Prediction of Wordle based on ARIMA and PLS model

Ziyu Wang*

School of Information Science and Engineering (ISE), Shandong University, Qingdao, China, 266237

*Corresponding author: 202000120207@mail.sdu.edu.cn

Abstract. Nowadays, the New York Times Wordle game gains much popularity in the United States. In this paper, PLS regression analysis is applied to find 5 potential factors to interpret the independent variable information and obtain the prediction distribution result. Extra features such as time series have stationarity and nouns make up the vast majority of the total are investigated through data mining. It is also found that the fluctuating trend of the number of players is consistent with the infectious disease model. The paper gives a prediction of the population of players that is worth reading for reference and a simple classification method of difficulty of words. The model inferred in the paper is expected to help puzzle editors to give a word of proper difficulty.

Keywords: ARIMA Model, PLS Regression, Natural Language Processing, Time series forecastingARIMA Model, PLS Regression, Natural Language Processing, Time series forecasting.

1. Introduction

Wordle is a popular puzzle given by the New York Times daily [1]. In this game, players are asked to guess a five-letter word in six tries or less by submitting a recognizable word. Feedback will be given by different colors of the tiles after submitting the word. Players can play in both regular mode or Hard Mode. In Hard Mode, the already-found correct letters in the word are required to be used in subsequent guesses. The model described in the subject requires a concrete measurement. It will be used to classify the difficulty of words and verify a certain word given on March 1, 2023 to make the model comparable.

K Koh et al. introduced ManiWordle, a Wordle-based visualization tool that revamps interactions with the layout by supporting custom manipulations [2]. M Bonthron proposed combining a rank one approximation and latent semantic indexing to a matrix representing the list of all possible solutions [3]. Scholars at home and abroad have applied stepwise regression models in different fields. Chen et al. used the stepwise multiple regression statistical method calculate the relation between algal chlorophyll a (Chla), total algal biomass (TB), Microcystis biomass (MB) and these environmental factors [4]. Zhang et al. multi-objective optimization of automobile safety based on stepwise regression model [5].

We developed models to predict Wordle results, considering the day people play the game and the difficulty of words. We used PLS regression analysis to obtain the prediction distribution result. We also predicted the expected tries that the word EERIE would take to guess it out correctly. We mined the data and investigated extra features. The flow chart of the model is shown in Figure 1.

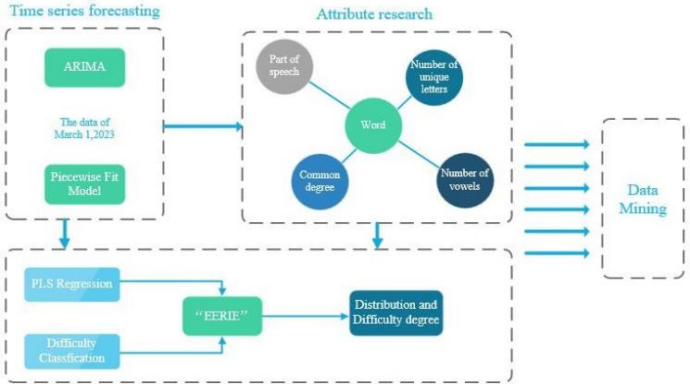


Figure 1. Flow chart

2. Preliminary

2.1. Assumptions

Time series is stationary in a certain range, which means there will be no rapid in- crease or decrease caused by accidental and irresistible factors. The data loss caused by PLS can be neglected. The players who report their results are equally well-educated.

2.2. ARIMA model

The ARIMA *(p, d, q)* model is known as the Auto regressive Integrated Moving Average Model [6], which is based on AR and MA models, and is often used to mine time series. Usually *p* is the autoregressive term. *d* is the number of differences when the time series is stationary. *q* is the number of moving average items. The formula is given in Equation 1 [7].

$$y = \mu + \sum_{i=1}^{p} \gamma iyi + \epsilon i + \sum_{i=1}^{p} \theta i \epsilon t - i \tag{1}$$

Where y_t is the current value. μ is the constant term. γ_i is the auto correlation coefficient. ε_t is the error term. θ_i is the coefficient of the error term.

2.3. Partial least squares (PLS) regression

Partial least squares (PLS) regression is a statistical method related to principal component analysis [8]. Unlike principal component analysis, which finds the hyperplane of maximum variance between the response variable and the independent variable, PLS finds a linear regression model by projecting the predictor and observed variables into a new space, respectively. PLS performs well when predictors are highly col-linear, or when there are more predictors than observations and ordinary least squares regression produces high standard errors of coefficients or fails entirely. Unlike multiple regression, PLS does not assume that the predictors are fixed. This means that the measurement of the predictors may be subject to error, making the measurement of PLS more uncertain.

The basic steps of PLS are as follows.

1.Confirm the number of principal components extracted, and confirm through the cross-validity and projection importance VIP analysis table.

2.Based on the number of principal components in the first step, regression analysis is performed.

The general multivariate underlying model of partial least squares as two matrix equations given in the Equation 2[9]. Among them, *X* is the prediction matrix, *Y* is the response matrix, *T* and *U* are the projections of *X* and *Y* respectively, *P* and *Q* are the orthogonal loading matrices, *E* and *T* are the error terms, which are independent and identically distributed normal distribution random variables. Decompose *X* and *Y* to maximize the covariance between *T* and *U*.

$$X = UQ^T + F \tag{2}$$

3. Experiments

3.1. The process of ARIMA model

Using the ARIMA model to predict time series requires the data to be stable. If the data is unstable, it is impossible to capture the regulation. Just like stock data, which cannot be predicted by ARIMA because that stock data is unstable and often fluctuates under the influence of policies and news. Figure 2 shows the process of the ARIMA model.

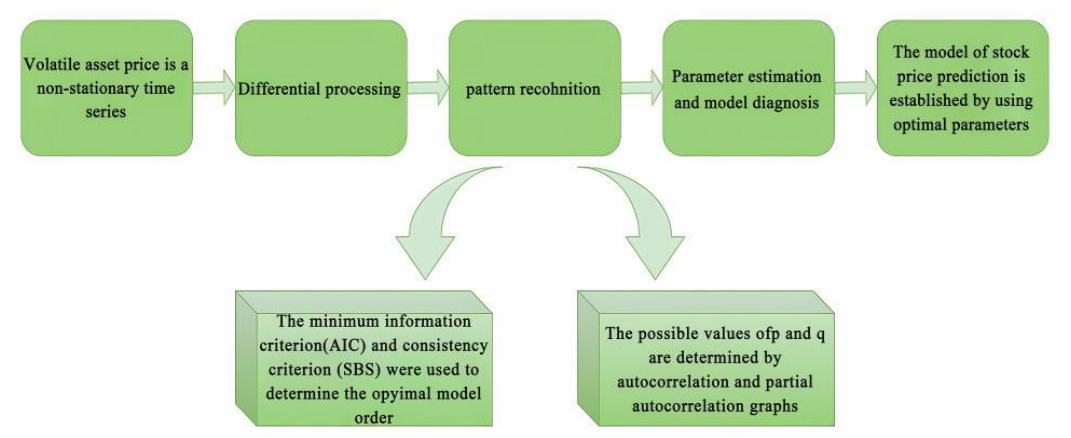
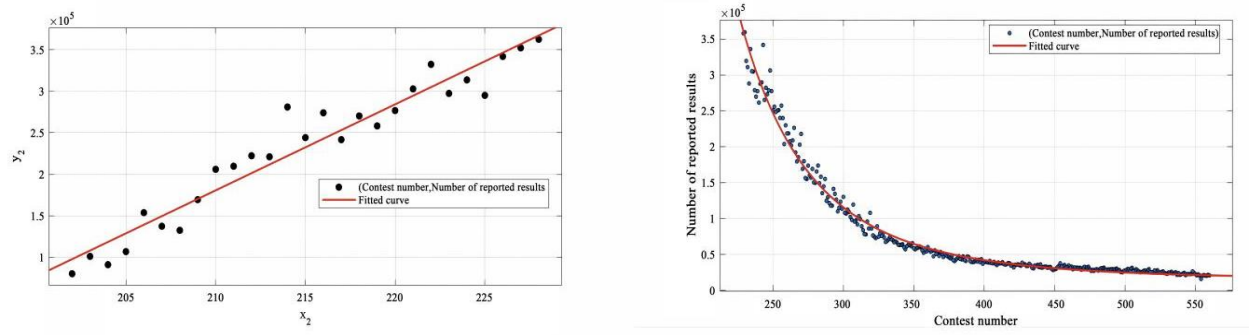


Figure 2. The flow chat of ARIMA

The Augmented DickeyFuller test[10] is often used to detect whether the time series satifies the stationary characteristics. If the calculated P value is less than 0.05, it does not satisfy the non-stationary null hypothesis, which proves that the time series is stationary. If the time series is not stationary, it needs to be differentially calculated to obtain a stationary series, and then use the ARIMA model. Figure 3(a) and Figure 3(b) shows the comparison between the processed data and unprocessed data.



(a) Stage1:202−233 (b) Stage2:234−560

Figure 3. Data processed vs Data unprocessed

3.2. Results

According to the test results in Table 1, When the difference is 0 order, the significance *P* value is 0.002***, the level is significant, rejecting the null hypothesis, and the series is a stationary time series. When the difference is 1st order, the significant *P* value is 0.001***, and the level is significant, rejecting the null hypothesis, and the series is a stationary time series. When the difference is 2nd order, the significance *P* value is 0.000***, the level is significant, the null hypothesis is rejected, and the series is a stationary time series.

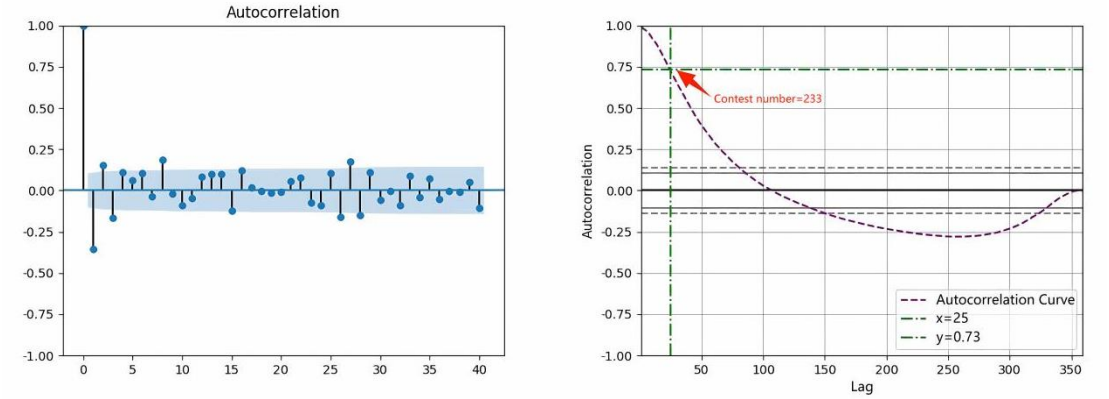
Based on the AIC information criterion, the automatic search for the optimal model is the ARIMA model (1,1,0) and the goodness-of-fit R_2 of the model is 0.982. The model performs well and basically meets the requirements.

Through the model ARIMA(1,1,0), the prediction of Number of reported results will reach at 10.454.17615061 on March 1,2023.

Table.1. The ADF test table of Number of reported results

Differential Order	*T* value	*P* value	AIC
0	-3.865	0.002	7185.006
1	-4.027	0.001	7176.733
2	-10.621	0.000	7156.157

By performing the first-order difference operation and autocorrelation calculation on the original time series[9], we obtained the following two figures, in which Figure 4a is the first-order difference sequence, which is as stationary as the original sequence, and Figure 4b is the autocorrelation curve, It can be seen that for the sequence as a whole, there is a strong correlation before the critical point (233,0.73), and the subsequent correlation will gradually decrease to the range of weak correlation.



(a) First difference sequence (b) Autocorrelation Curve

Figure 4. The first−order difference operation and autocorrelation calculation

3.3. PLS Regression and Difficulty Classification

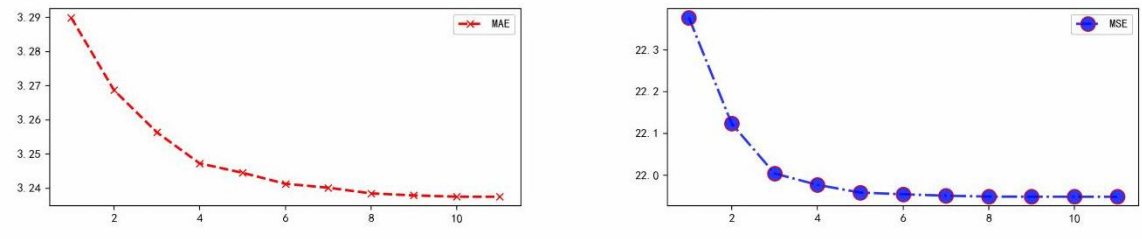
According to the steps of PLS regression, by extracting potential factors, it is found that when there are 5 potential factors, it can explain 80% of the information of the independent variables, and based on this, the maximum number of principal components of the parameters is determined. VIP (Cumulative Projection Importance) means that when the number of components is different, the interpretation importance of X to Y can also be used to refer to the maximum number of principal components. Among them, for in- dependent variables with a large VIP (greater than 1), it plays a relatively greater role in explaining potential factors (and thus explaining dependent variables).

Table 2 shows the factor variance explanation.

Table.2. The factor variance explanation

Potential factor	X variance	Cumulative χ variance	Y variance	Cumulative γ variance	Adjusted R^2
1	0.303	0.303	0.171	0.171	0.169
2	0.198	0.501	0.01	0.181	0.177
3	0.176	0.676	0.004	0.185	0.178
4	0.163	0.84	0.001	0.186	0.177
5	0.16	1	0.001	0.186	0.175

At the same time, in order to find the optimal number of latent factors, the latent factors are used as independent variables to calculate the mean absolute error and mean squared error of PLS, and obtain the changing trend, as shown in the figure 5.



(a) MAE (b) MSE

Figure 5. MAE and MSE Loss

By using PLS Regression input the characteristics of "eerie", a set of percentages of times will be output. We will get [0.29612914,5.42441962,18.17524991,30.03906772, 27.13005711,14.57541882, 4.35965767], substitute it into the formula 9. Then get output 439.44739116510743. So "eerie" is a word of moderate difficulty.

But in fact, the impact of this difficulty should take into account the main participants in the game and the commonness of vocabulary. The well-educated crowd will have a higher accuracy rate than others, or use fewer times to get the right answer.

3.4. Analysis

By using the fitting model and ARIMA model, the prediction value of March,1,2023 will be 14,740 or 10,454.176. Because the two models have their own weaknesses, the final prediction value interval should be [10, 454, 14, 740].

The study indicates that word commonness and the number of non-repeating letters in words were negatively correlated with the number of attempts in player difficulty mode and the former correlated more, while the number of vowel letters in words did not correlate with the number of studies.

By using PLS regression and weighing model, the difficulty is quantified and divided into three levels, corresponding to three difficulty levels of easy, medium and difficult. According to the model, the difficulty of the word EERIE is quantified as about 439.4. Therefore, EERIE is a word of moderate difficulty. The difficulty classification model applies a simple method but achieved a good indication of the difficulty of guessing the word. The prediction correspond the relationship between number of tries and difficulty of a certain word.

4. Conclusions

This paper predicted the results of Wordle game by two model, PLS regression and ARIMA model. PLS regression analysis is applied to find 5 potential factors to interpret the independent variable information and obtain the prediction distribution result. Extra features such as time series have stationarity and nouns make up the vast majority of the total are investigated through data mining. Through the approaches above, we have successfully found out the best solutions and strategy of the game. Besides, there are still more usage for ARIMA and PLS regression model. We tend to ultimate them for other predictive problems.

References

[1] Anderson B J, Meyer J G. Finding the optimal human strategy for Wordle using maximum correct letter probabilities and reinforcement learning [J]. 2022.

[2] Pang B, Feng S, Yin Y. Development of a dimensionless rod-bundle CHF correlation based on stepwise regression method, Part II: determination of correction factors accounting for the effect of unheated guide tube and non-uniform axial heat flux profile [J]. Annals of Nuclear Energy, 2022, 169:108899-.

[3] Sonkar R, Dhekne P Y, Londhe N D. Prediction of Peak Particle Velocity of Blast-induced Ground Vibrations using Boosted Regression Trees Authored [J]. Journal of Mines, Metals & Fuels, 2022.

[4] Borght, Koen Van Der, et al. Cross-validated stepwise regression for identification of novel non-nucleoside reverse transcriptase inhibitor resistance associated mutations. BMC Bioinformatics, 12,1 (2011-10-03) 12.1(2011): 386.

[5] A O S, B S J. Fast stepwise regression based on multidimensional indexes [J]. 2021.

[6] Agustín Maravall. A CLASS OF DIAGNOSTICS IN THE ARIMA-MODEL-BASED DECOMPOSITION OF A TIME SERIES [J]. 2022.

[7] Ferranti J S, Chaves A P, Cruz M, et al. APPLYING ARIMA MODEL IN THE CONTEXT OF AGRICULTURE AND INDUSTRIES [J]. 2022.

[8] Carvalho N R, Bazana L C G, Fuentefria A M, et al. Digital images coupled to PLS regression for pH prediction in sterile culture medium [J]. Biomedical Signal Processing and Control, 2022, 73: 103435-.

[9] Wold S, Sj-Str-M M, Eriksson L.PLS-regression: a basic tool of chemometrics [J]. Chemometrics & Intelligent Laboratory Systems, 2001, 58(2): 109-130.

[10] Joppen L. Flomatic announces AIS compliant MAXIAIR model [J]. Valve World, 2022(6): 27.