An Exploration of Wordle Game Data and Player Behavioural Habits Based on Time Series and GBDT Models

Yao Lu [1, *, #], Xue Yu [2, #], Yuqing Gu [3, #]

[1] School of Mathematics, Nanjing Audit University, Nanjing, Jiangsu, China, 211815

[2] School of Finance, Nanjing Audit University, Nanjing, Jiangsu, China, 211815

[3] School of Government Audit, Nanjing Audit University, Nanjing, Jiangsu, China, 211815

* Corresponding Author Email: 213080520@stu.nau.edu.cn

[#]These authors contributed equally.

Abstract. The focus of the whole problem is to explore interesting characteristics of Wordle' players' behavior data over the year. In order to explain the daily variation in reported results and to explore whether lexical attributes affect the percentage of scores in the difficult mode, this paper introduced a time series model for specific predictions, finally coming to the conclusion that there is a significant correlation between tries in hard mode and commonness, while there is no significant correlation directly with the number of repeated letters. In addition, based on previous research, a model was developed to predict the distribution of reported results for games at future dates. This paper introduced the prediction model based on Gradient Boosted Tree (GBDT) and carried out specific training, where the training accuracy reached more than 95%. The models are conducive to understanding the development trend of wordle games and players' game habits, and to providing better game suggestions for game developers.

Keywords: Time series; Correlation analysis; GBDT model.

1. Introduction

Wordle is a mind game that features players collecting words and combining words against other players in an English word scramble game, as well as gaining points through leaderboards and advancing to higher levels. With the popularity of wordle games, issues related to its characteristics, game rules, and players' behavior in the game have attracted the attention and research of scholars.

Scholars at home and abroad have carried out a lot of research on studying user object datasets for wordle or related domains. Wormley Alexandra S and Cohen Adam B explored the phenomenon of player cheating in wordle and concluded that players with high literacy levels have a behavioral tendency to cheat to win in the game [1]. Brown Keith examined active learning through the lens of the game Wordle, explaining both the process of active learning and describing the types of research questions that arise when using active learning for material research [2] . Chenlu Li et al. investigated Wordle-based interactive visualization of urban text distribution, and they used correlation analysis and visualization analysis to visualize restaurant data in Shanghai, China, to enhance urban planning strategies [3].

Based on time series with gray prediction and GBDT model, this article predicts the interval of game report results for a future date and the distribution results of attempts for a word, and classify all words by difficulty to study word attributes [4].

2. Time Series Forecasting &Attribute Correlation

2.1. Data Description

(1) Overall descriptive analysis of statistical indicators

Based on the Number of reported results shown in Table 1, the coefficient of variation (CV) is 0.982, which is greater than 0.15. There may be outliers in the current data, and it is recommended to analyze the indicators that are abnormal or have a more prominent performance.

Table 1. Indicator Description Table

Variable Name	Sample size	average	Standard deviation	Median	Variance	Kurtosis	Coefficient of variation	
Number of reported results	359	90918.87	89274.38	44578	7.97E+09	0.893	1.447	0.9819

(2) Analysis of indicators

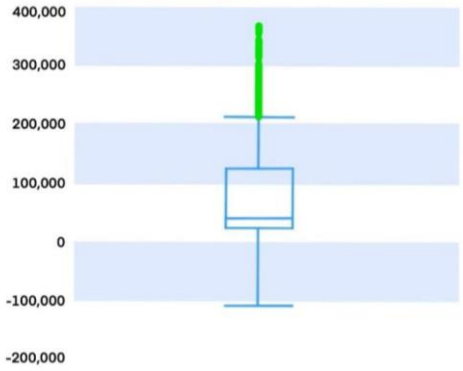

Figure 1. Number of box plots for reporting results

Figure 1 shows the results of the Number of reported results frequency analysis discrete trend analysis in the form of a box line plot. The discrete trend is measured by the difference of the data distribution using statistical indicators such as the extreme value, the minimal value, the 25% quantile, the median, and the 75% quantile. Because this paper gives the real raw data, the abnormal data can be excluded for abnormal data.

(3) Trend Chart

The trend in Figure 2 can be plotted and a slow slowdown from a surge in users can be found.

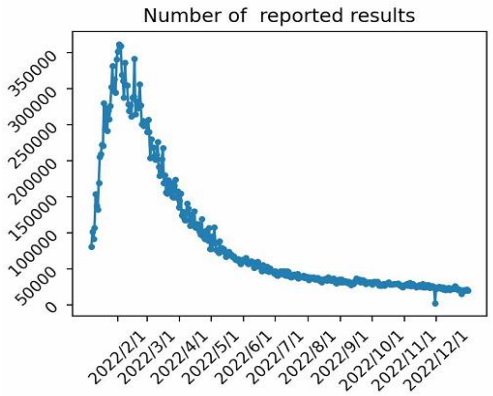

Figure 2. Trend Chart

2.2. Using time series analysis ARIMA

(1) Model parameters table and Time Series Final Forecast Chart

Table 2. Model parameters table

ARIMA model (1,1,0) test table		
Items	Symbols	Results
Sample size	Df Resifuals	356
	N	359
	Q6(P)	0.019(0.891)
	Q12(P)	24.347(0.000***)
Q-statistic	Q18(P)	54.224(0.000***)
	Q24(P)	82.246(0.000***)
	Q30(P)	98.258(0.000***)
Information Guidelines	AIC	7749.064
	BIC	7760.705
Goodness of fit	R^2	0.982

The system automatically finds the optimal parameters based on the AIC criterion, and the model results are the ARIMA model (1, 1, 0) test table (Table 2). From the analysis of the Q statistic results: Q6 does not present significance at the level and cannot reject the hypothesis that the residuals of the model are white noise series, while the goodness-of-fit R^2 of the model is 0.982, the model performs well and the model basically meets the requirements [5].

The model residual autocorrelation plots ACF and PACF are basically within the dashed line, and AR and MA are white noise series. The derived time series plots are partially as Figure 3 shows.

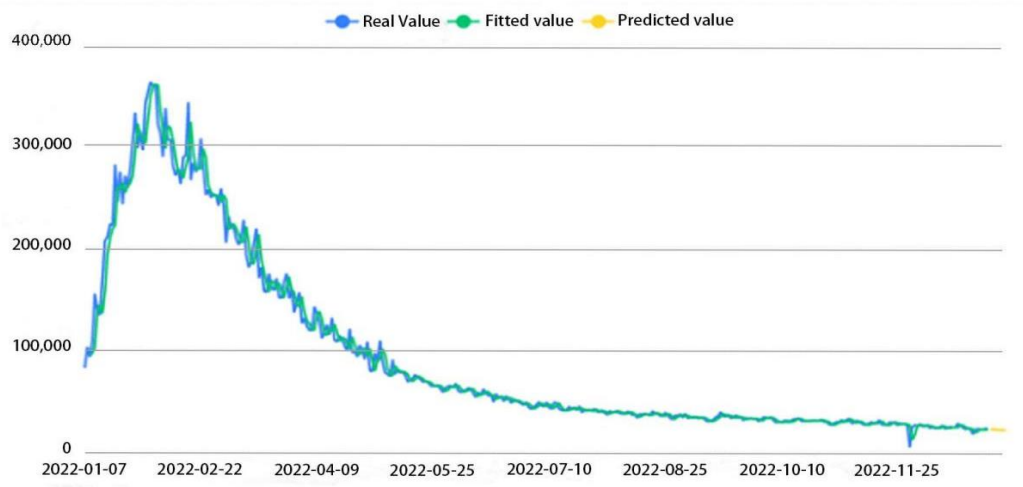

Figure 3. Time Series Chart

Table 3 shows the effect of predicting the last 60 data by the first 305 data in a year.

Table 3. Time series prediction table

Predicted value	
Order (time)	Predicted results
1	17200.285681608406
2	17105.931262292015
…	…
58	16958.269044425928
59	16933.972953756345
60	16917.742669670384

Then the interval is [16917.742669670384, 17200.285681608406].

2.3. The effect of word attributes on difficult mode scores

(1) Commonness of the words

Use Google books ngram viewer to calculate the commonness of these words. Google booksngram viewer is an online tool provided by Google that can be used to see how often a word appears in the Google books database over a period of time, and this tool contains the Google This tool contains the content of millions of books in the Google books database, so it is a good indicator of how often a word is used in English text.

By entering words into Google books ngram viewer, if a word appears more often in English text, then it is more common and more frequently used.

(2) The number of repetitions of letters in a word

Set 1 for 1 repetition and 0 for no repetition.

Then pre-process the data, some words are not common and not found in Google booksngram viewer, so the commonness is changed to 0. Because the topic says all five letters, but there are 4 non-five-letter words, I finally did not delete them because the data is given.

(3) Relationship between word attributes and difficulty mode scores

This article uses Pearson correlation analysis to output a correlation coefficient table (Table 4) and correlation coefficient heat map.

Table 4. correlation coefficient table

	Commonness	Number of repeated letters	Number in hard mode
Commonness	1 (0.000***)	-0.05 (0.341)	0.218 (0.000***)
Number of repeated letters	-0.05 (0.341)	1 (0.000***)	-0.059 (0.268)
Number in hard mode	0.218 (0.000***)	-0.059 (0.268)	1 (0.000***)

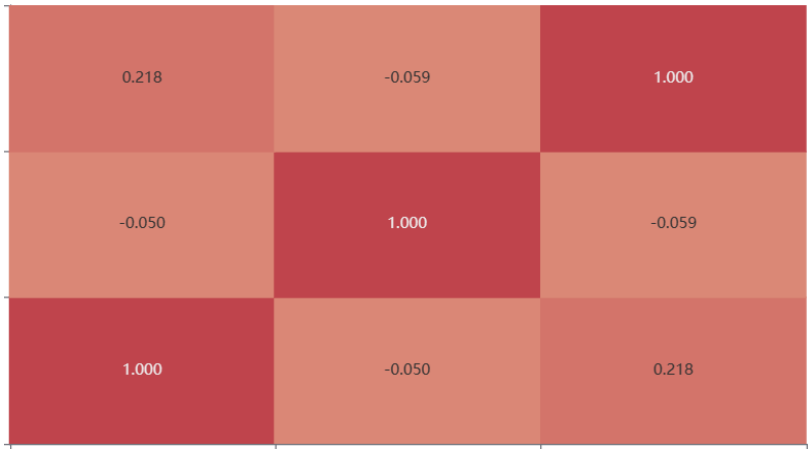

Figure 4. correlation coefficient heat map

From Figure 4, it can be concluded that there is a significant correlation between number in hard mode and commonness, while there is no significant correlation with the number of repeated letters.

3. GBDT model prediction tries report

3.1. Preliminary analysis

This article then aims to calculate the percentage of each number of guesses by the attributes of a word (commonness, number of repeated letters). Using these two attributes, a model is trained for each category of guesses to obtain a model. Eventually there will be seven trained models, which will be input seven times to get the final result, but the probability will not add up to 1. The question also says that the sum of the percentages is not always equal to 100%, so it is normal that all do not equal 1 [6-8].

3.2. Using gradient boosting tree GBDT prediction model

(1) Principle of gradient boosting tree prediction model

GBDT is an iterative decision tree algorithm consisting of multiple regression decision trees, and the output of all decision trees is summed up as the final output:

$$f_m(x) = f_{m-1}(x) + T(x;\theta_m) \tag{1}$$

For the regression problem, if squared error loss is used, then:

$$L(y, f_{m-1}(x_i) + T(x_i;\theta_m)) = [y - f_{m-1}(x) - T(x,\theta_m)]^2 \tag{2}$$

Where $r = y - f_{m-1}(x)$ is the residual (actual value - predicted value) of the current model on the training set, and each round of base model learning is actually fitting the residual of the current model.

The GBDT algorithm uses negative gradient to approximate the residuals, and the learning problem of the base model is transformed into fitting the current overall loss negative gradient, making each loss function available. sklearn's GBDT implementation, MSE, MAE, Huber and other losses are available for regression problems, and Deviance (logistic for binary classification and

softmax for multi-classification) is available for classification problems. softmax), and Exponential loss.

(2) One try training process

First, the parameter names and parameter values of the training procedure can be obtained. The value of the training time is 0.037 s and the data cut is 0.7. The loss function and node classification evaluation criterion is friedman_mse, the number of base learners is set to 100, and the learning rate is 10%. It is worth noting that the minimum number of samples for internal node splitting is 2, the minimum number of samples for leaf nodes is 1, the minimum weight is 0, and the maximum depth of the tree is 10. There is no threshold for node division impurity.

The importance of the features of commonness and the number of repeated letters (as Figure 5 shows) can be obtained.

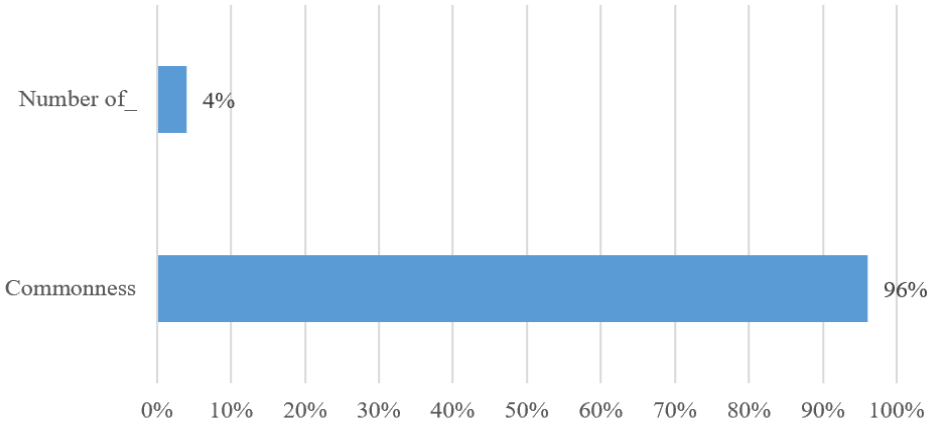

Figure 5. features of commonness and the number of repeated letters

This is consistent with the correlation analysis above, where the degree of commonness is more important, indicating that there is no problem with the analysis.

Finally, the model evaluation results and test data prediction results in Table 5 can be obtained.

Table 5. model evaluation results for 1 try

MSE	RMSE	MAE	MAPE	R^2
0.008	0.087	0.03	81.174	0.989

The R^2 of the training set is very close to 1, which indicates that the model is highly accurate (shown in Figure 6).

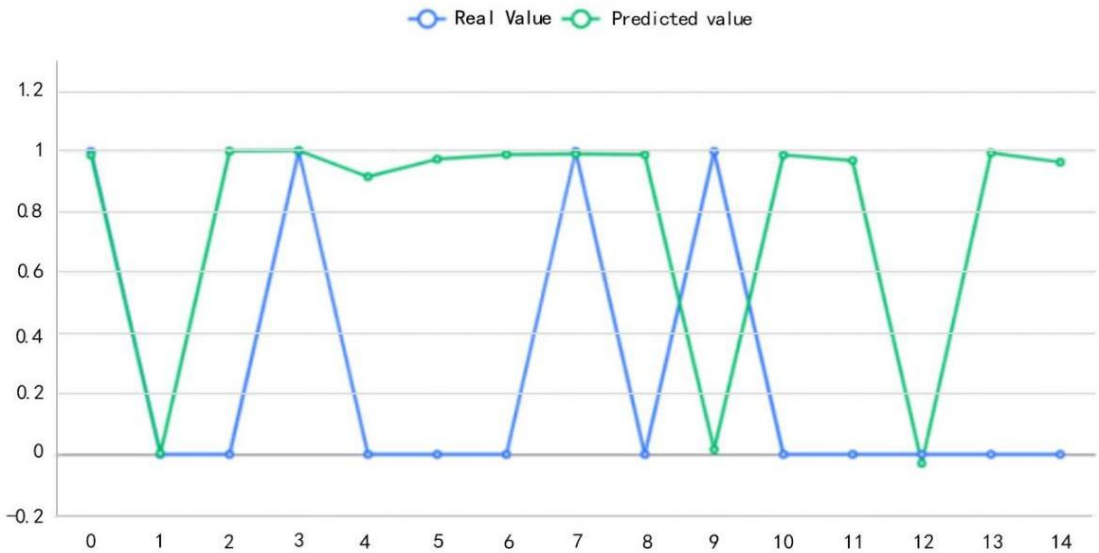

Figure 6. Test Data Prediction Chart for 1 try

(3) Apply the same training method to two or more tries

Table 6. model evaluation results for 2 tries

2 tries	MSE	RMSE	MAE	MAPE	R^2
Training set	0.141	0.375	0.255	7.008	0.991
Test set	25.179	5.018	3.588	168.276	0.478

Table 7. model evaluation results for 3 tries.

3 tries	MSE	RMSE	MAE	MAPE	R^2
Training set	1.575	1.255	0.73	3.863	0.972
Test set	108.091	10.397	8.309	49.585	-0.641

4 tries or more can be derived in the same way, and the graphs and tables are not listed due to space limitation. It is noticable that the accuracy of each model is very high[9].

3.3. Substitute the word EERIE into the model

First, the article calculates how often the word "EERIE" is used: according to Google Books Ngram Viewer, from 1800 to 2008, "EERIE" appeared 1821 times in English books In the same period, the average number of books published per year was 1,821. During the same period, the average number of words published each year was about 5.186 billion. Therefore, the number of occurrences of "EERIE" per million words is about (1821/5,186 million)*1000000≈0.0352, so it can be assumed that the frequency of this word is very low, about 0.0352 times per million words.

The number of repetitions of "EERIE" is 2.

Finally, seven calculations were performed to obtain Table 8.

Table 8. final prediction table

1try	2tries	3tries	4tries	5tries	6tries	7or more tries (X)	commonness	Number of repeated letters
0.003	3.937	10.906	28.533	34.943	20.035	4.5485	0.0352	2

The summation is 102.904931552%, which is very close to 100%. Because the performance of the training set is all above 90%, the accuracy is very good, so for this simulation the prediction accuracy is high.

Because the prediction of 3.1 was found in the first question, the score interval for everyone is [16917.742669670384, 17200.285681608406]. The answer to the first question was added because the prediction for the word 3.1 "EERIE" was found.

3.4. Extensions in terms of models

There are many models for decision making, not only GBDT models, but also classical models such as decision tree models and random forests [10] .Take random forest as an example, random refers to sample random (with put-back sampling) and feature random (generally take $K=log_2(d)$, d is the number of all features). The final prediction result can be determined by weighted average method, voting method, etc. In this question to study attribute correlation, it can be a good measure of feature importance and sample similarity without worrying about over-fitting. The results can be calculated using this model as well. However, this model was not chosen because the random forest does not give a continuous output and does not perform as well as the classification problem.

4. Conclusion

Based on time series with gray prediction and GBDT model, the article predicts the interval of game report results for a future date and the distribution results of attempts for a word, and classify all words by difficulty to study word attributes.

In order to explain the daily variation in reported results and to explore whether lexical attributes affect the percentage of scores in the difficult mode, this article introduced a time series model for specific predictions, respectively, and finally arrived at a result interval of [16917.742669670384, 17200.285681608406] for the March 1, 2023 report. After that, the passage chose the number of occurrences per million words in a standardized way as the characterization quantity of commonness, coming to the conclusion that there is a significant correlation between tries in hard mode and commonness, while there is no significant correlation directly with the number of repeated letters.

In addition, based on previous research, a model was developed to predict the distribution of reported results for games at future dates. This article introduces the prediction model based on Gradient Boosted Tree (GBDT) and carried out specific training, where the training accuracy reached more than 95%. The precision is good and the prediction results are proved to be reasonable.

References

[1] Wormley Alexandra S, Cohen Adam B. C-H-E-A-T: Wordle Cheating Is Related to Religiosity and Cultural Tightness. [J]. Perspectives on psychological science: a journal of the Association for Psychological Science, 2022.

[2] Brown Keith A... MODEL, GUESS, CHECK: Wordle as a primer on active learning for materials research [J]. npj Computational Materials, 2022, 8(1).

[3] Chenlu Li, Xiaoju Dong, Xiaoru Yuan. Metro-Wordle: An Interactive Visualization for Urban Text Distributions Based on Wordle [J]. Visual Informatics, 2018, 2(1).

[4] Malakouti Seyed Matin. Utilizing time series data from 1961 to 2019 recorded around the world and machine learning to create a Global Temperature Change Prediction Model [J]. Case Studies in Chemical and Environmental Engineering, 2023, 7.

[5] Ma Shaojuan, Hao Zheng, Li Man. Chatic time series prediction based on unbiased least squares support vector machines [J/OL]. Practice and understanding of mathematics: 1-10 [2023-05-31]. http://kns.cnki.net/kcms/detail/11.2018.O1.20230530.1026.026.html.

[6] Gao Wenzong, Li Zhao, Chen Qusen, Jiang Weiping, Feng Yanming. Modelling and prediction of GNSS time series using GBDT, LSTM and SVM machine learning approaches [J]. Journal of Geodesy, 2022, 96(10).

[7] Wang Xiaowei. Credit risk assessment of banks based on logistic regression and GBDT model [D]. Shantou University, 2021. DOI:10.27295/d.cnki.gstou.2021.000259.

[8] Chen Lu, Wu Hua. Design of GBDT-based ship fuel consumption prediction model [J]. Electronic Design Engineering, 2022, 30(02): 91-95. DOI: 10.14022/j.issn1674-6236.2022.02.020.

[9] Xu Yingzhuo, Guo Bo, Wang Liupeng. Research on game sales prediction model based on GBDT algorithm [J]. Intelligent Computer and Applications, 2023,13 (01): 182-185.

[10] Sridharan K., Komarasamy G... Retraction Note: Sentiment classification using harmony random forest and harmony gradient boosting machine [J]. Soft Computing, 2022, 27(2).