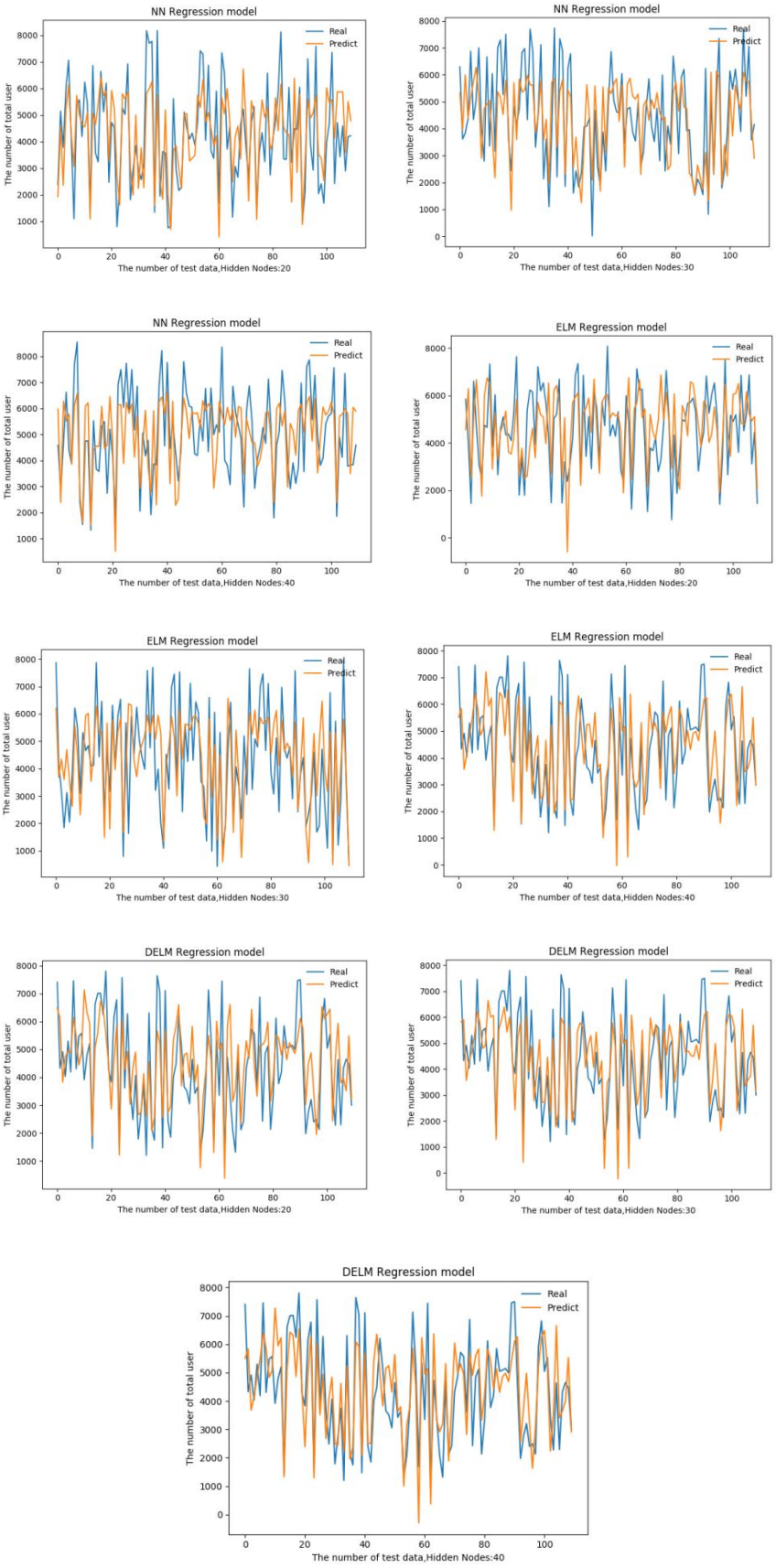


Figure 3. Regression model prediction for bike sharing total users

5. Conclusion and Discussion

Regression model is important method in predicting the number of total users. A good prediction can help managers manage their time, total users and investment. This paper mainly focuses on using ELM model to predict the bike sharing total users. In this paper, the main factors include weather, feeling temperature, weather sit, humanity and windspeed. These factors are chosen by using AIC methods. By analyzing average RMSEs of ELM, NN and Linear regression model, the ELM regression model can achieve has the smallest errors in the test data and the best performance. These methods for predicting the number of total users in bike sharing can become a management reference. However, because in the experiments, this paper only uses single layer hidden node to train the model, and every layer in the neural network uses no larger than 40 nodes. In the input layers, the factors include season, weather, feel temperature, humidity and windspeed. The output layers include the number of total users. In order to improve the accuracy of predicting the total number of bike usage, multiple neural network can be for future research direction. Furthermore, the time factor and previous total number of users should be taken into consideration for better prediction in the future research.

References

Artificial Intelligence. (2013). Springer Berlin Heidelberg.

Bouvrie, J. (2006). *Notes on convolutional neural networks.*

Campbell, A. A., Cherry, C. R., Ryerson, M. S., & Yang, X. (2016). Factors influencing the choice of shared bicycles and shared electric bikes in Beijing. *Transportation Research Part C: Emerging Technologies, 67*, 399-414.

Gutierrez, E. E., & Heming, N. M. (2018). *Introducing AIC model averaging in ecological niche modeling: a single-algorithm multi-model strategy to account for uncertainty in suitability predictions*.

Hadi, F.-T., & Gama, J. (n.d.). *Event labeling combining ensemble detectors and background knowledge*.

Huang, G. B., Ding, X. J., & Zhou, H. M. (2010). Optimization method based extreme learning machine for classification. *Neurocomputing, 74*(1-3), 155-163.

Miller, A. (2002). *Subset Selection in Regression.* London: Chapman & Hall.

Pan, Y., Zheng, R. C., Zhang, J., & Yao, X. (2019). Predicting bike sharing demand using recurrent neural networks. *Procedia Computer Science, 147,* 562-566.

Wang, W. (2016). *Forecasting Bike Rental Demand Using New York Citi Bike Data*.

Zhang, J., Zheng, Y., Qi, D., Li, R., & Yi, X. (2016). DNN-based prediction model for spatio-temporal data. *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.*

Copyrights