DOI:10.18686/ahe.v7i31.11564

# Prediction of Wordle Based on Machine Learning

**Yufei Shi**
Tianjin University , Tianjin 300100 China

**Abstract:** Wordle brings a lot of fun to people's daily lives. Players are required to guess words in less than or equal to six attempts, and people can choose difficult or easy modes according to their own wills to experience the game. A series of data generated from the number of sign-ups in different modes and the number of guesses of different words. We use machine learning to count and process the data of this game and classify the difficulty of different words to predict the situation that the players will face in guessing the words to be introduced in the future. We propose the following solutions for the questions posed[1].

For problem one, we derive a dataset to train the LSTM model by sliding window processing, and then predict the number of reported results for March 1, 2023, with a value of roughly 23321. After that we validate and normalize the model, and the computational results showed that our model has little error and strong prediction effect. For the proportion of word attributes on the number of enrollment in different difficulty modes, we counted all combinations of word letters and obtained prob (F) = 0.587 by the linear regression model, which means that the significance of the model is low, indicating that enrollment is not affected by word attributes.

For problem two, to predict the distribution of the given word results for one day in the future, considering the circumstance of multiple inputs and multiple outputs, we adopt a regression chain model. Then we train a random forest regression algorithm based on the model, and divide the samples into training and test sets. Finally we derive percentage data for seven attempts: {0.2,3.5,18.3, 31.3,27.5,15.8,2.9}, whose MAPE are within acceptable limits. We then construct a mapping set on the attributes of the given word EERIE and derive the predicted data for the word. What's more, we compare the result with that obtained from the data processed by the neural network algorithm and find that the model used in the former is better[2].

For problem three, we divide the difficulty into three levels by RSR method, and export the data after the evaluation process. Then we train the data set by three machine learning algorithms, namely, logistic regression, decision tree and XGBOOST, and draw the corresponding learning curves. There are underfitting and overfitting phenomena, and the logistic regression model with the best effect among the three still failed to show a better fit in the test set, with an F-score of 0.5. So we continue to use CNN for its classification prediction, and the final F-score of both training and test sets is about 0.8, which we think is a good effect. Finally, we analyze the difficulty of EERIE by this model, and the difficulty factor we get is 1, which means it's easy.

For problem four, we present the data in graphical form and analyze its relevant features through correlation and descriptive analysis methods[3].

**Keywords:** Prediction of attempt numbers; Machine learning; Significance test; Difficulty coefficient

## 1. Models

### 1.1 Model Establishment of question1

To create a forecast interval for the results reported on March 1, 2023, the following modeling

steps can be taken:

Step 1: First, the required information is retained from the cell through the forgetting gate

$$f_t=\sigma(b_f+W_f x_t+U_f h_{t-1}) \quad (1)$$

In the formula: σ Activate the function for sigmoid. Xt is the input vector at time t, ht- 1 is the hidden layer vector at time t- 1, bf

is the bias cycle weight, Wf is the input weight, and Uf is the forgetting gate weight.

Step 2: Reset the information in the cell. Gt Output gate between 0- 1 is controlled by sigmoid activation function

$$g_t=\sigma(b_g+W_g x_t+U_g h_{t-1}) \quad (2)$$

Step3: Update the cell state Ct on the basis of Ct-1

$$C_t=f*C_{t-1}+g_t*\tanh(b_c+W_c x_t+U_c h_{t-1}) \quad (3)$$

Step4: Control information output by output gate ot

$$h_t=o_t*\tanh(C_t) \quad (4)$$

Among them, output gate

$$o_t=\sigma(b_0+W_o x_t+U_o h_{t-1}) \quad (5)$$

Through the above steps, the LSTM model can efficiently use the input historical data, thus having the memory function.

Step4: Use the above in-depth learning LSTM model to train the moving average value obtained previously as the input, and iterate the data for 100 times.

The following steps can be taken to determine whether the word attribute affects the percentage of the reported score played in difficult mode：

Step 1: Some indicator features can be constructed to extract the initial data

Step 2: Set the reading code of 26 letters and extract the combined features of all letters

Step 3: Extract the percentage of difficulty in selecting the report results

$$\varepsilon=（\text{Number in hard mode}）/(\text{Number of reported results}) \quad (6)$$

Step4：Set x-y, and use the linear regression model to analyze the effect of word attributes (suchas word length) on the percentage of difficulty. The formula is as follows:

$$y=bx+a \quad (7)$$

$$b=[n\sum^{n}i=1xiyi-(\sum^{n}i=1xi)(\sum^{n}i=1yi)]/[n\sum^{n}i=1xi^2-(\sum^{n}i=1xi)^2] \quad (8)$$

$$a=(\sum^{n}i=1yi/n)-b(\sum^{n}i=1xi/n) \quad (9)$$

## 1.2 Model Establishment of question2

Step 1: We created the RegisterChain model, which implements a multi-layer perceptron (i.e., neural network) through MLPRegressor.

Step 2: Fit the model to the training data tr_ X and tr_ y.

Step 3: Evaluate the model, calculate the average absolute percentage error on the training set and test set, and output the results.

Step 4: Finally, predict the new input data cc (EERIE) and output the results[4].

(Where, Y represents all output variables, and there are 7 in total. tr_x, tr_y, te_x and te_y represent training set input, training set output, test set input and test set output respectively. MAPE is a user-defined function for calculating the average absolute percentage error. Convert the new input data cc to be predicted into an array of 1 row and n columns to meet the input requirements of the model.)

We also adopted the forest regression algorithm, and the solution steps are as follows:

Stochastic forest regression algorithm is a supervised machine learning algorithm for regression tasks. This is an integrated method, which constructs multiple decision trees and combines their outputs to produce a single prediction[5].

The following are the steps to build a random forest regression model:

Step 1: Split the data into training sets and test sets.

Step 2: Build a decision tree: build multiple decision trees using randomly selected feature subsets and randomly selected training data subsets.

Step 3: Combined decision tree: combine the output of the decision tree to make a single prediction, and take the average

value of the prediction.

Step 4: Evaluation model: use the average absolute error to evaluate the performance of the model on the test set. Adjust the super parameters of the model, such as the number of decision trees and the number of features in each tree, to optimize the performance.

Step 5: After the model is trained and adjusted, use it to predict the new data.

### 1.3 Model Establishment of question3

Step 1: Grading the difficulty in the Excel table[6];

Step 2: Establish a classification model based on logical regression algorithm, decision tree algorithm and XGBOOST algorithm for mechanical learning regression, and extract parameters MAPE, ACC, REC and F-score;

Based on the above three algorithms, the learning curve of three algorithm parameters F-score on progress is made;

Step 3: Establish a classification model based on CNN convolution neural network to do deep learning regression and extract F-score value;

Step 4: Compare the fitting effects of the above algorithms and choose the best model to predict the given word[7].

## References:


[1]Haihong Fan Application of SVM classification algorithm based on convolution neural network in image classification [J]. Science and Technology Bulletin, 2022,38 (08): 24- 28.DOI: 10. 13774/j.cnki.kjtb.2022.08.005.

[2]Xiaotong Hu, Chen Cheng. Time series prediction based on multi-dimensional and cross- scale LSTM model [J]. Computer Engineering and Design, 2023,44 (02): 440-446. DOI: 10. 16208/j.issn1000-7024.2023.02.017.

[3]Shishi Dong, Zhexue Huang. Analysis of random forest theory [J]. Integrated Technology, 2013 (1): 1-7.

[4] Lei Liu. Research on Classification of breast cancer Diagnostic Data Based on Logical Regression Algorithm [J]. Software Engineering, 2018,21 (02): 21-23+17. DOI: 10. 19644/j.cnki.issn2096- 1472.2018.02.007.

[5]Xun Wang, Jia Qiao, Yanping Yu. Risk assessment of gas pipeline based on decision tree classification algorithm [J]. Gas and Heat, 2022,42 (10): 41-43+46. DOI:10. 13608/j.cnki.1000-4416.2022.01.015.

[6] Jiqing Yan, Zhiyuan Shen, Jing Lv, et al. Automatic classification of bidding documents based on XGBoost and text focus model [J]. Journal of Wuhan University (Engineering Edition), 2022,55 (03): 310-318. DOI: 10. 14188/j.1671-8844.2022-03-013.

[7]Chicco D, Warrens M J, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation[J]. PeerJ Computer Science, 2021, 7: e623.