Analysis of the Wordle data based on a time-series prediction model

Yan Zhu*

Department of Information management and information systems, Capital University of Economics and Business, Beijing, China

*Corresponding author: yanzh2022@126.com

Abstract. Wordle Is a popular word game daily provided by the New York Times. In this paper, we make a further data analysis on Wordle by using the time-series model. First, a time series model was built to predict and test the results, and the correlation of word properties with the degree of difficulty was analyzed. Second, we build multiple multivariate regression models, take the percentages of 1,2,3,4,5,6, and X as output variables, find the relationship between word attributes and the percentage of each attempt and make predictions. Finally, words were classified by difficulty based on the extracted attributes and eerie by the resulting model and evaluated for accuracy.

Keywords: Time series prediction, Machine learning, decision tree, Regression decision tree classification, K-means clustering.

1. Introduction

A simple word game is the latest social media and pop culture phenomenon: Wordle. The task is to guess a five-letter word. You have had six chances. After each guess, the block changes the color which letters are not in the word (gray), which letters are in the word but wrong (yellow), and which word is correct and which is correct (green). The game was very popular: According to The New, as of January, more than 300,000 people played the game in the New York Times every day. The popularity may sound confusing, but there are some small details that make everyone go crazy. There is only one puzzle a day: this creates a degree of risk. You only have one chance first. If you screw it up, you have to wait until tomorrow to get a whole new puzzle.

2. Time-series prediction model and correlation analysis

2.1. Time-series prediction model

A model was created to predict the number of results reported on March 1,2023, and to explore the relationship between time and the number of reports per day, using a time series prediction model.

A time series is a collection of data collected at different points in time with constant time intervals [1], which are analyzed to understand long-term trends and to predict the future. Time series prediction method is a regression prediction method, which belongs to quantitative prediction. Its basic principle is: on the one hand, it acknowledges the continuity of the development of things, uses the past time series data for statistical analysis, and speculates the development trend of things; On the other hand, the randomness due to the influence of accidental factors is fully taken into account. In order to eliminate the influence of random fluctuations, historical data is used for statistical analysis, and the data is properly processed for trend prediction.

The ARIMA (p, d, q) model

$$y_t = \mu + \sum_{i=1}^{p} r_i y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \quad (1)$$

Here: y_t is the present value, namely the current value; μ is a constant term. p is the order of the index; y_i is the autocorrelation coefficient. ε_t is the error.

In this paper, the time series is preprocessed to test the stationarity and randomness of the observed series. After inspection, the sequence is a stationary non-white noise sequence, identified as an

ARIMA model, with the goodness of fit R^2 of 0.982, and the model performed well. In this paper, the date and the number of daily reported results are selected, and the time series model is used to predict the 60 days after December 31,2022. From the perspective of forecast results, after the number of reported results reached the peak on February 5,2022, it has shown a downward trend and finally leveled out (as shown in the figure). Therefore, this paper predicts the number of reported results on March 1,2023 to be 10342.

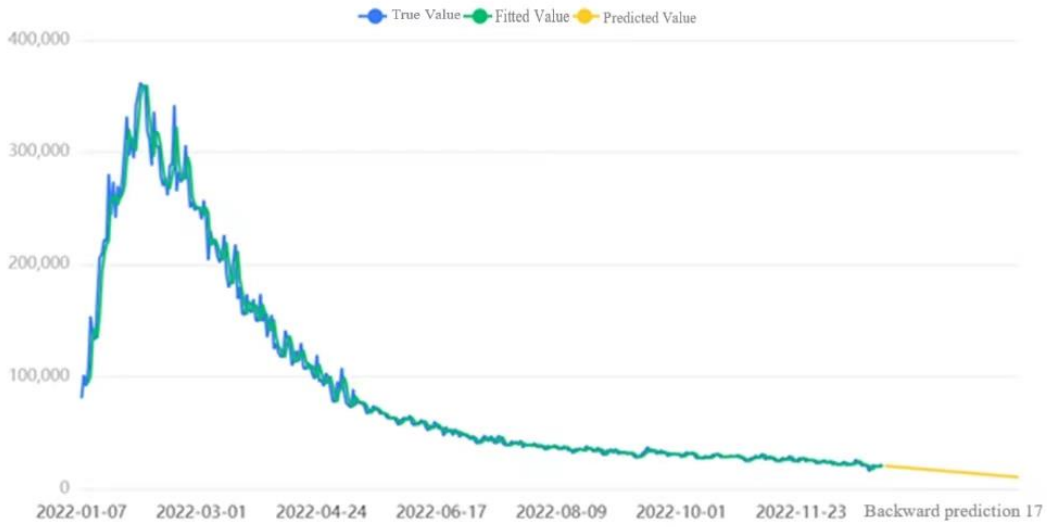

Fig. 1 Forecast results

2.2. Inquiry analysis of word attributes

Through common sense cognition such as daily experience and language habits, the frequency of the word itself, the frequency and location of letters in the word, and the distribution of vowels and consonants can all be used as the characteristic attributes of the words and have an impact on the results. We extracted the five words attributes as shown in the table below:

Table 1. Word attribute list

Word attribute	Symbolic representation
The letter attribute of a word	N
Letter frequency properties of words	F
How common the word is	w-com
Letter information entropy of a word	S
Structural properties of words	DEC

Gray correlation degree analysis based on the word attribute and the difficulty coefficient

2.3. Gray correlation degree analysis based on the word attribute and the difficulty coefficient

Grey Relation Analysis (GRA) is a multi-factor statistical analysis method. Simply put, it is a gray system, to understand the relative strength of a project under the influence of other factors. Because word attributes are different quality indexes, some numbers may be very large and some may be very small, and as a result of different dimensions, they need to be normalization without dimension. Grey correlation analysis formula is:

$$\xi_t(k) = \frac{\min_i \min_k |x_0(k)-x_i(k)| + \rho \cdot \max_i \max_k |x_0(k)-x_i(k)|}{|x_0(k)-x_i(k)| + \rho \cdot \max_i \max_k |x_0(k)-x_i(k)|} \quad (2)$$

The resolution coefficient $\rho \in (0, \infty)$, when $\rho \leq 0.5463$, the resolution is the best, take $\rho = 0.5$.

Grey correlation analysis was carried out for 4 words attributes (S, L-fre, DEC, w-com) and 355 data items, and l was taken as the "reference value "(parent sequence) to study the correlation degree between 4 words attributes (S, L-fre, DEC, w-com and L. The model results are shown in the following table:

Table 2. Relevance results

attribute	degree of relevance	rand
S	0.921	1
l-fre	0.885	2
DEC	0.758	3
w-com	0.682	4

It can be concluded from the table that the word attributes S, l-fre, DEC and w-com are correlated with the difficulty relation.

2.4. Pearson Correlation analysis and the Kendall consistency test

Explore whether any attribute of a word will affect the percentage of report scores played in difficult mode. In data processing, five attributes of a word have been obtained in this paper. Correlation analysis is made between the above attributes and the percentage of report scores played in difficult mode.

The following heat map is obtained:

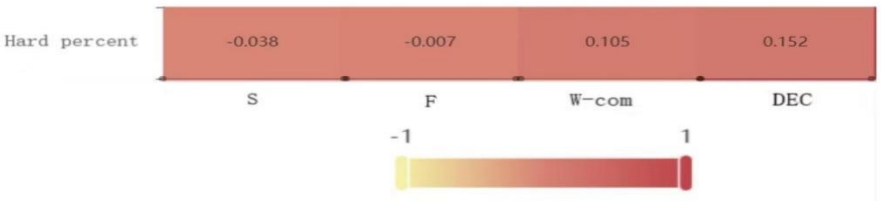

Fig. 1 Pearson correlation coefficient analysis Thermal map analysis

The above thermal maps reflect that the correlation coefficients between the word attributes S, F, W-com and DEC and Hard percent are all low. Therefore, none of the above attributes affect the percentage of report scores played in hard mode.

For letter attribute N of a word, since it contains 26 letters, it is not intuitive to explore the overall correlation between the occurrence times of 26 letters in each word and the correlation performance of Pearson heat map. Kendall consistency test is used here.

The difference between the two is [2]: Pearson compares the correlation between the two items; The Kendall coefficient is used to judge the correlation of the whole population (all data). It is applicable to the data of multiple columns of relevant grade data, which can be k raters to evaluate (N) objects, or the same person has evaluated N objects for k times. By obtaining Kendall harmony coefficient, we can choose good works or good graders objectively.

Test results it is concluded that the Kendall coefficient = 0.182, the significance of the overall data P value is 0.000***, level is significant, reject the null hypothesis, so the data consistency, at the same time the Kendall harmonious coefficient of model W value of 0.182, so the consistency of the degree of correlation is very low. It is concluded that the word attribute N does not affect the percentage of report scores played in hard mode.

3. Establishment and interpretation of multiple regression model

In data preprocessing, this paper has extracted The Times of a-z letters appearing in each word, taken these 356 pieces of data as the data set of input variables, and the percentage of 1, 2, 3, 4, 5, 6 and X as the output variable, in order to find the relationship between word attributes and the percentage of attempts and make predictions. This paper will try to establish multiple regression models in order to obtain the best adaptive model. They are respectively Neural network prediction [3], Linear regression [4], Decision tree regression, Random forest regression, LGBM [5].

Divide the test set and training set. Because machine learning requires the use of a large number of task-related data sets to train the model; The model is iteratively trained by the errors of the model on the data set, and a reasonable model fitting the data set is obtained. The amount of data in this paper is small, so there may be some errors.

After the above five regression predictions, the average absolute percentage error MAPE and fitting degree R^2 of 1, 2, 3, 4, 5, 6, X and the training set and test set of the five regression methods were obtained in this paper. MAPE is a percentage value. The smaller the value, the more accurate the model is. The closer the R^2 result is to 1, the higher the accuracy of the model. In this paper, the MAPE values and R^2 of 1-7 attempts were averaged to obtain the MAPE mean and R^2 means of the training and test sets under the five predictions, as shown in the table below.

Table 3. MAPE value of five regression test sets and training sets

	Neural network	Linear regression	Decision tree regression	Random forest regression	LGBM regression
Training set MAPE mean	0.44	0.61	0.00	0.30	0.52
Test set MAPE mean	0.73	0.45	0.49	0.60	0.76
Training set R^2	0.39	0.27	0.99	0.64	0.28
Test set R^2	-0.13	0.04	-1.31	-0.23	-0.33

According to the results of the training set of the model, the MAPE value of the training set is small, indicating that the error of the model is small and the uncertainty is low.

Confidence in the predictions of the model. As can be seen from Table 3, R^2 of training set and test set is small, and R^2 represents the proportion of variance explained by the model. It is a relative metric that you can use to compare with other models trained on the same data. To get an idea of the performance of a model. This value is generally low in the test set of the above model, indicating that the degree of fit is not high. In this paper, it is believed that the R^2 value is small due to the small amount of data, so our confidence in the model prediction has decreased.

4. Establishment and solution of the classification model based on word attributes

4.1. K-means clustering analysis of the coefficient of word difficulty L

4.1.1. Principle and establishment of the K-means clustering model

K-mean clustering algorithm (k-means clustering algorithm) is a cluster analysis algorithm for iterative solution. The step is that the data is divided into K groups, then K objects are randomly selected as the initial cluster center, and then the distance between each object and each seed cluster center is calculated, and each object is assigned to the cluster center closest to it. Cluster centers and the objects assigned to them represent a cluster. For each sample assigned, the cluster center of the cluster is recalculated based on the existing objects in the cluster. This process is repeated until a certain termination condition is met. The termination conditions can be the minimum number of

objects being reassigned to different clusters, the minimum number of clusters center changed again, with the sum of squared error being a local minimum. In this question, the difficulty of all the words is divided into three difficulty levels: "easy", "medium" and "difficult". Meanwhile, the k value is determined according to the "elbow method". The results are shown in the figure:

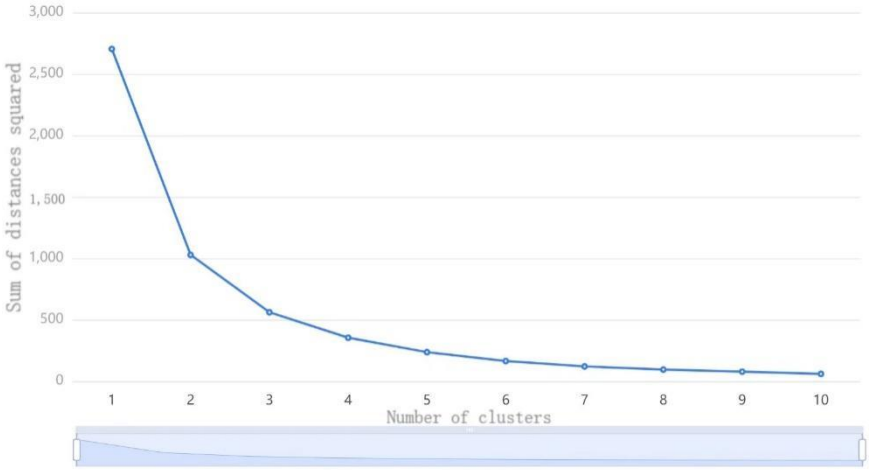

Fig. 2 K-means clustering analysis diagram

The results showed that the number of clustering centers was also 3, which was consistent with the difficulty classification.

4.1.2. Cluster results and the evaluation

According to the model, wordle words are classified and summarized in the following table by difficulty:

Table 4. Word difficulty classification table

Cluster categories	Center value
medium	26.579
difficulty	22.759
easy	29.734

Classification results are divided into three categories, and the percentages of clustering category frequency are shown as follows:

Table 5. Clustering classification frequency proportion table

Cluster categories	**Number of words**	**percentage**
easy	101	28.451%
medium	176	49.577%
difficulty	78	21.972%

The results of the classification of each word are as follows (in part):

Table 6. Word classification table

Contest number	202	203	204	205	206	207
Word	slump	crank	gorge	query	drink	favor
Cluster categories	medium	medium	difficulty	medium	easy	difficulty
Contest number	202	203	204	205	206	207

For the difficulty coefficient L, the significance P value is 0.000***, showing significance at the level. The null hypothesis is rejected, indicating that the difficulty coefficient L has significant differences among the categories divided by cluster analysis, and the clustering effect is good. The coefficient of difficulty calculated according to the percentage of attempts of model eerie in the

second question is 28, which belongs to "easy". The results can be used to test and evaluate the results of attribute - based classification.

4.2. EERIE explores the difficulty classification based on attributes

4.2.1. Principle and establishment of decision tree classification model

Decision tree is a tree structure, in which each internal node represents a test on an attribute, each branch represents a test output, and each leaf node represents a kind of category decision tree model. The logic of decision tree model is to judge each feature of an instance from the root node, and assign the instance to its child nodes according to the judgment results. At this time, each node corresponds to a value of this feature. In this way, instances are judged and allocated recursively until instances are allocated to leaf nodes, which basically follows the simple and intuitive strategy of "divide and rule". The essence of decision tree learning is to summarize a group of classification rules from the training data set, which may have none or many decision rules. In this case, it is necessary to choose a decision tree rule with little contradiction with the data set and good generalization effect.

4.2.2. The result of decision tree classification model

In this paper, 80% of the data is taken as the training set and the rest as the test set. The classification effect is shown in the table below:

Table 7. Decision tree classification table

	Accuracy	Recall	Harmonic averaging of precision and recall
Training set	0.789	0.795	0.79
Test set	0.507	0.527	0.511

Because the data given in this question is less for machine learning, the accuracy of the test set is lower. The prediction results for the eerie are as follows:

Table 8. Decision tree prediction table

Final prediction results	l-fre	w-com	DEC	S
easy	50.7	8350	27	1.53

4.3. Accuracy of model

The characteristics and quantity of training data is the most important factor to determine the performance of a model. In general, machine learning training sets often reach tens of thousands of pieces before good results and small deviations appear. The amount of data in this question is small, so the accuracy is poor, and the prediction of classification in the test set is biased.

According to common sense analysis, eerie has multiple identical vowels, which is easier to guess than other words. Meanwhile, by comparison with K-means clustering results, eerie is divided into "easy" category. Therefore, the accuracy of the model is not bad in the case of less data.

5. Model Evaluation

5.1. Strengths

The model structure is simple and easy to implement.

It is easy to generalize the model through professional software calculation.

5.2. Weaknesses

Being affected by outliers may lead to inaccurate ranges of estimates.

References

[1] Wang Yan. Application of time series analysis [M]. Beijing: China Renmin University Press 2005.

[2] Xu Weichao. Review of correlation coefficients [J]. Journal of Guangdong University of Technology,2012,29(3):12-17.

[3] Zhou Zhihua. Machine Learning [M]. Tsinghua University Press, 2016.

[4] Draper, N.R. and Smith, H. Applied Regression Analysis. Wiley Series in Probability and Statistics. 1998.

[5] Meng Q . LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 2018.