*Artificial Intelligence Technologies and Applications*
*C. Chen (Ed.)*

*doi:10.3233/FAIA231400*

# Predictive Modelling for Wordle Game: Analysing Word Attributes and Future Trends

Jingdi HE[1]
*Xi'an Jiaotong-liverpool University*
*Suzhou, Jiangsu Province, China*

**Abstract.** Wordle, an engaging online crossword puzzle, challenges players to decipher a hidden five-letter word. In this manuscript, our primary focus lies in predicting the distribution of reported results in the Wordle game. To accomplish this, we devised and compared multiple predictive models, seeking the most accurate and stable solution. The foundation of our research revolves around training 31 distinct word attributes through five diverse multi-output regression data processing models. Then we choose a model whose average error of the test set error is very small, indicating that the prediction accuracy of our model is better, and also more stable than other models. Finally, to illustrate the practical application of our model, we provide a concrete example of predicting the associated percentages of (1, 2, 3, 4, 5, 6, X) for the word "EERIE" in future game instances.

**Keywords.** Wordle game, Mathematical modeling, Random forest regression, XGBRegressor, Light gradient boosting machine

## 1. Introduction

Wordle is a globally popular charades game available in more than 60 languages. In this game, players have six chances to guess a five-letter word, and each time it must be a real English word. The game provides feedback for each guess in the form of a color tile indicating whether the guessed letter is in the word and whether it is in the right or wrong position. Players can choose between "normal mode" or "hard mode", which adds extra difficulty by requiring players to use the letters they guessed correctly in subsequent guesses. Wordle has a single daily solution that lets all players try to guess the same word. Each day's game draws a random word from a list of 2,309 words (there are about 13,000 five-letter words in English) [1]. The New York Times offers a daily Wordle crossword puzzle [2], and players often share their scores on social media platforms like Twitter. The daily reported results include the date, the contest number, the word of the day, the number of people reporting their scores that day, the number of players in hard mode, and the percentage of players who either guessed the word or were unable to solve the puzzle in 1, 2, 3, 4, 5, 6 attempts (indicated by X).

There have been some interesting analyses and applications of Wordle. The Wordle games have been studied from the perspective of complexity, demonstrating the NP-

[1] Corresponding author: Jingdi HE, Xi'an Jiaotong-liverpool University, Suzhou, Jiangsu Province, China; email: swagygrandehe@163.com

hardness of their natural formalization [3]. Some mathematical tricks are discussed to win Wordle wisely [4]. Liu designs and evaluates several strategies for solving Wordle, aiming to guess all of the known answers to the current Wordle [5]. How to determine the best starting word has been proposed [6,7]. Some researchers explore correlates of cheating on Wordle using data from Google Trends and Twitter [8]. Wordle has been used to study active learning and explain the active learning process [9]. The new reinforcement learning methods are presented to address the solution of the popular Wordle puzzle [10]. These papers point to a growing interest in Wordle as both a puzzle game and a research tool in fields such as computer science, statistics, and linguistics.

In this paper, focusing on modelling data about Wordle and predicting future trends, we develop a model to predict the distribution of reported results for a given Wordle puzzle at future dates and the uncertainties associated with the model and predictions. Then, we provide a specific example of a prediction about the word "EERIE" in future date.

## 2. Model Building

Aiming to predict the distribution of reported results in the Wordle game, 31 word attributes are analysed. These attributes included word frequency, phonetic writing harmony, number of repeated letters, number of parts of words, and the ratio of vowels to consonants. We constructed attribute vectors for 26 English letters based on these attributes. Through correlation analysis, we identified attributes with significant impact on the percentage of scores in the hard mode. Subsequently, we compared five data processing models and found that the random forest regression model demonstrated the best performance with a low average error of 0.410428571. Hence, we utilized the random forest regression model to train the word data (353 days of data for Wordle.).

### *2.1. Correlation Analysis of Word Attributes*

In the following, we give five attributes of words, namely

- F: Word frequency: This attribute refers to the number of times a given word appears in a given scene(https://books.google.com/ngrams).
- P: Phonetic writing harmony (word length/phonetic symbol length): The degree of consistency between the pronunciation and the form of a word. Define P as the ratio of the length of a word to the length of the phonetic symbol of the word, that is, P=VL/PN. The VL is the number of letters in a word, and PN is the number of phonetic symbols in a word.
- R: Number of repeated letters: The number of repeated letters in the spelling of a word.
- C: Number of parts of words: The number of all possible parts of speech in a word.
- V: Number of vowels/number of consonants: The ratio of vowels to consonants in a word. Define V as the ratio of the length of a word to the length of the phonetic symbol of the word, i.e., V=VY/VF. The VY is the number of vowels and VF is the number of consonants in the word.

| | a | b | c | d | e | f | g | h | i | j | ... | q | r | s | t | u | v | w | x | y | z |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **1** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **3** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **348** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | ... | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **349** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | ... | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| **350** | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 0 | 0 | ... | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **351** | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **352** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

353 rows × 26 columns

**Figure 1.** An example of 26 English letters attributes.

Then, equipped with these attributes, we introduce a method as follows to construct the attribute of 26 English letters.

- Step 1: Construct a vector with character 0 and length 26.
- Step 2: Record the corresponding position of the letter appearing once in the array of each word as 1; if it appears twice, record it as 2, and so on. As shown in Fig. 1.
- Step 3: Get 353 vectors with 26 English letter attributes.

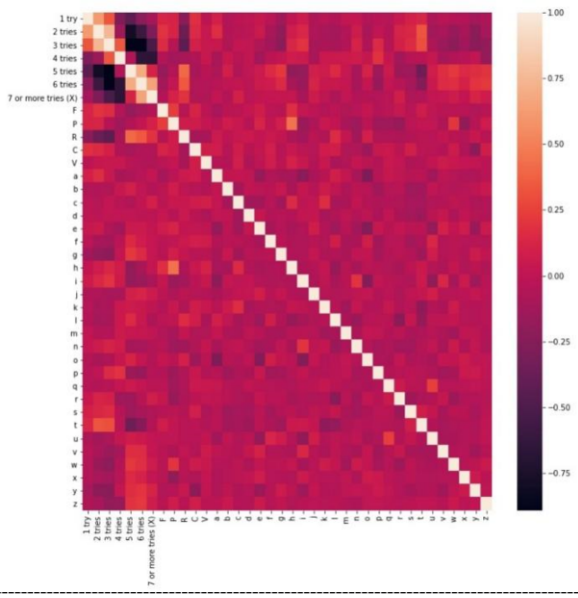


**Figure 2**. The heat map corresponding to the correlation table.

In order to show the correlation more clearly, we use the heat map in Fig. 2, which is more in terms of numerical value or color, to accurately describe the relationship between the various attributes. Since players choose the hard mode before starting the

game, the percentage of players who play hard mode (hard percent) is independent of the word attribute. Moreover, through correlation analysis, the correlation between hard percent and different answer times is very low, so this influence can be ignored. Then, we use the least square method to qualitatively analyse the correlation between 31 word attributes and the proportion of 7 different answers, and the results are shown in Fig. 2. It can be seen that different word attributes have different degrees of correlation to the 7 result values. For example, the correlation between "R" and "3 tries" and "t" and "2 tries" is obvious.

The specific correlation coefficient and P-value are shown in Fig. 3. We analyse in detail how the attributes of words affect the percentage of scores in hard mode. When P-value is less than 0.05, it indicates that this variable has a strong correlation with the score ratio. We take "2 tries" as an example, the "P", "R", "a", "g", "i", "n", "t", "w" on scoring ratio have higher correlation. Therefore, it shows that word attributes have a significant correlation with the percentage of scores.

### *2.2. Model Selection*

In accordance with the word attributes utilized in Section 2.1, we begin by employing five different data processing models to train the word attributes. Subsequently, we compare their performance against the test set to identify a model with favourable results for solving. The first 318 words serve as the training set, and the last 35 words as the test set to obtain a well-performing model. The following five different data processing models are selected.

(1) Neural network (NNs): It is an algorithm mathematical model that imitates the behaviour characteristics of animal neural networks and carries out distributed and parallel information processing.

(2) Decision tree regression (DTR): It refers to classification and regression tree (CART) algorithm. The value of internal node features is "yes" and "no", which is a binary tree structure. The corresponding output value is determined according to the eigenvector. A Regression tree is to divide the feature space into several units, and each unit has a specific output.

(3) Random forest regression (RFR): It is composed of multiple regression trees, and there is no correlation between each decision tree in the forest. The final output of the model is determined jointly by each decision tree in the forest.

(4) Light Gradient Boosting Machine (LGBM), a framework that realizes Gradient Boosting Decision Tree (GBDT) algorithm, supports efficient parallel training and has the advantages of faster training speed, lower memory consumption, better accuracy, supporting distributed and processing massive data quickly, etc.

(5) XGBRegressor (XGBR): The basic idea is the same as GBDT, but with some optimization, such as the second derivative to make the loss function more accurate; regular term to avoid tree overfitting; Block storage can be parallel computing, etc. XGBoost has been widely used in data mining, a recommendation system, and other fields [11].

In summary, 14 groups of data are obtained according to the above five methods, as shown in the following Table. 1. And we expressed the results in this table as a line graph in Fig. 4. As can be seen from Fig. 4, the random forest regression model has a good effect, and the average error of its 7 results is 0.410428571, which is small and stable compared with other models. Therefore, the random forest regression model is used to train the word data.

| Corr | 1 try | | 2 tries | | 3 tries | | 4 tries | | 5 tries | | 6 tries | | 7 or more tries (X) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | coef | p value | coef | p value | coef | p value | coef | p value | coef | p value | coef | p value | coef | p value |
| F | 0.00 | 0.32 | 0.02 | 0.05 | 0.04 | 0.05 | -0.01 | 0.74 | -0.03 | 0.07 | -0.02 | 0.30 | -0.01 | 0.35 |
| P | 0.68 | 0.00 | 3.72 | 0.00 | 0.82 | 0.61 | -3.87 | 0.01 | -2.96 | 0.01 | -0.56 | 0.71 | 2.49 | 0.04 |
| R | -0.37 | 0.00 | -2.90 | 0.00 | -6.25 | 0.00 | -0.87 | 0.14 | 4.63 | 0.00 | 4.18 | 0.00 | 1.57 | 0.00 |
| C | 0.19 | 0.00 | 0.55 | 0.03 | 0.34 | 0.46 | -0.61 | 0.17 | -0.32 | 0.35 | 0.06 | 0.89 | -0.21 | 0.54 |
| V | -0.07 | 0.54 | 0.39 | 0.42 | 1.02 | 0.23 | 0.84 | 0.31 | -0.27 | 0.67 | -1.11 | 0.16 | -0.85 | 0.19 |
| a | 0.17 | 0.06 | 1.89 | 0.00 | 3.86 | 0.00 | 3.07 | 0.00 | 1.21 | 0.01 | -0.19 | 0.75 | -0.33 | 0.49 |
| b | -0.23 | 0.12 | -0.31 | 0.60 | 2.46 | 0.02 | 5.45 | 0.00 | 3.17 | 0.00 | 0.06 | 0.95 | -0.98 | 0.21 |
| c | -0.05 | 0.63 | -0.01 | 0.98 | 3.35 | 0.00 | 3.70 | 0.00 | 1.30 | 0.02 | 0.68 | 0.32 | 0.74 | 0.18 |
| d | 0.00 | 1.00 | 0.39 | 0.38 | 4.20 | 0.00 | 4.30 | 0.00 | 1.29 | 0.03 | -0.11 | 0.88 | -0.18 | 0.76 |
| e | 0.04 | 0.62 | 0.84 | 0.02 | 2.67 | 0.00 | 3.50 | 0.00 | 2.44 | 0.00 | 0.54 | 0.37 | -0.47 | 0.34 |
| f | -0.07 | 0.61 | -0.68 | 0.19 | 0.41 | 0.66 | 3.94 | 0.00 | 4.17 | 0.00 | 1.73 | 0.05 | 0.31 | 0.66 |
| g | -0.25 | 0.02 | -1.32 | 0.00 | -0.29 | 0.72 | 4.70 | 0.00 | 4.67 | 0.00 | 2.18 | 0.00 | -0.05 | 0.93 |
| h | -0.24 | 0.06 | -0.39 | 0.46 | 3.66 | 0.00 | 5.63 | 0.00 | 1.39 | 0.05 | 0.25 | 0.78 | -0.31 | 0.66 |
| i | 0.13 | 0.22 | 1.41 | 0.00 | 4.89 | 0.00 | 4.03 | 0.00 | 0.66 | 0.24 | -0.85 | 0.23 | -0.67 | 0.24 |
| j | -0.40 | 0.29 | -1.51 | 0.32 | -4.52 | 0.10 | -1.08 | 0.68 | 6.65 | 0.00 | 7.68 | 0.00 | 2.93 | 0.15 |
| k | -0.11 | 0.40 | -1.08 | 0.04 | 0.37 | 0.70 | 4.78 | 0.00 | 3.78 | 0.00 | 1.85 | 0.04 | -0.06 | 0.93 |
| l | -0.06 | 0.43 | 0.15 | 0.66 | 3.20 | 0.00 | 4.72 | 0.00 | 2.25 | 0.00 | 0.14 | 0.80 | -0.73 | 0.11 |
| m | 0.01 | 0.94 | -0.33 | 0.42 | 1.08 | 0.14 | 4.55 | 0.00 | 3.16 | 0.00 | 1.58 | 0.02 | 0.10 | 0.86 |
| n | 0.01 | 0.88 | 0.86 | 0.02 | 4.13 | 0.00 | 4.36 | 0.00 | 1.18 | 0.02 | -0.14 | 0.83 | -0.21 | 0.68 |
| o | 0.08 | 0.37 | 0.94 | 0.01 | 3.97 | 0.00 | 4.30 | 0.00 | 1.86 | 0.00 | -0.54 | 0.37 | -1.03 | 0.04 |
| p | -0.13 | 0.19 | 0.00 | 1.00 | 5.00 | 0.00 | 6.46 | 0.00 | 0.66 | 0.24 | -1.46 | 0.04 | -0.33 | 0.56 |
| q | -0.36 | 0.30 | -1.80 | 0.20 | -0.95 | 0.71 | 5.00 | 0.04 | 5.97 | 0.00 | 2.53 | 0.28 | -0.12 | 0.95 |
| r | 0.03 | 0.66 | 1.08 | 0.00 | 3.98 | 0.00 | 2.50 | 0.00 | 0.51 | 0.24 | 0.54 | 0.32 | 0.95 | 0.03 |
| s | 0.10 | 0.26 | 1.04 | 0.01 | 4.69 | 0.00 | 3.59 | 0.00 | 0.94 | 0.06 | -0.15 | 0.81 | -0.48 | 0.34 |
| t | 0.08 | 0.32 | 1.96 | 0.00 | 5.90 | 0.00 | 4.12 | 0.00 | -0.46 | 0.30 | -1.48 | 0.01 | -0.38 | 0.40 |
| u | -0.04 | 0.74 | -0.23 | 0.63 | 2.16 | 0.01 | 4.16 | 0.00 | 2.68 | 0.00 | 0.95 | 0.24 | -0.16 | 0.81 |
| v | -0.14 | 0.36 | -1.87 | 0.00 | -2.73 | 0.02 | 2.28 | 0.04 | 6.18 | 0.00 | 5.17 | 0.00 | 0.81 | 0.34 |
| w | -0.27 | 0.08 | -2.40 | 0.00 | -2.88 | 0.01 | 3.65 | 0.00 | 6.58 | 0.00 | 4.28 | 0.00 | 0.82 | 0.33 |
| x | -0.19 | 0.54 | -2.06 | 0.10 | -6.15 | 0.01 | 0.67 | 0.75 | 8.67 | 0.00 | 7.24 | 0.00 | 1.75 | 0.30 |
| y | -0.05 | 0.64 | -0.90 | 0.05 | -0.38 | 0.64 | 3.54 | 0.00 | 3.64 | 0.00 | 3.26 | 0.00 | 0.84 | 0.17 |
| z | -0.71 | 0.04 | -4.14 | 0.00 | -9.37 | 0.00 | -2.86 | 0.22 | 11.78 | 0.00 | 12.47 | 0.00 | 2.78 | 0.13 |

**Figure 3.** Qualitative analysis of the correlation between 31 word attributes and 7 different answer times.

**Table 1.** The train results of five methods.

| | | NNs | DTR | RFR | LGBMR | XGBR |
|---|---|---|---|---|---|---|
| 1 try | train | 0.965 | nan | 0.719 | 0.799 | 0.63 |
| | test | 3.696 | nan | 1.823 | 2.225 | 1.3 |
| 2 tries | train | 0.783 | nan | 0.153 | 0.292 | 0.02 |
| | test | 0.448 | inf | 0.447 | 0.292 | 0.461 |
| 3 tries | train | 0.18 | 0 | 0.114 | 0.164 | 0.002 |
| | test | 0.231 | 0.343 | 0.222 | 0.219 | 0.22 |
| 4 tries | train | 0.114 | 0 | 0.07 | 0.092 | 0 |
| | test | 0.116 | 0.196 | 0.136 | 0.121 | 0.125 |
| 5 tries | train | 0.129 | 0 | 0.085 | 0.141 | 0 |
| | test | 0.166 | 0.304 | 0.206 | 0.215 | 0.197 |
| 6 tries | train | 0.319 | 0 | 0.234 | 0.293 | 0 |
| | test | 0.4 | 0.638 | 0.506 | 0.535 | 0.557 |
| 7 or more tries (X) | train | 0.922 | nan | 0.58 | 0.765 | 0.077 |
| | test | 1.048 | nan | 1.237 | 1.404 | 1.517 |

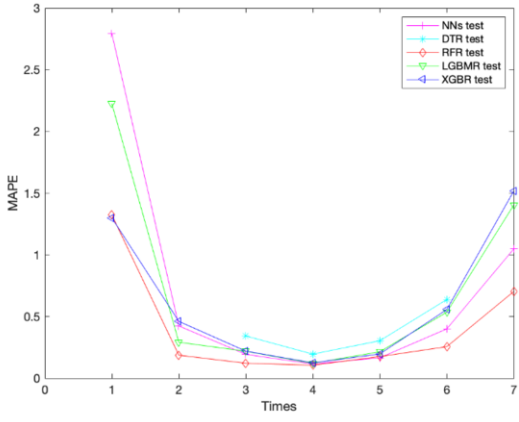


**Figure 4.** A line graph of Table. 1.

## 3. Example: Distribution Prediction

As a demonstration of the model's capabilities, we provided an example of predicting the associated percentages of (1, 2, 3, 4, 5, 6, X) for the word "EERIE" in the future. This showcases the practical applicability of our research findings and serves as a testament to the potential real-world utility of our predictive approach. First, calculate the 26 letter attributes of the word EERIE and the 5 attributes presented in Section 2.1. We get "EERIE" can be expressed as (vector of length 5+26=31): $[0.2243, 1.667, 3, 1, 4, 0, 0, 0, 0, 3, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0]^T$ . By substituting the calculated word attributes into the trained random forest regression model, Finally, the ratio of (1, 2, 3, 4, 5, 6, X) is [0.61073173, 8.30230422, 28.25678445, 33.83707426, 19.98188209, 8.1649541, 0.65665061]. The sum of all attempts is 99.8104%, which is closer to 100%.

Since the average error of this random forest regression model is 0.410428571 and the total prediction ratio is 99.8104, we are full of confidence in this model, and there is a significant correlation between the selected attributes and the percentage of scores. From the above analysis, it can be seen that the word attributes selected by us can be used as the standard for predicting the score percentage of New York Times.

## 4. Conclusion

In conclusion, our study focused on predicting the distribution of reported results in the Wordle game by employing a comprehensive set of 31 word attributes and evaluating them through five different multi-output regression data processing models. Through rigorous experimentation and analysis, we identified a specific model that demonstrated outstanding performance in terms of prediction accuracy and stability on the test set. The success of our model offers promising prospects for enhancing the gaming experience of Wordle players. By accurately predicting the distribution of potential outcomes, we can potentially provide players with insights into the likelihood of different word combinations, enhancing their strategic decision-making process. Moreover, the developed model can be employed as a valuable tool for analyzing and understanding word attributes and their impact on Wordle outcomes. We hope that our findings will inspire further research in this domain and foster the development of more accurate and sophisticated models for predicting outcomes in word-based puzzles and beyond.

## References

[1] Daniel Victor. Wordle is a love story. The New York Times, 1 2022.
[2] Alexis Benveniste. The sudden rise of wordle. The New York Times, 1 2022.
[3] Daniel Lokshtanov and Bernardo Subercaseaux. Wordle is NP-hard. arXiv preprint arXiv:2203.16713, 2022.
[4] Martin B Short. Winning wordle wiselyor how to ruin a fun little internet game with math. The Mathematical Intelligencer, 44(3):227–237, 2022.
[5] Chao Lin Liu. Using wordle for learning to design and compare strategies. In 2022 IEEE Conference on Games (CoG), pages 465–472. IEEE, 2022.
[6] Nisansa de Silva. Selecting seed words for wordle using character statistics. arXiv preprint arXiv:2202.03457, 2022.
[7] Benton J Anderson and Jesse G Meyer. Finding the optimal human strategy for wordle using maximum correct letter probabilities and reinforcement learning. arXiv preprint arXiv:2202.00557, 2022.

[8] Alexandra S Wormley and Adam B Cohen. Cheat: Wordle cheating is related to religiosity and cultural tightness. Perspectives on Psychological Science, 2022.
[9] Keith A Brown. Model, guess, check: Wordle as a primer on active learning for materials research. NPJ Computational Materials, 8(1):97, 2022.
[10] Siddhant Bhambri, Amrita Bhattacharjee, and Dimitri Bertsekas. Reinforcement learning methods for wordle: A POMDP/adaptive control approach. arXiv preprint arXiv:2211.10298, 2022.
[11] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794, 2016.