Research on Wordle based on ARIMA and GBDT models

Zhongnan Nie, Hang Wu *, Yongqiang Song

Xi'an University of Posts and Telecommunications, School of Electronic Engineering, Xi'an, China, 710119

* Corresponding Author Email: 13772658858@163.com

Abstract. Wordle is a very popular word guessing at the moment, where players have a total of six chances to guess a five-letter word. Based on the AMIRA time series model, this paper analyzes the data and predicts the changes in the number of people who submit reports in the future, and then analyzes the word attributes that affect the accuracy of word guessing according to a large number of actual data, and establishes a Gradient Boosting Decision Tree (GBDT) model according to the determined word attributes to predict the correct guess rate of any word, and finally predicts and analyzes the word "EERIE".

Keywords: AMIRA time series model, Gradient Boosting, Decision Tree (GBDT) model.

1. Introduction

Wordle is a very popular word guessing at the moment,where players have a total of six opportunities to guess a five-letter word. Each time you guess a word,you will be given some hints based on how well the guessed word matches the correct answer. There are three possible hints for each letter: gray,yellow, and green. Gray means that the current letter does not appear in the final answer, yellow means that the current letter appears in the final answer but is not in the correct position, and green means that the current letter appears in the final answer and the position is also correct. The Wordle is about using hints to narrow down the scope and find the final answer by guessing again and again. Although Wordle has simple rules, it actually comes with many properties that promote propagation. [1]

2. AMIRA time series

2.1. Model Building

In order to model non-stationary time series, some scholars have proposed the differential autoregressive moving average model (ARIMA), that is, the non-stationary time series is converted into a stationary time series through differential operation, and then the ARMA model is established for the differential series. In order to improve the fitting and prediction effect of the model, it is necessary to improve the heteroskedasticity of the residual series of the model [2].

Autoregressive Moving Average - The ARMA model is a commonly used model for dealing with stationary time series in time series analysis, which is realized by time series data through two processes: autoregressive (AR) and moving average (MA). An existing set of stationary time series {xt}, with a sequence of observations {xt}, and a model that satisfies the following structure is called an autoregressive moving average model ARMA(p, q). [3]

$$\begin{gathered} x_t = \phi_0 + \sum_{i=1}^{p} \phi_i x_{t-i} \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \\ \phi_p \neq 0, \theta_q \neq 0 \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma^2 \\ E(\varepsilon_t \varepsilon_t) = 0, s \neq t \\ E(x_s \varepsilon_t) = 0, \forall s < t \end{gathered} \tag{1}$$

The ARMA model is only suitable for stationary time series data, and for stationary data, d-order difference operations are required first. If the differentially processed time series is suitable for the

ARMA model, then the original time series data is said to be suitable for the ARIMA(p,d,q) model. [4]

This model has the following advantages:

1) There is no need to analyze the past and future connections of things with the help of the causal relationship of the development of the thing;

2) It has a small amount of information, and can have good prediction results for trend, random and correlated data;

3) The structure of the model is relatively simple, and the prediction principle is easy to understand, requiring only endogenous variables and not some other types of exogenous variables.

2.2. Results

The autocorrelation and partial correlation analysis of the differentially processed data were carried out, and p and q were preliminarily determined by the function images.

Final Differential Data Autocorrelation Plot (ACF), as shown in Fig 1.

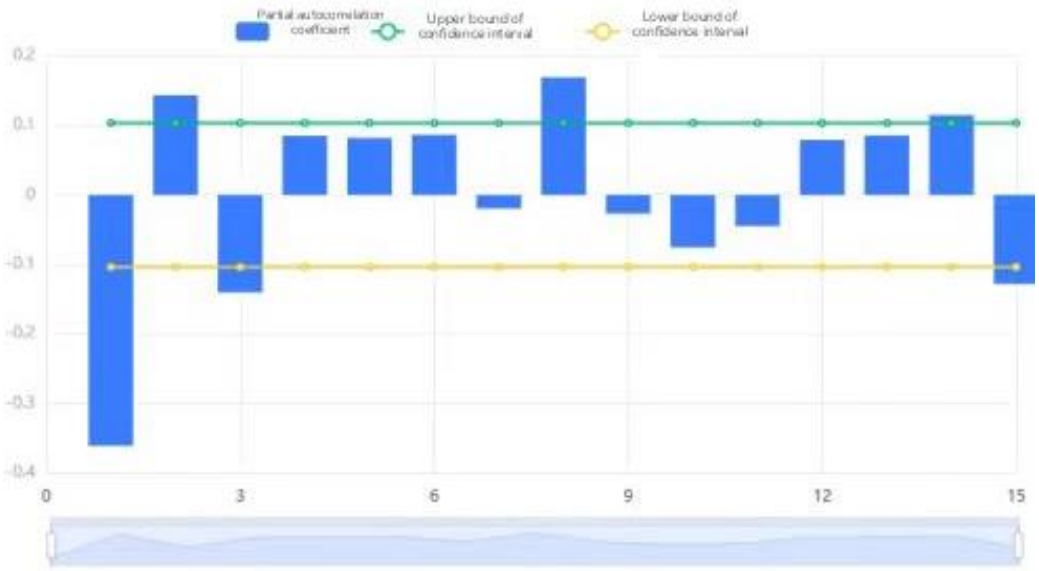

Figure 1. Differential Data Autocorrelation Graph (ACF)

The figure above illustrates the autocorrelation plot (ACF), including coefficients, upper and lower confidence bounds.

The horizontal axis represents the number of delays, and the vertical axis represents the autocorrelation coefficient. The autocorrelation (ACF) diagram is truncated at the q order, and the partial autocorrelation (PACF) diagram is tailed, and the ARMA model can be simplified to the MA(q) model.

Figure 2. Partial Autocorrelation Graph (PACF) of Differential Data

The figure above illustrates a partial autocorrelation plot (PACF) with coefficients, upper and lower confidence bounds. The partial autocorrelation (PACF) diagram is truncated at the p-order, and the autocorrelation (ACF) diagram is tailed, and the ARMA model can be simplified to the AR (P) model. If both the autocorrelation and partial autocorrelation plots are tailed, the most significant order (minimum value) in the PACF and ACF graphs can be combined as the p and q values.

The optimal parameters were found by graph analysis, and the final model result was ARIMA model (1, 1, 0).

2.3. Residual white noise test

For the residual sequence obtained by fitting the model, it is necessary to perform a white noise test to determine the correctness of the functional relationship. If the results show that the residual sequence is white noise, the fitting is considered appropriate to obtain the model results, but if not, it may be due to insufficient information extraction from the sequence data, which makes the sequence not appear normally distributed and needs to be retested. [5] The test results are shown in the Table 1:

Table 1. AMIMA model (1, 1, 0) test table

item	sign	value
	Df Residuals	356
Sample size	N	359
Q statistic	Q6 (P)	0.003 (0.958)
	Q12 (P)	24.316 (0.000***)
	Q18 (P)	52.473 (0.000***)
	Q24 (P)	80.244 (0.000***)
	Q30 (P)	97.19 (0.000***)
Information guidelines	AIC	7749.352
	BIC	7760.994
Goodness of fit	R^2	0.982
Note: ***, **, and * represent the significance levels of 1%, 5%, and 10%, respectively		

R^2 in the graph represents the degree of fitting of the time series, and the closer to 1, the better. Information criterion AIC and BIC values are used for multiple model comparisons (lower is better)

The probability of the Q test statistic is less than the significance level of 0.05), that is, the null hypothesis is accepted, the residual sequence of the model is a white noise sequence, and the Q test

passes,as shown in Table 2.

Table 2. Time series forecast table

predicted value	
Order (time)	predict the outcome
51	11971.34167911301
52	11803.045588443427
53	11634.749497773844
54	11466.453407104262
55	11298.15731643468
56	11129.861225765097
57	10961.565135095514
58	10793.269044425931
59	10624.972953756349
60	10456.676863086766

Chart Description:

The table above shows the data forecasts for the last 60 periods of the time series model (Due to space constraints, only the last ten pieces of data were retained), where the forecast value for period 60 is the forecast value as of March 1, 2023.

3. Gradient boosting decision tree

3.1. Model building

GBDT is an iterative decision tree algorithm based on boosting ensemble learning with continuous fitting residuals, which is a process in which the boosting tree uses the additive model and the forward step-by-step algorithm to achieve learning optimization. It is suitable for dense data, can be calculated in parallel, with fast calculation speed and strong generalization ability. [6] Proposed in 1999 by Jerome H. Friedman, a professor of statistics at Stanford, the algorithm is centered on influencing the structure of the evaluator by continuously fitting the residuals. The GBDT algorithm is composed of a plurality of decision trees, and when the model predicts, it will first give an initial value to the sample prediction value, and then traverse each decision tree, and each tree will adjust and correct the prediction value, and the final result is to accumulate the results of each decision tree to obtain the final prediction result. [7] The loss function is used to evaluate the performance of the model, and it is believed that the smaller the loss function, the better the performance. By letting the loss function descend in the direction of the gradient, you can continuously improve the performance of the model. [8] In order to maximize the reduction of each iteration of the loss function and accelerate the local or global convergence of the model, the model is continuously optimized, and the fitting direction of the residuals is the negative gradient direction of the loss function of the previous weak evaluator, and the output of the weak evaluator gradually approaches the true value after each round of fitting [9].

The GBDT algorithm trains the model on the input features, and optimizes the model under the minimization of the loss function. The input training dataset D={(x_i, y_i), (x_2, y_2),……, (x_n, y_n)}, where $x_i \in x \in R_n$, where x_i is the input feature variable, i.e., word-related attribute parameters, which belong to the sample space R_n. $y_i \in Y \in R$, where y_i is the target value of the ith training sample. If the loss function is defined as L(y, f(x)), the specific iteration process of the GBDT algorithm is as follows. [10]

(1). Initialize the 1st regression tree $f_0(x)$.

$$f_0(x) = \arg\min_c \sum_{i=0}^{n} L(y_i, c) \tag{2}$$

During the ceremony: c is the constant when the loss function is minimized, and $f_0(x)$ is the constant value of the initial C ART regression tree, that is, c; n is the number of samples.

(2). Iterative processing. m =1,2,..., M , where M is the number of regression trees, i.e., the number of iterations.

Firstly, for sample x_i , the negative gradient value of the loss function of the current model is fitted as the residual, and the loss function is minimized.

$$\backslash r_{m,i} = -\left[\frac{\partial L\left(y_i, f\left(x_i\right)\right)}{\partial f\left(x_i\right)}\right]_{f(x)=f_{m-1}(x)} \tag{3}$$

During the ceremony: $r_{m,j}$ is the residual fitting of sample x_i by the mth regression tree, and $f_{m-1}(x)$ is the predicted value of the mth regression tree.(x_i , y_i)

And then, substituting (x_i , y_i) into $r_{m,j}$, fitting the mth round of training dataset, that is, the mth CART regression tree, and obtaining the leaf node ensemble region $r_{m,j}$.

Then, the value of the node j region is calculated by Eq. (8) to minimize the loss function.

$$c_{m,j} = \arg\min_{c} \sum_{x_i \in R_{m,j}} L\left(y_i, f_{m-1}\left(x_i\right) + c\right) \tag{4}$$

During the ceremony: $c_{m,j}$ is the best fitting value for the j node regions of the mth regression tree under the minimization of the loss function.

Finally, the prediction results are updated by Eq. (9) to obtain the predicted value $f_m(x_i)$. of the mth regression tree.

$$f_m\left(x_i\right) = f_{(m-1)}\left(x_i\right) + \sum_{j=1}^{J} c_{m,j}\theta, x \in R_{m,j} \tag{5}$$

(3). The $c_{m,j}$ values obtained by all weak evaluators were summed together in the same leaf node region, and the final regression prediction value $F(x)$ was obtained

$$F\left(x\right) = f_M\left(x_i\right) = f_0\left(x\right) + \sum_{m=1}^{M}\sum_{j=1}^{J} c_{m,j}\theta, x \in R_{mj} \tag{6}$$

3.2. Result

The feature importance is calculated by feeding the data into the model, as shown in Fig 3.

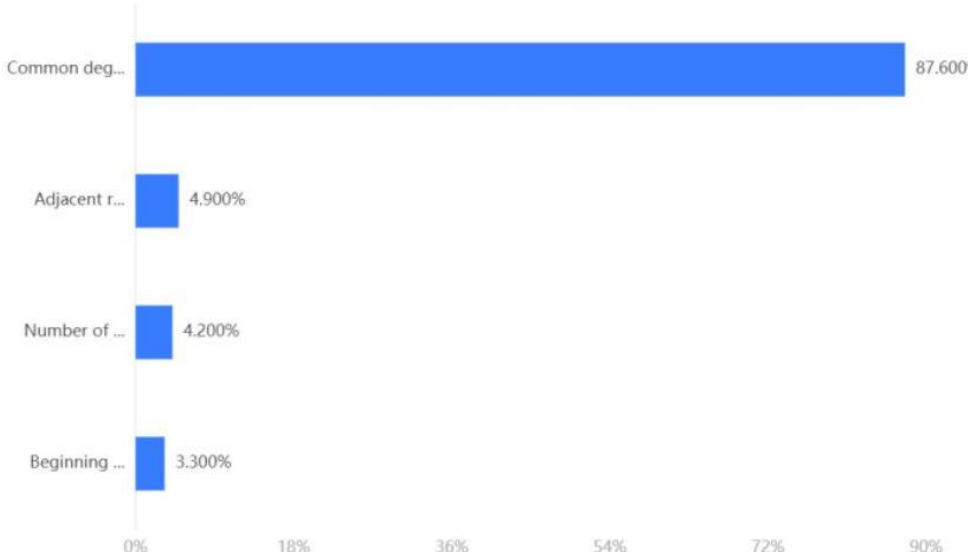

Figure 3. The characteristic importance of each word attribute

The upper column chart or table shows the proportion of importance of each feature (independent variable).

The Gradient Boosting Tree (GBDT) classification model was established by training set data, and the established Gradient Boosting Tree (GBDT) classification model was applied to the training and test data to obtain the model evaluation results, as shown in table 3.

Table 3. GBDT model evaluation results table

	Accuracy	Summons rate	Precision	F1
Training set	0.993	0.993	0.993	0.993
Test set	0.625	0.625	0.704	0.643

After that, the word "EERIE" is input into the model for prediction, and the prediction result is as follows table.4:

Table 4. Prediction results table for the word "EERIE".

1try	2tries	3tries	4tries	5tries	6tries	7or more tries	Number of repeated letters	Beginning of vowel	Common degree	Adjacent repeating letters
0	7	7	27	38	14	2	15351	1	2	1

As can be seen from the table above, adding the percentage of the seven predictions to the number of guesses gives 95, which is very close to 100, and the prediction is credible. Moreover, the prediction model training set performs well, with an accuracy rate of more than 90%, so the model is accurate and credible.

4. Conclusion

Based on the analysis of the AMIRA time series model, based on the analysis of the word attributes that affect the accuracy of word guessing, it was found that none of the four attributes selected had a significant correlation with the percentage of the Hardmode score, but increased with the increase of the date, that is, any attribute of a word would not affect the percentage of its Hardmode score. According to the determined word attributes, a Gradient Boosting Decision Tree (GBDT) model was established to predict the correct guessing rate of any word, and it was found that the difficulty of EERIE was moderate.

References

[1] NetEase: Anagram Wordle is now available on The New York Times' Crossword app. https: //www.163.com/dy/article/HFK1MKQ90511BLFD.html.

[2] ZHANG Wenhua. Improvement and simulation of ARIMA model for time series forecasting [J]. Information & Computer (Theoretical Edition), 2021, 33 (05): 53 - 56.)

[3] Wang Yan. Applied time series analysis [M]. Beijing: Chinese Renmin University Press 2016.

[4] Yang Yang, Tian Dingsheng, Zhang Baoan, et al. Research on urban economy and population forecasting based on ARIMA model [J]. Integrated Transportation, 2023, 45 (11): 79 - 85+97.)

[5] Chen Kexiu, Liu Juan. Sales forecast of new energy vehicles based on ARIMA model [J]. Modern Industrial Economy and Informatization, 2022, 12 (03).

[6] Xu Zixi, Tang You, Zhong Wenyu, et al. Research on corn yield prediction model in Jilin Province based on GBDT algorithm [J]. Smart Agriculture Guide, 2024.

[7] Research on big data risk control model based on GBDT algorithm [J]. Journal of Zhengzhou Institute of Aeronautical Industry Management, 2020.

[8] ZHANG Yaofang, CHEN Jian. Short-term prediction model of highway traffic flow by vehicle type based on GBDT algorithm [J]. Highway, 2022.

[9] Ke Guolin. Research on Gradient Boosted Decision Tree (GBDT) Parallel Learning Algorithm [D]. Xiamen: Xiamen University, 2016.

[10] ZHA Wei, DONG Yanwu, JIANG Zhouhua, et al. Prediction model of manganese content at the end point of vacuum self-consuming ingots based on GBDT algorithm [J]. Metallurgical China Society, 2022.