Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2313988

Little Wordle, Big Data: a Prediction and Classification Model Behind Wordle

Summary

Last year, Wordle was very popular all over the world. The three members of our team have been devoted followers of this small yet superb puzzle game for years. We find it to be highly intriguing and meaningful to study this game.

In order to evaluate the difficulty of words, we creatively put forward two special indicators, which are shown to have excellent effects on the evaluation of wordle game guessing after experimental analysis.

For the first problem, we first analyze the whole trend of the reported numbers. We found that there would be a weekly change rule except from the whole trend of the data set. So we used the Prophet to explain the variation. The whole trend of the curve is to go up first rapidly, then go down and tend to be stable. While the weekly trend of the curve is to go up and go down. It reaches the peak on Wednesday. Then we compare the MAE, MSE, MAPE between ARIMA, Holt-Winter and Prophet. We found Prophet is the best of all. By using Prophet we predicted that interval would be (18580, 24940) and the expected value is 20253.1. None of the attributes of the word would affect the percentage of scores reported that were played in Hard Mode mainly because the player couldn't know the word before so they choose Hard Mode by their own willing.

For the second problem, we first observed the distribution of the data. After calculation of the skewness, we found the skewness is very small. So we considered the distribution obey the normal distribution. So if we can find out the mean value and the variance of a specific day, we can then use the normal distribution to give a prediction. We used our model VARMAX to give a prediction for the word EERIE on March 1, 2023. The result is in the **text**. There might be some uncertainties in the model. One of the uncertainties would be cheaters who just like to show how brilliant they are. Another uncertainty is that a lot of people take part in this competition and get to know Wordle so there might be a little rise in the number of players which might affect the final results. We also compared the VARMAX with lots of machine learning ways. The comparison shows that VARMAX is the best of all which make us quite confident in our prediction.

Thirdly, we used K-means to divide all the words into three groups: low difficulty, medium difficulty and high difficulty. Then we used SVM to do another classification. At last, we put the EERIE into the model to classify. The result shows that the EERIE is in medium difficulty group. As for the accuracy, our model SVM has the best performance. The accuracy of its prediction is 0.74.

During the time we do lots of researches on this data set, we figured out some other interesting features. For example, we found that nouns are most likely to be guessed correctly while the verbs are difficult to be guessed correctly.

Last but not least, as three loyal fans of Wordle, we sincerely hope it could be better and better. So we wrote a letter to the Puzzle Editor of the New York Times to give some advice.

Keywords: Gini coefficient; PCF; Prophet; Varmax; K-means; SVM;

Contents

1 Introduction

1.1 Problem Background

Wordle is a very popular puzzle game which is provided daily by the ***New York Times***. Players attempt to guess a five-letter word within six tries to solve the puzzle. After every guess you will have a feedback. Specifically, if the tile becomes gray, it means the letter in that tile does not exist in the word; if the tile becomes yellow, it means the letter in that tile exists in the word but in a wrong place; if the tile becomes green, it means the letter in that tile exists in the word and in the right place. Players can play in regular mode or "Hard Mode". Wordle's "Hard Mode" makes the game more difficult by requiring that once a player has found a correct letter in a word (the tile is yellow or green), those letters must be used in subsequent guesses. **Figure 1** provides an example solution where the correct result was found in 5 tries.

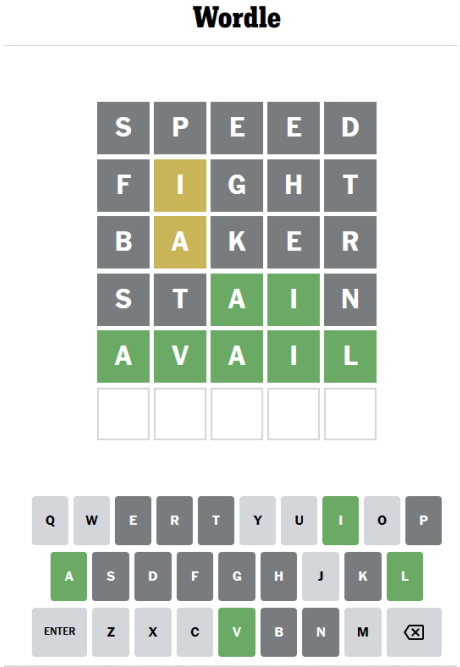


Figure 1: Example solution of Wordle Puzzle from February 18, 2023

Wordle has been public since October 2021. Then it became a hit since its birth rapidly. Countless people played it together. Then it was purchased by the New York Times in January 2022[3]. Since then, players have complained that the daily Wordle puzzles have become more difficult. The New York Times responded that they did not intentionally alter the difficulty of Wordle – if anything, they made the game easier by eliminating more obscure words[4]. However, it is evident that some Wordle puzzles have a significantly high average number of attempts, indicating that some words are more difficult to guess than others[5].

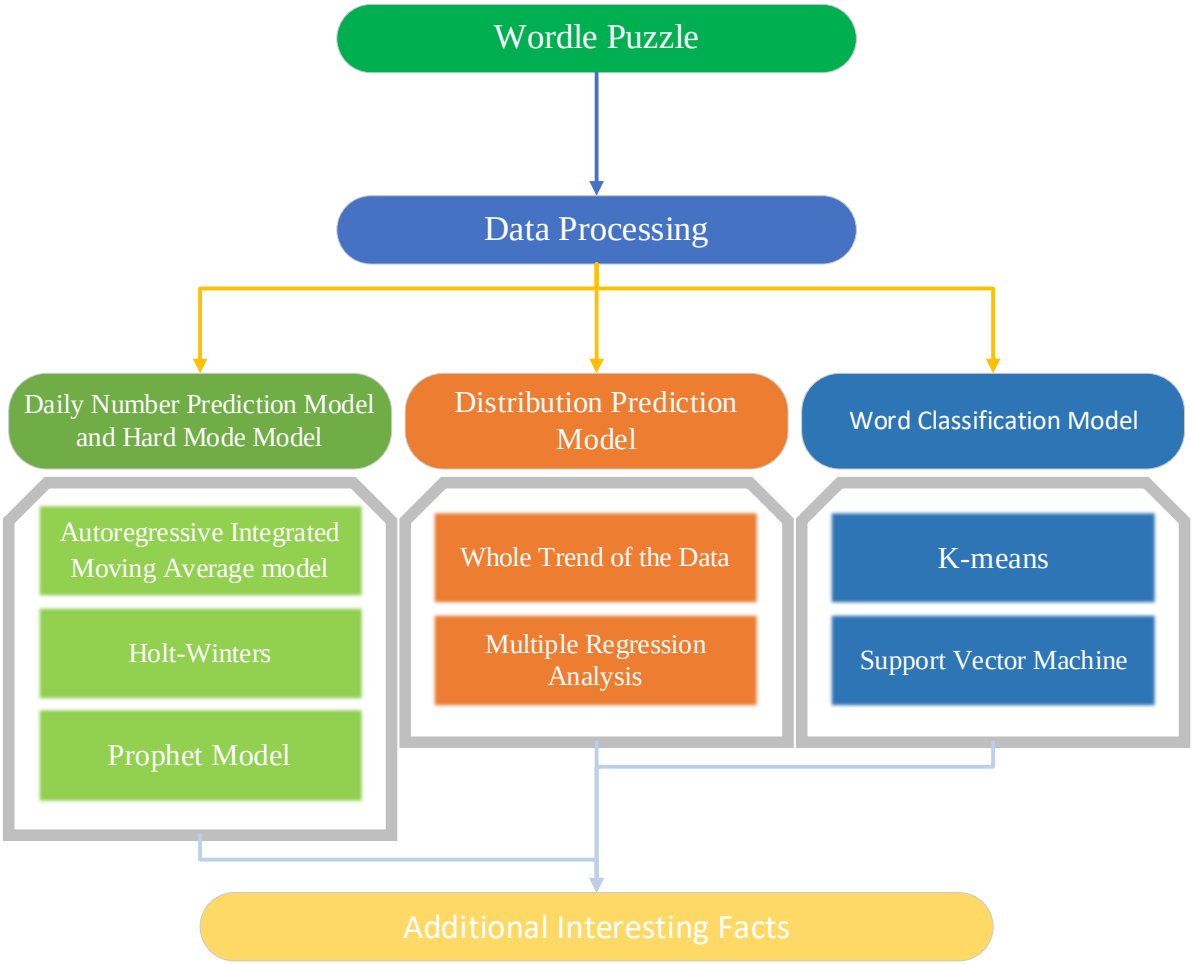
Based on this popular puzzle game, we will do some research to see some interesting facts about it.

1.2 Restatement of the Problem

Considering the background information and restricted conditions identified in the problem statement, we need to solve the following problems:

- Build a math model to explain why the number varies daily and use it to predict an interval for the number on March 1, 2023. Then analyze if any attributes of the word affect the percentage of scores reported that were played in Hard Mode and explain the analysis.
- Develop a model to predict the percentages of (1,2,3,4,5,6,X) for a future date and use it to give a prediction for the word EERIE on March 1, 2023. Also required to evaluate the model and expound the uncertainties of the model.
- Build a model to classify the words by difficulty and identify the attributes of a given word. Then using the model to estimate the difficulty of EERIE and discuss the accuracy of the model.
- Show some other interesting features of the data.

1.3 Our Work



2 Assumptions and Justifications

Assumption: There is no sudden increase or decrease in the report due to some unexpected circumstances.

Justification: Using the data set provided in this topic, we can only build some models which do not consider unexpected circumstances. Also, there is very low chance that some unexpected things happen and influenced the result significantly.

Assumption: People would not use the internet to get to know the answer before they play.

Justification: We know that there would be a lot of people showing the correct answer on Tiwitter every day, so it's possible to get to know the answer before playing. But most people wouldn't do this kind of thing and it's hard to find how many people would like to cheat.

Assumption: People's willingness to report on Twitter has nothing to do with the difficulty or the results of their answers today.

Justification: For all of us, we tend to show our achievement to others. So the people who get a good result will be more likely to report their results on Twitter. But this is just a leisure game, people would not value their results so much.

Assumption: The skills of people playing wordle will remain at a good level after the rapid growth period.

Justification: When the wordle is at its rising period, most people play it for their curiosity. After the period, the people who insist on playing it really love this game. It's obvious that they will have a high level of skills.

Assumption: Everyone has the same level of knowledge to make assumptions and explanations about words.

Justification: Most people would not study some extra knowledge to solve the puzzle. Most of them are just normal people. Their understanding of how to solve the puzzle is thus pretty similar.

Assumption: The difficulty of the word has no relationship with the specific date.

Justification: Every day's word is quite random. It's not very likely that wordle would give a word according to the date.

3 Notations

The key mathematical notations will be explained in the article.

4 Data Processing and Features Extracting

4.1 Data Processing

1. We first sorted the data according to chronological order to check if there is any missing item or exceptional item. We found an exceptional item about the number of reported results on November 30, 2022 and replace it with real data. (2569 -> 25569)

2.We then went through all the words in the data set. We concluded that there are four words which should be corrected as there is misspelling. Specifically, they are trash(tash) on April 29, 2022; marsh(marxh) on October 5, 2022; clean(clen) on November 26, 2022; probe(rprobe) on December 16, 2022.

3.We recalculated the percentage of each try to make the sum of each percentage of different tries is actually 100%.

4.We put another tag on each word to state the day of the week of itself.

5.Considering the subsequent evaluation and prediction, etc., we divided the data in December into tests, and the data of the previous months as modeling data, so that we can reasonably evaluate the model. But when we forecast the data on March 1, 2022, we will take the data of December into consideration.

4.2 Words' Features Extracting

We extract features from words to find some attributes of words. In general, the difficulty of a solution word with the same length in Wordle can be influenced by a combination of factors, including its complexity[6], rarity, and the amount of specialized knowledge required to solve it.

We predict that frequency, the structure of the word, and part of speech will all have a noticeable impact on how many guesses people need to solve a Wordle. We believe that obscurity and word structure will have a significant effect, because players will be less familiar with obscure words, requiring them to make more guesses. Additionally, since players typically use familiar spelling patterns to make better guesses, we expect that words with letters in an unusual order will also be more difficult to guess.

We consider the impact of the following aspects.

4.2.1 Word Frequency

More rare words (i.e., less frequently used) will be more difficult because they may be unfamiliar, or at least not the most important when we search for new guesses. Word frequency can be regarded as a statistical feature of lexical difficulty. That is, the number of lexical repetitions in a certain number of real corpora. It determines the degree of vocabulary use and learners' familiarity with vocabulary, so it is an important factor in quantifying vocabulary difficulty. Generally speaking, the higher the frequency of word use, the more common it is for learners, and the less difficult it is to remember; on the contrary, the lower the frequency of word use, the lower the familiarity of learners with it, and the more difficult it is to memorize. Given that some words seem to be more challenging than others, one possibility is that less common words are more difficult because even if the cues we've learned single them out, they are unlikely to be the primary consideration.

By using the data from Wolfram[7], we can get the word frequency. After importing the frequencies of the words with five letters, we sort the word frequencies from large to small, give the rank of the words' frequencies as attribute feature 1, and regularize it.

We are inspired by the game mode. If a word is quite similar to some other words, it might bring some effects to guess which one it is. On word frequency and similarity data, it is easier to infer the morphology of other words from the letter patterns contained in high word frequency data. At the same time, considering that words with the same letter position are easier to infer from each other, we propose a new index called Pattern Compare Frequency(PCF).

To propose PCF, we need to propose Pattern Compare String(PCS) and Bool Pattern Compare String(BPCS).

For PCS, we consider two 5-letter-words. They are $word_a = (a_1, a_2, a_3, a_4, a_5)$ and $word_b = (b_1, b_2, b_3, b_4, b_5)$. PCS(a, b) would be as followings:

$$c_i = \begin{cases} 0, a_i \neq b_i \\ 1, a_i = b_i \end{cases}, i = 1, 2, 3, 4, 5 \tag{1}$$

The BPCS of two PCSs would be as followings:

$$BPCS(PCS_1, PCS_2) = \begin{cases} 1, \text{if for every} PCS_1[i] \geq PCS_2[i] \\ 0, else \end{cases} \tag{2}$$

The PCF is as followings:

$$PCF(word) = \sum_{i=1}^{n} \frac{word_{frequency}}{\sum_{j=1}^{n} BPCS(PCS(word, word_i), PCS(word, word_j))} \tag{3}$$

We use this PCF as attribute feature 1.

4.2.2 Complex Letter Patterns

Let's imagine the word we will guess is EERIE. It's hard to guess there are three same letter 'e' in this word. Or maybe the word we're guessing is GHOST. It's hard to guess there is a 'gh' in this word. From these two simple examples we can know that if there is any complex letter patterns in the word would make this word hard to guess.

One way to measure the complexity of letter patterns in a word is to use a metric such as Shannon entropy. Entropy is a measure of the amount of uncertainty or random-ness in a system. In the case of a word, the entropy can be calculated by determining the probability of each letter appearing in the word and then using that to calculate the entropy of the overall distribution.

To calculate the entropy, we first counted the frequency of each letter in the word. Then we calculated the probability of each letter by dividing its frequency by the total number of letters in the word. Lastly, we used the probability of each letter to calculate the entropy of the distribution using the formula:

$$H = -\Sigma(p \log_2 p) \tag{4}$$

In the formula, H is the entropy, p is the probability of each letter, and Σ is the sum over all letters in the alphabet.

Another approach is to use a metric such as the Gini coefficient, which is often used in economics to measure income inequality. In the context of letter patterns, the Gini coefficient can be used to measure the degree to which the frequency of letters in a word is evenly distributed. A higher Gini coefficient would indicate that the word has a more uneven distribution of letters and thus more complex letter patterns.

To calculate the Gini coefficient of the words, we first calculated the Gini's mean difference.

$$\Delta = \frac{\sum_{j=1}^{n}\sum_{i=1}^{n}|Y_j - Y_i|}{n^2}, 0 \leqslant \Delta \leqslant 2u \tag{5}$$

In the formula, Δ is the Gini's mean difference, $|Y_j - Y_i|$ is the absolute value of the sample difference, n is the length of the word(under this circumstance n is 5), u is the mean value.

Then we calculated the Gini coefficient by using the formula:

$$G = \frac{\Delta}{2u}, 0 \leqslant G \leqslant 1 \tag{6}$$

In the formula, G is the Gini coefficient.

We use the Gini coefficient of words as attribute feature 2.

Last but not least, we used another index as word_length which means the different letters the words have. We use it as attribute 3.

4.2.3 Extra indexes

Except from the three features above, we also use a lot of other indexes.

For example, we use the number of consonants, the longest continuous consonant, begin with a vowel or not and so on.

4.3 Wordle game's difficulty feature

We choose four evaluating coefficients to use.

4.3.1 Expected number of tries

For people who have successfully solved the problem ,the simplest approach is to look at the expected number of tries(Expect tries) for each target word, on the grounds that less tries point to easier words. The problem is that a simple expected can disguise a lot of what is going on.

4.3.2 Proportion of people solved problems

An obvious fact is that the proportion of people who successfully solve problems to all people is an important indicator to measure the difficulty of words. We have reason to believe that a word puzzle will be simpler if more people can solve it.

P_Win means the probability of successfully guessing the word, that is, games that are completed within 6 rounds

4.3.3 Ratio of long and short games

If a word is associated with mostly short games then it is easier than a word that is associated with mostly long games. According to the distribution of tries, if we assume that the ratio of the people who spend more than 4 chances to guess the word correctly is a and the ratio

of the people who spend less than 4 chances to guess the word correctly is b, then we define Difficulty Ration (DR) is as following.

$$DR = (b - a)/a \tag{7}$$

Because of the calculation of DR, we can measure the complexity of the word on that day.

5 The Daily Number Prediction Model and Hard Mode Model

5.1 The Establishment and Solution of Daily Number Prediction Model

At the beginning, we drew an image which reflects the variation of the number of the reported results and the number of "Hard Mode" over time. **Figure 2** shows our result.

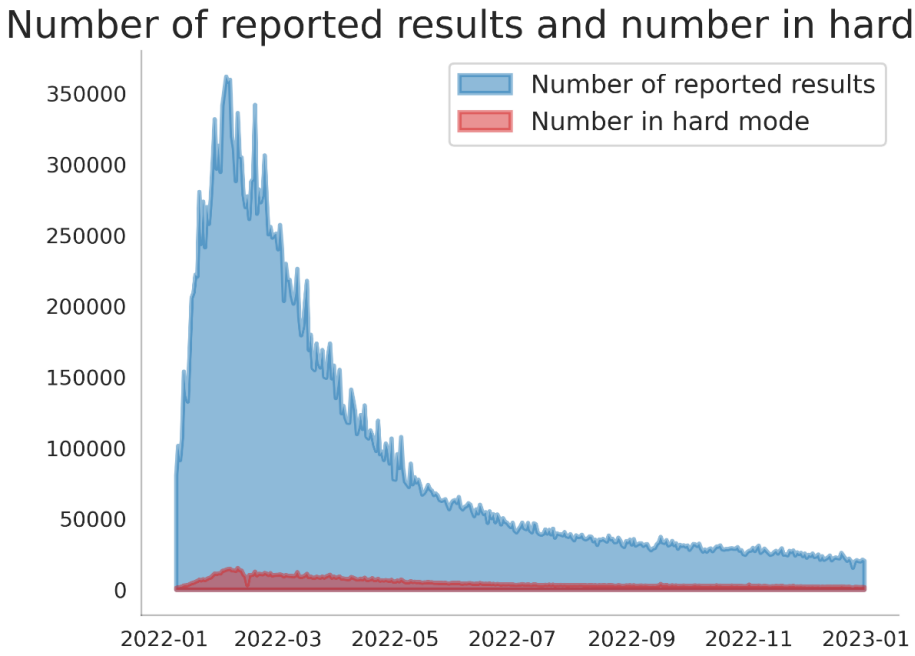


Figure 2: Number of reported results and number in hard mode

Also we give a conjecture that the day of the week would do some influences to the number of the people who play it. So we use the pie chart to find out if there is any connection between the day of the week and the number. **Figure 3** shows the pie chart.

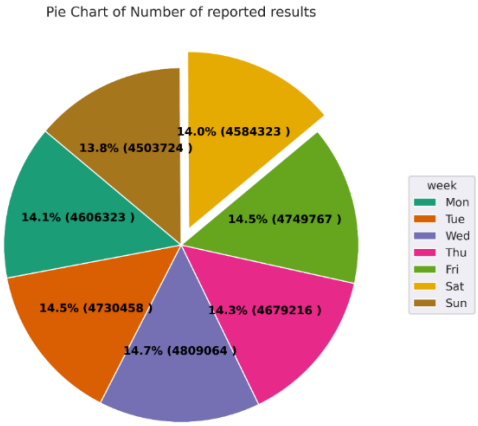


Figure 3: the pie chart of number of the reported results

Clearly, there are the most people to play Wordle on Wednesday which means on 1 March, 2023 there would be a small rise on its number.

5.1.1 The Autoregressive Integrated Moving Average model[1]

Taking this picture into consideration, we can clearly know that the number first went up and then went down. During the whole process, it reached its peak in February. So clearly we need a model which can fit this kind of curve.

The Autoregressive Integrated Moving Average model(ARIMA) is the most common way to complete some time series prediction tasks. The ARIMA model is based on the autocorrelation of time series data to describe the short term memory of the series, so it has the nature of short term prediction. We can see that every point have a strong connection with its nearby points. So ARIMA is quite suitable for this prediction.

For AR, it describes the relationship between the current value and the historical value, and uses the historical time data of the variable to predict itself. The formula definition of p-order autoregressive process is as **(8)**:

$$y_t = \mu + \Sigma_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t \tag{8}$$

In the formula, y_t is the current value, μ is the constant, p is the order-number, γ_i is autocorrelation coefficient, ϵ_t is the error.

For MA, it focuses on the accumulation of error terms in the autoregressive model. The formula definition of p-order autoregressive process is as **(9)**:

$$y_t = \mu + \epsilon_t + \Sigma_{i=1}^{q} \theta_i \epsilon_{t-i} \tag{9}$$

The notations have the same meaning as **(8)**.

ARMA is a combination of the AR model and the MA model. The formula definition of ARMA is as **(10)**:

$$y_t = \mu + \Sigma_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t + \Sigma_{i=1}^{q} \theta_i \epsilon_{t-i} \tag{10}$$

I is a differential model, and ARIMA is a differential ARMA model, which ensures the stability of the data. The formula definition of ARIMA is as **(11)**:

$$y_t' = \mu + \Sigma_{i=1}^{p} \gamma_i y_{t-i}' + \epsilon_t + \Sigma_{i=1}^{q} \theta_i \epsilon_{t-i} \tag{11}$$

where

$$y_t' = \Delta^d y_t = (1-L)^{\,d} y_t \tag{12}$$

p, d, q are three parameters of ARIMA.

5.1.2 Holt-Winters Model and The Prophet Model[2]

As the pie chart shows that people would like to play Wordle on Wednesday the most, we think that there would probably be some connection between the day of the week and the reported number.

To reflect this weekly variation, we chose Holt-winters and Prophet models to do it as they can reflect periodic changes.

For Holt-Winters, we will go from single exponential smoothing algorithm to triple exponential smoothing algorithm.

For single exponential smoothing algorithm, we have:

$$F_{t+1} = \alpha x_t + (1-\alpha) F_t \tag{13}$$

which means:

$$F_{t+1} = \alpha x_t + (1-\alpha)\alpha x_{t-1} + (1-\alpha)^2 \alpha x_{t-2} + (1-\alpha)^3 \alpha x_{t-3} + \cdots + (1-\alpha)^t F_1 \tag{14}$$

In the formula, F_{t+1} is the prediction value, x_t is the actual observed value, α is the weighted value.

For double exponential smoothing algorithm, we have:

$$\begin{cases} S_t^{(1)} = \alpha x_t + (1-\alpha) S_{t-1}^{(1)} \\ S_t^{(2)} = \alpha S_t^{(1)} + (1-\alpha) S_{t-1}^{(2)} \end{cases} \tag{15}$$

The prediction value is $x_{t+T} = A_T + B_T T, A_T = 2S_t^{(1)} - S_t^{(2)}$

For triple exponential smoothing algorithm, we have:

$$\begin{cases} S_t^{(1)} = \alpha x_t + (1-\alpha) S_{t-1}^{(1)} \\ S_t^{(2)} = \alpha S_t^{(1)} + (1-\alpha) S_{t-1}^{(2)} \\ S_t^{(3)} = \alpha S_t^{(2)} + (1-\alpha) S_{t-1}^{(3)} \end{cases} \tag{16}$$

The prediction value is $x_{t+T} = A_T + B_T T + C_T T^2$

For Prophet model, it uses a decomposable time series model consisting of trend and seasonality.

$$y(t) = g(t) + s(t) + \epsilon_t \tag{17}$$

In the formula, $y(t)$ is the items to be predicted, $g(t)$ is a trend function, representing the value of non-periodic changes, $s(t)$ indicates periodic changes (such as weekly and annual seasonality), ϵ_t is the error term.

Specifically, to calculate $g(x)$, we can use the following formula:

$$g(t) = \frac{C}{1 + \exp(-k(t-m))} \tag{18}$$

In the formula, C is the bearing capacity, k is the rate of accretion, m is the offset coefficient.

For $s(t)$,

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi n t}{P}\right) + b_n \sin\left(\frac{2\pi n t}{P}\right) \right) \tag{19}$$

In the formula, P represents the period.

By using Prophet Model, we can see the trend of the curve which determined by the date and the day of the week.

5.1.3 Results Analysis

We can see that there is a sudden rise in February. And after August the curve enters the flat period. So we use the data from 2022/1/7 to 2022/11/30 as the first modeling data and use the data from 2022/8/1 to 2022/11/30 as the second modeling data. Last, we use the data in December as the validation set.

By using the Prophet, we can get a prediction curve shown in **Figure 4**.

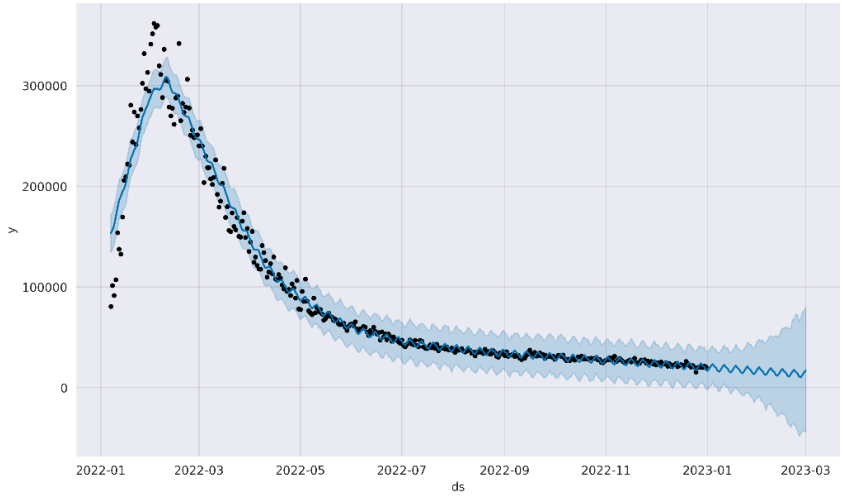


Figure 4: The prediction line by Prophet

Also, we could get a curve which would tell us the statistics vary because of the day of the week.

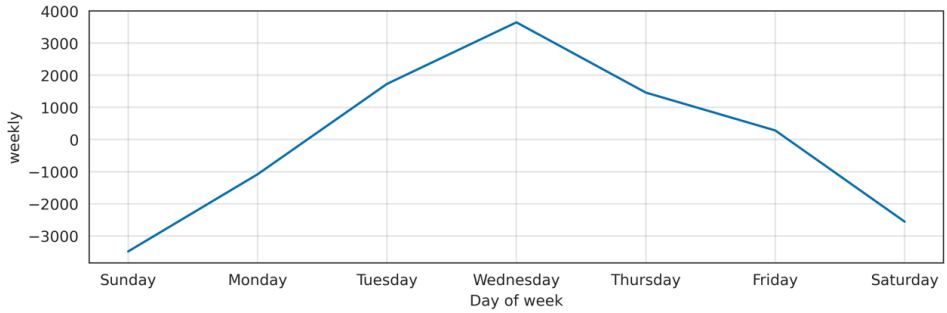


Figure 5: The prediction line by Prophet for the day of the week

According to these two pictures, we can explain the variation now. First, there is a general trend that the number of the players would go up and then go down. Second, there would be some differences caused by the day of the week. There would be more people playing Wordle on weekdays and people would like to play Wordle on Wednesday the most.

By using the ARIMA, we can get a prediction curve shown in **Figure 6**.

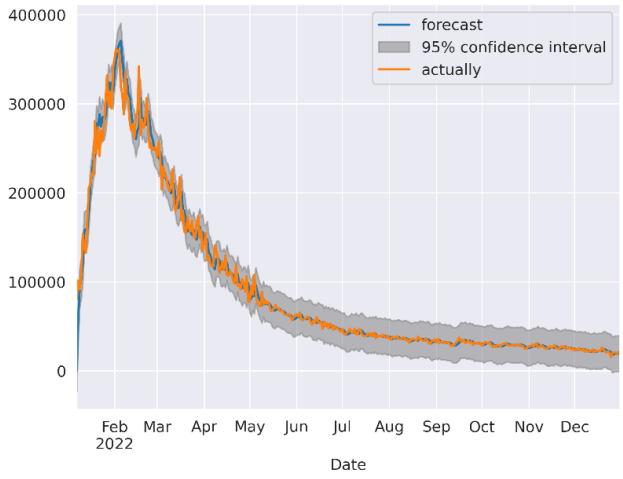


Figure 6: The prediction line compared with actual line by ARIMA

We can see that in this image, the prediction data are very close to the actual data. The actual data are always in the 95% confidence interval except some points. So we can consider our estimation to be quite appropriate.

Additionally, we used two other methods to do this prediction job. They are Holt-Winter and Modified exponential curve. To compare these three methods, we used the Mean Absolute Error, Mean Squared Error and Mean Absolute Percentage Error to measure the approximation of the curve. The result is in the **Table 1** below.

Table 1: Comparison among three prediction models

Models	MAE	MSE	MAPE
ARIMA	1321.8783	3118803.1233	6.0942
Holt-Winter	1305.9844	3201299.5985	5.9624
Prophet	**1183.5048**	**2838605.2470**	**5.4842**

Compared with Holt-Winter and ARIMA, Prophet has the smallest MAE, MSE and MAPE which indicates that Prophet has the best prediction accuracy.

By using the Prophet and the whole data set, we predict the number of reported results on March 1, 2023 would be in the interval(10627, 29873), the expected results would be 25751.9. Apparently, this could be a bad interval because the interval is too large. We think this is because the data before August has a strong impact on the final result.

We can see that the curve tend to be stable after August, so we chose to use the data after August 1, 2023 to do the prediction. This time we predict the number of reported results on March 1, 2023 would be in the interval(15293, 20940), the expected results would be 17851.3.

5.2 The Establishment and Solution of Hard Mode Model

At first we will calculate the percentage of hard mode and place them in a chronological order. **Figure 7** shows the whole trend.

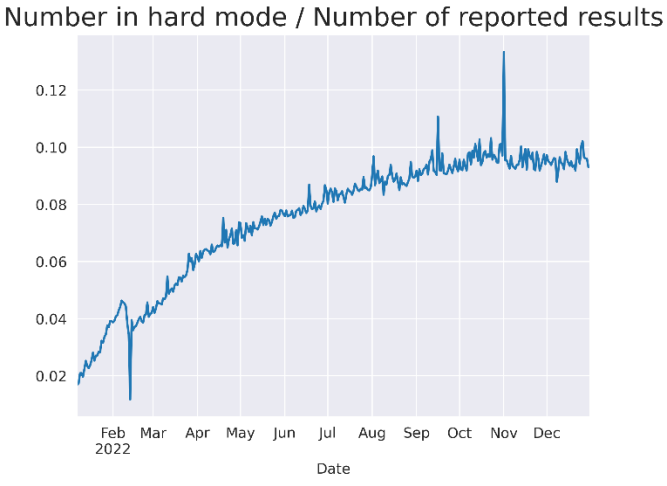


Figure 7: Percentage of hard mode over time

We can see that the percentage of the hard mode gradually grows over time which means the players are more likely to play hard mode over time. We give a conjecture that the people who play it for fun leave and the remaining players would like to challenge themselves more. To analyze the relationship between the word and the hard mode percentage, we shall remove

the influence of the time at first. We use the partial correlation coefficient method to evaluate the influence of word attributes. Also, we found that there is a smooth zone after October so we use the data after October to do some research.

5.2.1 Correlation Coefficient Method

We use the first-order partial correlation coefficient to evaluate. Here is the formula for First-order partial correlation coefficient.

$$r_{ij\cdot h} = \frac{r_{ij} - r_{ih}r_{jh}}{\sqrt{(1-r_{ih}^2)\,(1-r_{jh}^2)}} \tag{20}$$

In the formula, r_{ij} is the simple correlation between variable x_i and x_j, r_{ih} is the simple correlation between variable x_i and x_h, r_{jh} is the simple correlation between variable x_j and x_h.

Null hypothesis of partial correlation coefficient test is that in the population the partial correlation coefficient between the two variables is 0. The hypothesis test formula is as followings by using the t test method.

$$t = \frac{\sqrt{n-k-2}\cdot r}{\sqrt{1-r^2}} \tag{21}$$

In the formula, r is the corresponding partial correlation coefficient, n is the number of sample, k is the number of controllable variables, $n-k-2$ is the degree of freedom. When $t > t_{0.05}(n-k-2)$ or $p < 0.05$, we reject the null hypothesis.

5.2.2 Word Attribution

We extract features from words to find some attributions of words, so that we can analyze them.

In general, the difficulty of a solution word can be influenced by a combination of factors, including its length, complexity, rarity, and the amount of specialized knowledge required to solve it.

Some have set out to determine what exactly makes a Wordle word difficult. Many players point out that words with two occurrences of the same letter are particularly difficult to guess. Forbes observes that a surprising number of the most difficult words begin with 'F,' an uncommon starting letter. WordFinder highlights two characteristics about difficult words: they are obscure or do not follow predictable spelling patterns. In contrast, linguistics professor Dr. Matthew Voice points out that words with especially common n-grams (when letters are organized in common patterns) can be some of the hardest Wordles to solve.

We predict that obscurity, the structure of the word, and hints will all have a noticeable impact on how many guesses people need to solve a Wordle. We believe that obscurity and word structure will have a significant effect, because players will be less familiar with obscure words, requiring them to make more guesses. Additionally, since players typically use familiar spelling patterns to make better guesses, we expect that words with letters in an unusual order will also be more difficult to guess.

We consider the impact of different aspects which have been proposed in section **4.2**.

5.2.3 Results Analysis

Considering the influence of word characteristics, we use the whole data set as the modeling set for the partial regression analysis and use the data from 2022/10/1 to 2022/12/30 as the modeling set for regression analysis.

We did OLS regression on Per(percentage of Hard Mode) with all the attribute feature. As the results is too long, we store it in the **appendix 1**.

We can see that Per(the percentage of hard mode) has no obvious correlation with any attribution. We could only get a conclusion that none of the attributes of the word affect the percentage of scores reported that were played in Hard Mode. We consider this conclusion to be quite reasonable because you can only know what the word is after finishing playing Wordle. So the people who chose to play the Hard Mode is just because they wanted but not because some characteristics of the word to be guessed.

6 Distribution Prediction Model

First, we used the whole data set to analyze the whole characteristics.

We found that the distribution is quite likely to obey normal distribution. Also the skewness is always very small which is shown in **Figure 8**.

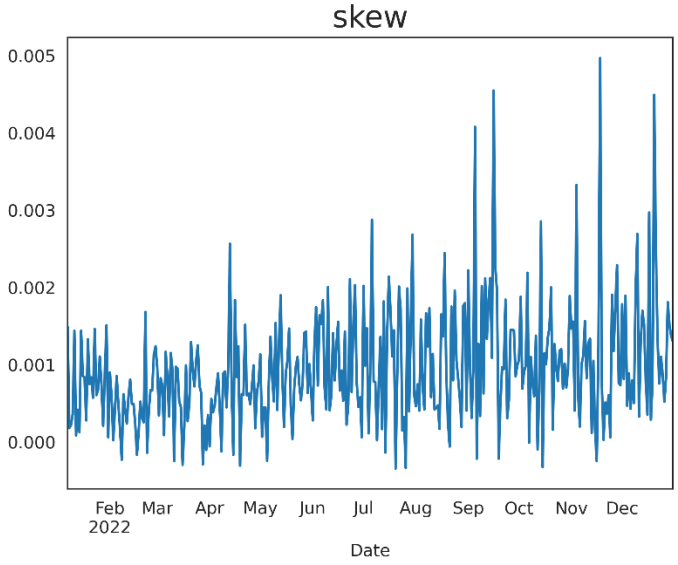


Figure 8: the skewness of the whole data

So we could assume that the distribution obey the unbiased normal distribution.

So if we can predict the expected value and variance, we can get to know the prediction of the distribution.

6.1 VARMAX

As described **above**, AR describes the relationship between the current value and the historical value, and uses the historical time data of the variable to predict itself. MA focuses on the accumulation of error terms in the autoregressive model. ARMA is a combination of the AR model and the MA model. The VARMA[8] method is an extension of ARMA to multiple parallel time series, such as multivariate time series. A finite order VAR process with a finite

order MA error[9] term is called VARMA. The VARMAX method is an extension of VARMA[10] which introduces exogenous variables.

6.2 Results Analysis

We use the data from 2022/8/1 to 2022 / 11 / 30 as the training set to do regression analysis to the mean and variance of the distribution over time. And the full set was used for prediction. Then we use lasso and other means to use the subsequent word characteristics to predict the distribution mean variance.

By using the multiple regression analysis model, we get the prediction of the expected value and variance. Then we get the prediction on the distribution of EERIE on March 1, 2023 as **Table 2**.

Table 2: The predicted distribution of EERIE on March 1, 2023

	1	2	3	4	5	6	X
percentage	0.00467	0.05846	0.22738	0.32940	0.23649	0.11555	0.02816

Of course there are some uncertainties in the predictions. There might be some people choose to cheat to get the correct answer. Also, the people like us who is taking part in this competition would like to play Wordle on March 1, 2023 which might make the percentage of 1try become higher if EERIE would be on March 1, 2023.

We used a lot of machine learning ways such as Ridge, Lasso[12] and so on to predict the distribution in order to evaluate the accuracy of our model. **Table 3** shows the result.

Table 3: The comparison between VARMAX with other machine learning ways

	Ridge	Lasso	Elastic Net[13]	SVM	KNeigh-bors[14]	Random-Forest[15]	**VARMAX**
R-square score	0.363	0.1677	0.16739	-0.8039	0.144	0.39324	**0.421**
1 try	-0.00189	0.00102	0.0013	-0.0385	-0.0024	0.006	**0.004511**
2 tries	0.00352	0.04631	0.0451	0.013	0.0481	0.01699	**0.055874**
3 tries	0.10804	0.21654	0.2316	0.214	0.254	0.0991	**0.22093**
4 tries	0.3159	0.3953	0.3721	0.4145	0.3746	0.3063	**0.327423**
5 tries	0.3259	0.2632	0.2645	0.2314	0.2293	0.34839	**0.240982**
6 tries	0.1919	0.0943	0.0830	0.0321	0.1046	0.1911	**0.120322**
7 or more tries	-0.30637	-0.01667	0.0024	0.1335	-0.0082	0.03212	**0.029958**

Our model defeats all the machine learning ways above. So we are quite confident with our in our model's prediction.

7 Word Classification Model

In the previous studies, we have already designed four evaluating indicators. Now we'll use the Expect_tries, P_Win and DR to do some further research because they can reflect the difficulty of the word.

7.1 K-means[16]

Different from classification, sequence labeling and other tasks, clustering is to divide samples into several categories through the intrinsic relationship between data without knowing any sample labels in advance, so that the similarity between samples in the same category is high, and the similarity between samples in different categories is low (that is, increasing class cohesion and reducing class spacing).

Clustering belongs to unsupervised learning, and K-means clustering is the most basic and commonly used clustering algorithm. Its basic idea is to find a partition scheme of K clusters by iteration, so that the loss function corresponding to the clustering result is minimized. The loss function can be defined as the sum of squared errors of the distance between each sample and the cluster center:

$$J(c,\mu)=\Sigma_{i=1}^{M}||x_i-\mu_{c_i}||^2 \tag{22}$$

In the formula, x_i represents the i-th sample, c_i is the cluster which x_i belongs to, μ_{c_i} represents the cluster corresponding to the center point of c_i, M is the total number of the samples.

To use K-means, we need to select K clusters as $\mu_0,\mu_1\cdots\mu_K$.

Then use the $J(c,\mu)$ as the loss function.

For each sample, it is assigned to the nearest cluster.

At last, calculated the center point again until $J(c,\mu)$ convergence.

7.2 Support Vector Machine

Support vector machine (SVM) is a kind of generalized linear classifier that classifies data by supervised learning. The decision boundary is the maximum margin hyperplane for solving the learning samples. Kernel functions are often used in the calculation process to map low-dimensional data to high-dimensional space, so that the results that were originally linearly inseparable in low-dimensional space are linearly separable in high-dimensional space. In this way, we can solve an interface that makes the two types of samples farthest from the boundary hyperplane, so that we can have higher accuracy when judging unknown data.

To use the SVM, we need to set training samples as $\{x_i,y_i|i=1,2,...,N\}$ where x_i is the eigenvector of the sample and y_i is the type of the sample, N is the number of the samples.

The sample will satisfy the following formula:

$$\begin{cases} w^T\phi(x_i)+b>0 \ when \ y_i=+1 \\ w^T\phi(x_i)+b<0 \ when \ y_i=-1 \end{cases} \tag{23}$$

Here w and b are a set of vectors, which together establish a hyperplane $y=w^Tx+b$ as the boundary basis, and $\phi(\ )$ is the rb kernel function, which can map the low-dimensional data to the high-dimensional space, so that we can The data that is linearly inseparable in the low-dimensional space can be separated in the high-dimensional space to meet the requirements of the SVM model.

After regularization we can have:

$$\begin{cases} w^T\phi(x_i)+b \geqslant 1 \ when \ y_i=+1 \\ w^T\phi(x_i)+b \leqslant -1 \ when \ y_i=-1 \end{cases} \tag{24}$$

It's just like the **Figure 9:**

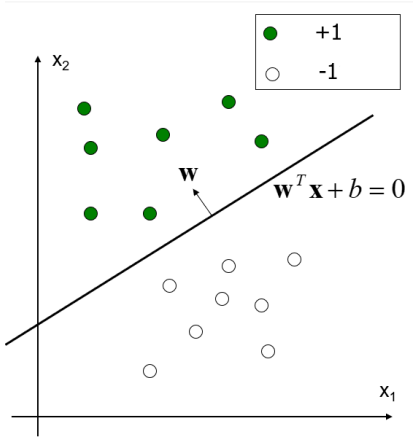


Figure 9: SVM classification diagram

At this time we can obtain the interval between the two lines is $\frac{2}{||w||}$.

We require the solution to be the maximum interval under the above conditions:

$$\max \frac{2}{||w||}$$
$$\begin{cases} w^T\phi(x_i)+b \geqslant 1 \ when \ y_i=+1 \\ w^T\phi(x_i)+b \leqslant -1 \ when \ y_i=-1 \end{cases} \tag{25}$$

Then we can get a nice line to separate samples of different categories.

7.3 Results Analysis

By using the K-means, we can get the results **Figure 10**. The Silhouette Coefficient is 0.52194.

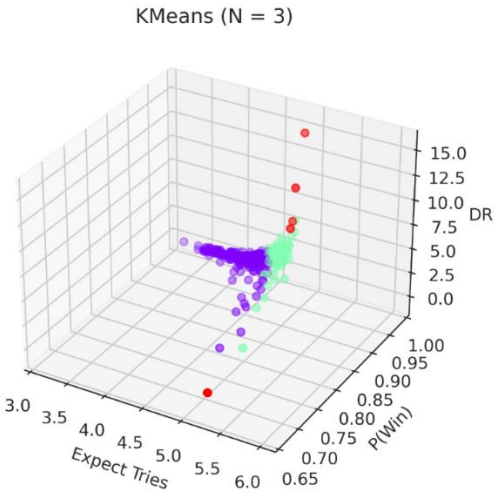


Figure 10: Graphic model by using K-means

So the solution words can be divided into three types : low difficulty, medium difficulty and high difficulty. We think the Expected tries of a word, the percentage of success and coefficient DR(designed by us in formula **7**) are three most important attributes that are associated

to the classification. Then we combined DR, P_Win and Expect tries into Mixed Difficulty(MD).

By using the SVM model while using the coefficient word_length, Gini coefficient, PCF which has been introduced in session **4.2**, we could get the Graphic model showed in **Figure 11**.

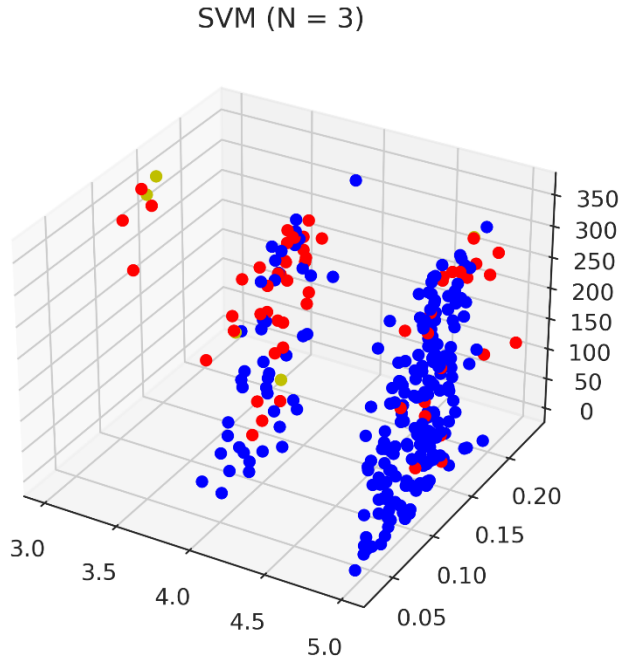


Figure 11: Graphic model by using SVM

By using our model, EERIE belongs to the medium difficulty group.

We divide the data set into training set and test set by 9 : 1 and compare it with Random forest and AdaBoost for evaluation :

Models	**Accuracy**	**Precision**	**Recall**	**F1**
SVM(ours)	**0.74**	**0.6314**	**0.5576**	**0.5751**
Random forest	0.62	0.5193	0.5897	0.5233
AdaBoost	0.53	0.5576	0.5577	0.5576

SVM gets the best effect.

8 Additional Interesting Features

We found that most people would like to play Wordle on weekdays, especially on Tuesday, Wednesday, Thursday. But we assume that there would be more people play Wordle on weekends. This might because in America people have to work in office from Tuesday to Thursday so that they may play Wordle when the bosses don't notice.

Although many players have claimed that the difficulty becomes more and more difficult after New York Times bought it. But after our study about this data set we found that the difficulty does not change significantly. People might just have some prejudice on New York Times.

9 Model Evaluation and Further Discussion

9.1 Strengths

1.We innovatively used the Gini coefficient to describe the words and get a good result which indicates a relation ship between its expected tries and its Gini coefficient.

2.We creatively design an index named Pattern Compare Frequency because of the Wordle's characteristics.

3.We noticed there would be a weekly change rule combined with the whole date change rule. So we use the Prophet to do the prediction which could explain the variation more accurately.

4.To make the results more reliable, we always use multiple models to solve the problems.

9.2 Weaknesses

1.To make the problems easier to establish models to analyze we make some assumptions which could lead to some slight difference with the real situation.

2.In problem 2, we ignore the skewness because the skewness is quite small. But it will do some influence to the final results.

10 Conclusion

After doing research on the variation, we found that Wordle surely became a hit last year. Like any other internet celebrity products, the reported number has been declining since March, 2022. Fortunately, Wordle seem to have gained a group of loyal fans, they insist on playing this game so that the curve of the players tend to be stable. Also, it seems that the number of the people wants to challenge themselves with the hard mode has no relationship with the word to be guessed. It's quite easy to understand because you can only know the solution word after finishing the game when you have already made a decision that you will enjoy hard mode or not.

11 A Letter to the Puzzle Editor of the New York Times

Dear editors,

We hope everything goes well.

Last year, we were obsessed with the Wordle. We would look forward to the update of the Wordle every day and try our best to solve it in less steps. We try to start the game in a good way in order to get the correct answer easily. So we did a lot to get a better start words. During this period, we became more and more interested in knowing about this fantastic game, so we did some research about this game.

As we all know that DAU(daily active user) is quite important for a game. So we first did some research on this aspect. We found that the reported number of the players has reached its peak in the February, 2023. At that time, more than 350,000 people played this game together! But after that, the reported number has kept declining. Recently, the number goes down to near 20,000. While the fortunate thing is that Wordle has a group of loyal fans who insist on playing Wordle and would like to share their results every day. Maybe you can do some interesting changes on the game to attract some new fans and hold on to loyal fans. Also we found that the percentage of people who want to play in Hard Mode is quite low. Maybe you can encourage more people to play Hard Mode when the word is not so hard to guess.

Because we always wanted to solve the puzzle in least steps, we care about the distribution of different times of tries a lot. Most people can guess the word with 4 times. This is quite suitable. We heard that a lot people said that the difficulty has increased. But there is no clear evidence to prove it. Maybe you can give the people who can't figure out the answer after 4 tries a little more tips to let them to figure out the words successfully. We think in this way more new players would like to enjoy the Wordle.

Lastly we divided the words into three groups which are low difficulty, medium difficulty and high difficulty. Most of the words are in the low difficulty. Of course the difficulty is relative. We think this kind of arrangement could make more normal players to enjoy the game. But as We discovered that you have a group of loyal players who really like playing this kind of games. Maybe you could add different words to the Wordle. If you correctly guess the easier word, you can keep guessing a harder one. We believe this could make some loyal fans more likely to enjoy the harder game and new players could also enjoy the excitement after solving a problem.

We are three loyal Wordle fans and would like to make the game more dynamic. But the fact is that the number of the people who enjoy this game become less and less day by day. We are quite distressed. We believe there could be some interesting changes which could attract more and more people to play this.

We wish Wordle could be a hit again! If you need more information, please feel free to write to us. Looking forward to a better Wordle!

Yours Sincerely,
Team # 2313988

References

[1]Taylor SJ, Letham B. 2017. Forecasting at scale. PeerJ Preprints 5:e3190v2 https://doi.org/10.7287/peerj.preprints.3190v2
[2]Box G E P, Pierce D A. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models[J]. Journal of the American statistical Association, 1970, 65(332): 1509-1526.
[3]https://www.washingtonpost.com/media/2022/01/31/new-york-times-wordle/
[4]https://www.today.com/popculture/popculture/is-wordle-getting-harder-fans-questions-nytimes-rcna16623
[5]https://wordfinder.yourdictionary.com/blog/the-hardest-wordle-puzzles-to-date-what-does-tacit-even-mean/
[6] https://www.techradar.com/news/why-todays-wordle-answer-is-so-hard-according-to-the-experts
[7] https://reference.wolfram.com/language/ref/WordFrequencyData.html
[8]K.Lakshmi,J.Prawin.Decentralized damage diagnostic technique for tall buildings using VARMAX model[J].Earthquake Engineering and Engineering Vibration,2022,21(02):417-439.
[9]https://en.wikipedia.org/wiki/Mean_absolute_error
[10]Greg Hundley,Sergio Koreisha. The specification of econometric strike models: A VARMA approach[J]. Applied Economics,1987,19(4).
[11] Hastie, T., Friedman, J., & Tisbshirani, R. (2017).The Elements of statistical learning: data mining, inference, and prediction.
[12] Santosa, Fadil; Symes, William W. (1986). "Linear inversion of band-limited reflection seismograms". SIAM Journal on Scientific and Statistical Computing. SIAM. 7 (4): 1307–1330. doi:10.1137/0907087
[13] Zou, Hui; Hastie, Trevor (2005). "Regularization and Variable Selection via the Elastic Net". Journal of the Royal Statistical Society, Series B. 67 (2): 301–320. CiteSeerX 10.1.1.124.4696
[14] Cover, Thomas M.; Hart, Peter E. (1967). "Nearest neighbor pattern classification" (PDF). IEEE Transactions on Information Theory. 13 (1): 21–27. CiteSeerX 10.1.1.68.2616
[15] Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282. Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016.
[16] Kriegel, Hans-Peter; Schubert, Erich; Zimek, Arthur (2016). "The (black) art of runtime evaluation: Are we comparing algorithms or implementations?". Knowledge and Information Systems. 52 (2): 341–378. doi:10.1007/s10115-016-1004-2. ISSN 0219-1377. S2CID 40772241

Appendices

Appendix 1

Introduce: The tables of OLS regression

	Word_frequency			The number of consonants			Longest continuous consonant		
	R^2	AIC	RIC	R^2	AIC	RIC	R^2	AIC	RIC
Per	0.012	-1715	-1707	0.006	-1713	-1705	0.002	-1711	-1703
Ex-pect_tries	0.025	346.2	354	0.003	354.3	362.1	0.001	354.9	362.7
P_Win	0.003	-1269	-1261	0.004	-1269	-1262	0.001	-1268	-1260
DR	0.012	1453	1461	0.006	671.9	679.7	0.003	1456	1464
MD	0.010	419.7	427.5	0.005	421.4	429.1	0.000	423.0	430.8

	Word_length			Begin with a vowel or not			Part-of-speech		
	R^2	AIC	RIC	R^2	AIC	RIC	R^2	AIC	RIC
Per	0.006	-1713	-1705	0.014	-1715	-1707	0.026	-1702	-1659
Ex-pect_tries	0.165	290.6	298.4	0.001	355	362.8	0.028	363.2	405.9
P_Win	0.038	-1282	-1274	0.009	-1271	-1264	0.011	-1254	-1211
DR	0.209	428.6	436.3	0.003	1456	1464	0.017	1470	1512
MD	0.194	345.6	353.3	0.001	422.7	430.5	0.070	397.1	404.8

	compatibility ratio			Log_frequency			gini coefficient		
	R^2	AIC	RIC	R^2	AIC	RIC	R^2	AIC	RIC
Per	0.021	-1718	-1710	0.012	-1714	-1707	0.019	-1717	-1709
Ex-pect_tries	0.184	18.64	21.44	0.025	-5176	-5169	0.45	6.822	9.624
P_Win	0.004	-1269	-1262	0.023	1488	1496	0.002	-1269	-1261
DR	0.029	1447	1445	0.012	-5172	-5164	0.068	1432	1440
MD	0.027	413.4	421.2	0.230	22.08	24.88	0.035	410.3	418.1

	Frequency_rank		
	R^2	AIC	RIC
Per	0.009	-1713	-1706
Ex-pect_tries	0.089	197.3	204.3
P_Win	0.017	-1274	-1267
DR	0.058	90.18	92.98
MD	0.123	83.26	88.23