Problem Chosen
C

2023
MCM/ICM
Summary Sheet

Team Control Number
2322161

Predicting Wordle Results

Summary

Wordle is a popular puzzle game on the Internet. It receives a lot of attention and raises a lot of discussions on Twitter. In this article, we analyze the daily report results from Twitter and predict future user volume and gaming results.

In Model I, we establish a mathematical model describing the population growth of the game. We generate a Mathematical Time Series Wordle Prediction Model by combining the Logarithmic Normal and Logistic Growth Models. Using the model, we successfully predicted the number of reported results on March 1, 2023.

We construct a Monte Carlo Simulation Model in Model II to simulate the Wordle process. This model contains three algorithms responsible for judgment, selection, and processing. Using this model, we obtain a prediction dataset for all the words that appear in the reported data. The prediction dataset shows homogeneity with the reported data. Then we use the model to predict the word "EERIE".

In Model III, using the K-Means and Gaussian Mixture Model, we build a clustering model to determine the best number of clusters and classify the difficulty into three levels. By comparison, K-Means Model offers a better accuracy on the samples. We can predict the difficulty using our model and the generated prediction distribution from Model II of the word "EERIE".

Finally, we conclude and write a letter to the Puzzle Editor to report our models and corresponding answers to the questions asked by the New York Times.

Keywords: Logarithmic Normal Model; Logistic Growth Model; Monte Carlo Simulation Model; Random Forest Classification Model; K-Means Clustering Model; Gaussian Mixture Model;

Contents

1 Introduction 3

1.1 Problem Background . . . 3

1.2 Restatement of the Problem . . . 4

2 Assumptions and Justifications 5

3 Data Processing 5

3.1 Data Acquisition . . . 5

3.2 Data Preprocessing . . . 6

4 Notations 6

5 Model I 7

5.1 Logarithmic Normal and Logistic Growth Model . . . 7

5.1.1 Logarithmic Normal Model . . . 7

5.1.2 Logistic Growth Model . . . 7

5.2 Mathematical Time Series Wordle Prediction Model . . . 8

5.2.1 Model Components . . . 8

5.2.2 Model Analysis . . . 8

6 Model II 10

6.1 Limitations Based on the Reported Data . . . 10

6.2 Logic of Wordle . . . 11

6.3 Monto Carlo Process for Simulating the Wordle Process . . . 11

6.3.1 Data Preprocessing . . . 11

6.3.2 Algorithm . . . 11

6.3.3 Adjustment on Data and Algorithm . . . 12

6.4 Model Assessment . . . 12

6.4.1 Random Forest Classification . . . 12

6.4.2 Error Estimation . . . 13

7 Model III 13

7.1 Number of Difficulty Levels . . . 14

7.2 K-Means and GMM Model . . . 14

7.3 Model Accuracy . . . 15

8 Conclusion 18

8.1 Solution . . . 18

8.1.1 Task 1 . . . 18

8.1.2 Task 2 . . . 18

8.1.3 Task 3 . . . 18

8.1.4 Task 4 . . . 18

8.2 Overall Evaluation . . . 19

8.2.1 Strengths . . . 19

8.2.2 Weaknesses 19
8.2.3 Future Work 19

9 A Letter to the Puzzle Editor 20

References 21

1 Introduction

1.1 Problem Background

Wordle is an online puzzle provided by the New York Times. The popular game is updated daily and requires players to fill in a five-letter word puzzle with a maximum of six opportunities each day. Three colors denote denote the states of the game process. For instance, a green tile indicates the letter and word are in the correct position. A yellow tile shows that the letter belongs to the word but is in the wrong position. And a gray tile indicates that the letter isn't in the word. Figure1 shows the corresponding color tiles. Moreover, a "Hard Mode" increased the difficulty by forcing players to include their correct guesses (tiles in green or yellow) in the next tries.

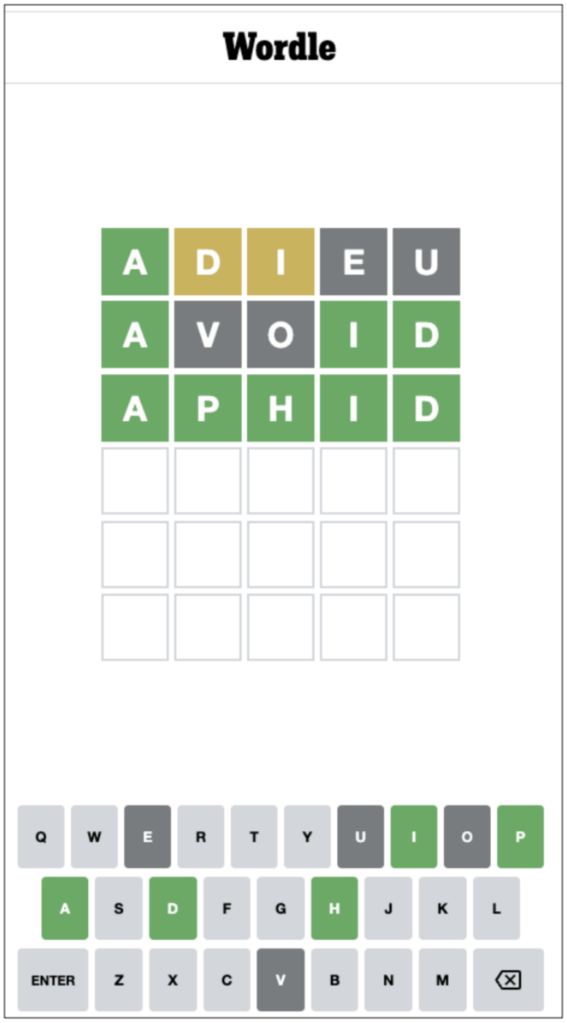
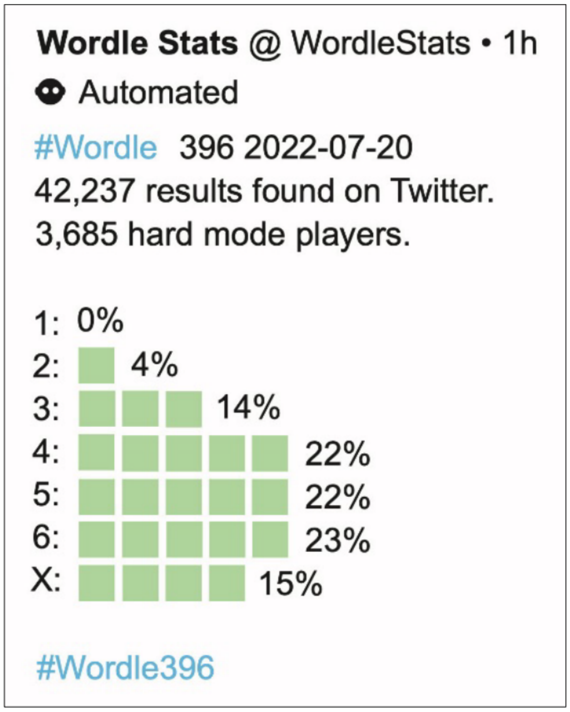


Figure 1: Wordle Game Interface

Many users are sharing their scores on a social media named Twitter. Figure 2 is an example. Daily reports from January 7, 2022 to December 31, 2022 are given, which includes the date, contest number, word of the day, the number of people reporting scores that day, the number of players on hard mode, and the percentage that guessed the word from one to six tries, or could not solve the puzzle (denoted by X).



Wordle Stats @ WordleStats • 1h

Automated

#Wordle 396 2022-07-20
42,237 results found on Twitter.
3,685 hard mode players.

1: 0%
2: 🟩 4%
3: 🟩🟩🟩 14%
4: 🟩🟩🟩🟩🟩 22%
5: 🟩🟩🟩🟩🟩 22%
6: 🟩🟩🟩🟩🟩 23%
X: 🟩🟩🟩🟩 15%

#Wordle396

Figure 2: Twitter Report Example

1.2 Restatement of the Problem

According to the data provided by MCM and information identified in the problem statement, we need to establish several models and complete the following tasks with the models:

- Develop a model to explain the daily variation of the report results and create a prediction interval on March 1, 2023. Also, determine and explain whether the attributes of the word affect the percentage of scores reported that were played in Hard Mode.
- Predict the distribution of the report results based on a model and a given target word EERIE on March 1, 2023. The analysis gives the uncertainties and confidence of the model's prediction results.
- Summarize a model to make word classification based on their difficulty, using the attributes of a given word "EERIE" for classification. Meanwhile, discuss the accuracy of the classification.
- Display and explain some other interesting features of the report results dataset.
- Design a one- to two-page letter to the Puzzle Editor of the New York Times to summarize the results.

2 Assumptions and Justifications

Assumptions are made as follows to simply the problem, each of which is appropriately justified.

- **Assumption 1:** Assuming the players are rational and devoted, they will try their best to solve the puzzle based on the condition.

  Justification: Rational users tend to take a predictable and feasible approach to solving the Wordle puzzle, which means their gaming pattern can be tracked and modeled. We need to add this restriction to better fit our model into the dataset to simplify the circumstance.

- **Assumption 2:** Assuming the players are honest, the report results are highly credible.

  Justification: Although fake information on social media exists in the real world, the problem requires us to generate a prediction model based solely on the information provided. With this assumption, we can analyze the results correctly.

- **Assumption 3:** Assuming users' average level (e.g. vocabulary) doesn't vary significantlyday day by day.

  Justification: The game's results vary significantly according to the users' vocabulary. We must impose such a restriction to eliminate the influence of generating a simplified and predictable model.

- **Assumption 4:** Assuming a "Stable User Volume", which denotes a user group that will not change significantly concerning the total user volume.

  Justification: There is a terminology called user stickiness that compares active users' engagement over a narrower time frame with their attention over a broader time frame. It denotes the ability to retain users over time of software. In the real world, a whole user group with "stable users" is always active daily to support the user stickiness theory.

- **Assumption 5:** Assuming that players will feel more complicated if the word is harder to guess, which results in more significant average steps to think of the solution.

  Justification: If some attributes of a word make it harder to guess, more people will spend more steps thinking about the answer, resulting in more significant average steps.

3 Data Processing

3.1 Data Acquisition

We included all the words that appeared on Wordle, totally $12,974$ words, in seek of the overall word pattern and potential difficulty behind the original game source data.

Despite the data provided by MCM and Wordle, we also downloaded English Word Frequency (EWF) dataset from Kaggle[1]. EWF contains $333,333$ most frequent English words on the Internet and as derived from the Google Web Trillion Word Corpus. This dataset allows us to better analyze the difficulty of a word by taking its frequency on the Internet as a feature.

3.2 Data Preprocessing

- **MCM's Dataset:** For the dataset provided by MCM, we found several errors which influenced the overall performance of our model. To achieve data cleaning and ensure our model's correctness, we modified the source data. Detailed information is addressed in Table 1.

Position	Original Data	Modified Data
D18	rprobe	probe
D38	clen	clean
D90	marxh	marsh
D249	tash	trash
D356	favor [1]	favor
E34	2569	25690[2]

Table 1: Error Correctness of MCM's Dataset

- **Wordle & EWF Dataset:** Since there are many non-five-letter words in EWF dataset, we first filter the satisfied word out and make a crossing with the letters in the Wordle dataset. After the operation, we access 8092 words with their frequency on the Internet.

4 Notations

Notations are shown in Table 2.

Symbol	Description
μ_0	Expected value of the Logarithmic Normal Model
σ_0	Standard deviation of the Logarithmic Normal Model
RMSE_0	Root Mean Squared Error of the Logarithmic Normal Model
μ_1	Expected value of the Mathematical Time Series Wordle Prediction Model
σ_1	Standard deviation of the Mathematical Time Series Wordle Prediction Model
RMSE_1	Root Mean Squared Error of the Mathematical Time Series Wordle Prediction Model
$P(0)$	Initial population of the Logistic Growth Model
M	Maximum population of the Logistic Growth Model
R	Growth rate of the Logistic Growth Model
k	Gradient of the Linear Model
x_0	X axis intercept of the Linear Model

Table 2: Parameter and RMSE of Logarithmic Normal Model

[1] The word "favor" here is followed by a blank space.

[2] we added a 0 for correcting the entry error. An error in 10 units doesn't influence the prediction results of our models.

5 Model I

In this section, we proposed a mathematical model to explain the variation of the reported results and predict future outcomes. We derived a prediction model combining the two based on the logarithmic normal distribution and logistic growth model.

5.1 Logarithmic Normal and Logistic Growth Model

5.1.1 Logarithmic Normal Model

Let Z be a standard normal variable, and let μ and $\sigma > 0$ be two real numbers. The distribution follows

$$X = e^{\mu+\sigma Z}$$

is called the log-normal distribution with parameters μ and σ.

To get the mean μ_X and variance σ_X^2, we have

$$\mu = \ln\left(\frac{\mu_X^2}{\sqrt{\mu_X^2+\sigma_X^2}}\right) \text{ and } \sigma^2 = \ln\left(1+\frac{\sigma_X^2}{\mu_X^2}\right).$$

The probability density function of the distribution can be denoted by

$$\begin{aligned} f_X(x) &= \frac{d}{dx}\Pr(X \le x) = \frac{d}{dx}\Pr(\ln X \le \ln x) = \frac{d}{dx}\Phi\left(\frac{\ln x-\mu}{\sigma}\right) \\ &= \phi\left(\frac{\ln x-\mu}{\sigma}\right)\frac{d}{dx}\left(\frac{\ln x-\mu}{\sigma}\right) = \phi\left(\frac{\ln x-\mu}{\sigma}\right)\frac{1}{\sigma x} \\ &= \frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right). \end{aligned}$$

5.1.2 Logistic Growth Model

A logistic function is a S-shaped curve (sigmoid curve) with equation

$$f(x) = \frac{L}{1+e^{-k(x-x_0)}},$$

where

x_0, the x value of the function's midpoint;
L, the supremum of the values of the function;
k, the logistic growth rate or steepness of the curve.

Specifically, we use the logistic function to set up a logistic growth model. We can mathematically model the logistic growth by using the equation

$$\frac{dP}{dt} = r(M-P)P = rMP - rP^2.$$

$$\frac{d^2P}{dt^2} = \frac{d}{dt}(rMP - rP^2) = rM\frac{dP}{dt} - 2rP\frac{dP}{dt} = r(M-2P)\frac{dP}{dt}.$$

Separation of variables:

$$\frac{dP(t)}{P(t)(M-P(t)} = \frac{1}{M}\left(\frac{dP(t)}{P(t)} + \frac{dP(t)}{M-P(t)}\right) = rdt.$$

Integration on both sides:

$$\ln\frac{P(t)}{P(0)} - \ln\frac{M-P(t)}{M-P(0)} = rMt.$$

Solving this equation, we can get the solution

$$P(t) = \frac{MP(0)}{P(0)+(M-P(0))e^{-rMt}}.$$

5.2 Mathematical Time Series Wordle Prediction Model

5.2.1 Model Components

We propose the Wordle Predict Model by combining the knowledge of the Logarithmic Normal and Logistic Growth Model. It has two main components: a log-normal distribution and an incremental denoting the "stable users" declared in Assumption 4.

5.2.2 Model Analysis

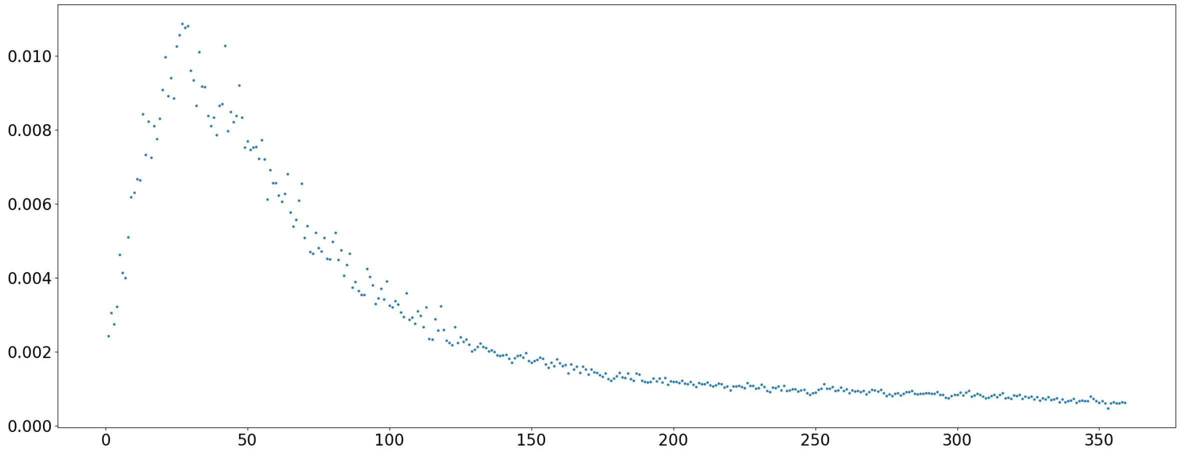


Figure 3: Original Data Scatter

Considering the real-world situation and the original data shown in Figure 3, we adopt the logarithmic normal distribution as the basis of our model. Figure 4 shows the initial solution using only the Logarithmic Normal Model. It is evident that with Logarithmic Normal Model, there are many errors when talking about report results 200 days after January 7, 2022. This model cannot successfully estimate the Wordle report results, especially the future results, which is asked in task one.

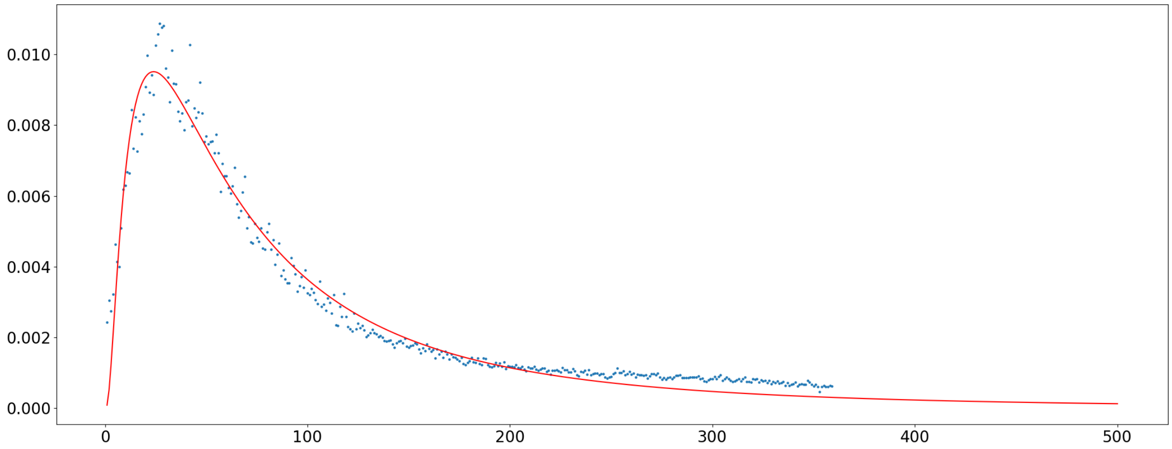


Figure 4: Fitting Result with Logarithmic Normal Model

In order to eliminate the prediction error, we adjust our model. Based on the knowledge that software has stable users, we add an incremental model consisting of a Logistic Growth Model and a Linear Model. We apply a Logistic Growth Model with initial population $P(0) = 100$, with temporary maximum stable user volume $M = 10000$, and a growth rate $R = 75$. Meanwhile, the Linear Model constantly decreases the stable user volume after 100 days beyond the start period of the dataset. Figure 5 displays the relevant result of our improved model, and the green line denotes the incremental component of the model. Table 3 and Table 4 show the parameters and Root Mean Squared Error (RMSE) of the Logarithmic Normal Model and our Mathematical Time Series Wordle Prediction Model. From the RMSE, we got relatively better results for predicting the report results.

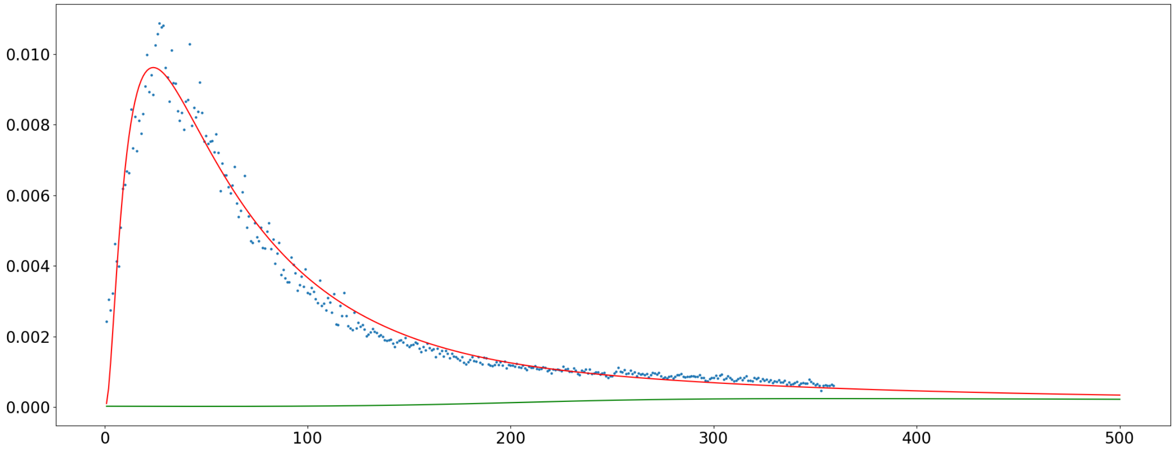


Figure 5: Fitting Result with Our Model

Parameter	Value
μ_0	4.236833315740967
σ_0	1.0293129975722861
RMSE_0	15369.368407498454

Table 3: Parameter and RMSE of Logarithmic Normal Model

Parameter	Value
μ_1	4.227476555132943
σ_1	1.024076057989859
$P(0)$	100
M	10000
R	75
k	0.0000002
x_0	100
RMSE_1	14227.475101826194

Table 4: Parameter and RMSE of Mathematical Time Series Wordle Prediction Model

6 Model II

The concrete embodiment of attributes of the words on the given data is the distribution of the seven groups, representing the number of trials players have done to solve the Wordle problem. The most significant attribute of a word is the order of letters. However, Wordle is not a single permutation and combination problem. Every trial should be a correct English word. Meanwhile, it is hard to quantify this attribute. Thus, the order of letters should not be the primary attribute for us to analyze words in Wordle. Despite the order of letters, the frequency is considered an excellent main attribute of a word to analyze in this game.

We first consider the K-Means Clustering Algorithm (K-Means) based on the Distribution of the Number of Trials (DTNT). Through the model, we classified the words into three different groups. The less guesswork, the easier the word. Then, we construct a Monto Carlo Simulation Model (MCSM) to simulate the process of a person playing Wordle based on the frequency of the word. By repeating the process many times, we could receive a particular distribution of the number of trials of a given the word. To examine the preciseness of the prediction model, we generate the distribution of the number of trials concerning each word that has already appeared in Wordle. From K-value clustering, we obtained the difficulty labeling each word and combined it with our prediction results of the DTNT for each word. We established a Random Forest Classification Model (RFCM). Finally, we calculate Root Mean Square Error (RMSE) to give a fundamental error analysis on our prediction of the word "EERIE".

6.1 Limitations Based on the Reported Data

We analyze the direct correlations between the DTNT and words' frequency. We build a K-means Clustering Model to classify the word according to its DTNT. When we choose k = 3, the clustering center of the three groups shows a significant difference. We rank the groups' difficulty into easy

(E), medium (M), and difficult (D), concerning a numerical rank as 1 (E), 2 (M), and 3 (D). The less guesswork, the easier the word. Then we try to do a regression on the frequency. However, the result shows that the regression model needs to work better. The p-value of each parameter is much greater than 0.05.

We summarized the failure and recognized that the Wordle process is a complex combination of independent events. Only the first guess can be viewed as a separate random case. The latter guesses are highly correlated with the former ones. Thus, the Wordle problem needs to be solved numerically and statistically. A logical model of the Wordle process is necessary.

6.2 Logic of Wordle

The first selection of the word is random for players, who receive instructional information in a color form. A yellow tile indicates the letter in that tile is in the word but in the wrong location. A green tile indicates that the letter in that tile is in the word and is in the correct location. A gray tile indicates that the letter in that tile is not included in the word. Then players will try another word within the range of words that satisfies the requirement that the instructional information has given. The same goes for steps behind. The process indicates that the times of guesswork depend not only on the words' frequency but also on the letter formation and order of other words. Thus, to precisely predict a word's DTNT, we built a logic model to simulate the detailed process for Wordle.

6.3 Monto Carlo Process for Simulating the Wordle Process

6.3.1 Data Preprocessing

We calculate the probability of selecting a word within a specific range of words by calculating its frequency as a proportion of all word frequencies as following:

$$P_i = \frac{\text{frequency}_i}{\text{total frequency of word range}}.$$

Then we construct a probability interval for each word as

$$\text{Interval}_i = (\sum_{j=1}^{i-1} P_j, \sum_{j=1}^{i} P_j].$$

6.3.2 Algorithm

We first generate seven variables to represent the number of samples with different trials. We construct a judgment algorithm that compares the selected and solution words and outputs the color-formed instructional information. We also construct a selection algorithm that can select words that strictly follow the instructional information from the judgment algorithm.

For each simulation, we do the coding following the algorithm below:

- Generate a uniform random variable x_1 within the interval $(0, 1]$.

- Select a word whose probability interval contains the value of x_1.
- If the selected word is the solution, add 1 on the number of samples with one try.
- If the first selected word is not the solution, while the guess word is not correct:
    - Select words from the latest range of words using the selection algorithm and update their probability intervals such that the summation of their probabilities equals 1.
    - Generate a uniform random variable x_2 within the interval $(0, 1]$.
    - Select a word whose probability interval contains the value of x_2.
    - Compare the selected word with the solution with the judgment algorithm.
- Record the number of cycles and add 1 to the variable

We allow the number of loops for each word to equal the number of people who participate on that word's date and calculate the DTNT of each word.

6.3.3 Adjustment on Data and Algorithm

In order to simulate a more realistic model, we consider some other behavioral factors based on players' habits when playing Wordle. For instance, players tend to guess a word that is more commonly used and contains more different letters at the first trial. We limit the first guess's word range to the first thousand commonly used words and overweight the words that contain five and four different letters in the first trial.

Also, the word's frequency cannot directly represent the probability. While the difference in frequency between a commonly-used word and a hardly-used word is too significant. For example, the frequency of the word ABOUT is 1226734006, and the word GAWKY is 31136. It is unrealistic to say that the probability of ABOUT is 39,399 times that of GAWKY. Thus, we try to limit the difference between probabilities to 50 times using the equation below:

$$\text{adjustedFrequency}_i = [\log(\text{frequency}_i)]^5$$

6.4 Model Assessment

6.4.1 Random Forest Classification

From the previous K-Means clustering, we have a classification of the words based on real-world data. We want to examine if the predicted data can follow the same classification when trained by itself. Thus, we use a Random Forest Classification Model (RFCM) to test the prediction data's accuracy in classification. RFCM can be directly accessed through the SPSSPRO[2], an online data analysis platform. We input the class generated by K-means as the nominal level variable and the predicted DTNT, which contains seven variables as a quantitative variable. The training results are shown in Figure 6 and Table 5.

The RFCM results show that the prediction data's classification is harmonious with the reported data, which means our simulation model performs well.

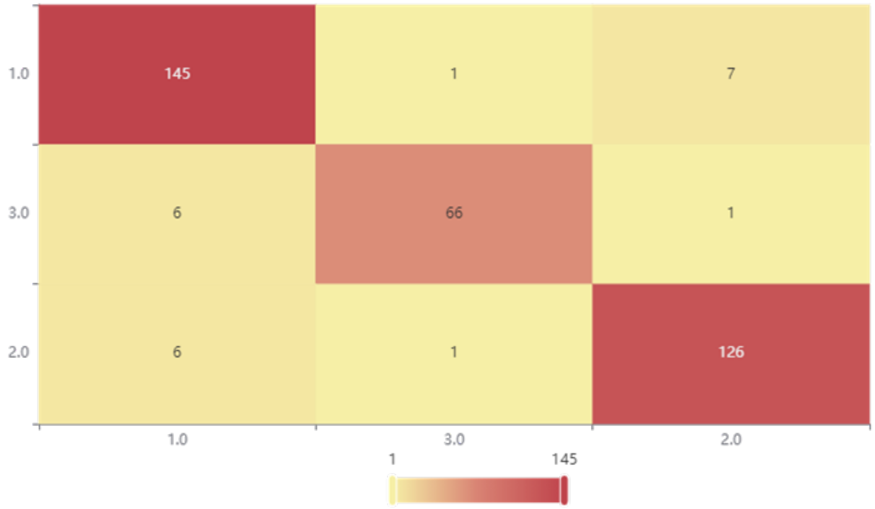


Figure 6: Confusion Matrix Heat Map

Accuracy Rate	**Recall Rate**
93.9%	93.9%

Table 5: Evaluation Result of Model II

6.4.2 Error Estimation

In order to give a more flexible prediction of future data on a specific word, we decided to construct data intervals for the predicted DTNT. In case we consider the word into three difficulty classes, by calculating the root of square error between the reported data and the predicted data, denoted by σ_{ij}, we give different intervals for words that belong to different classes.

$$\sigma_{ij} = \sqrt{\frac{\sum_{k=1}^{n} (\text{ReportedPercent}_{ijk} - \text{PredictedPercent}_{ij})^2}{n}},$$

$$\text{and Interval}_{ij} = (\text{prediction} - \sigma_{ij}, \text{prediction} + \sigma_{ij}).$$

In which i denotes the difficulty classes w.r.t. $i = 1(E), 2(M), 3(D)$, and j denotes the number of steps to guess the solution, which $j = 1, 2, ..., 7$.

7 Model III

In this section, we proposed a model to classify solution words by difficulty, using the K-Means clustering method and Gaussian Mixture Model (GMM) to make the classification. We compared and chose the better one of these two models to identify the word "EERIE"'s difficulty.

Intuitively speaking, the number of times a player tries can directly reflect how difficult the word is to guess. Since the number of attempts by the player is distributed as an integer between 1 and 6, the number of times greater than or equal to 7 will be counted as 7 times. So we can use clustering algorithms to classify the data from column G to M in the excel file.

7.1 Number of Difficulty Levels

Since too few difficulty levels mean that the model needs to be more accurate to classify the difficulty level, at the same time, too many difficulty levels may make the model over-fitting due to the small amount of data. So we used the Silhouette Coefficient and Rand Index to determine the best number of classification levels using the K-Means algorithm. As Figure 7 shows, it is observed that the best choice for it is 3, where both evaluation methods are relatively closer to 1. Therefore, We classified the difficulty into three levels: easy, medium, and difficult in the following part.

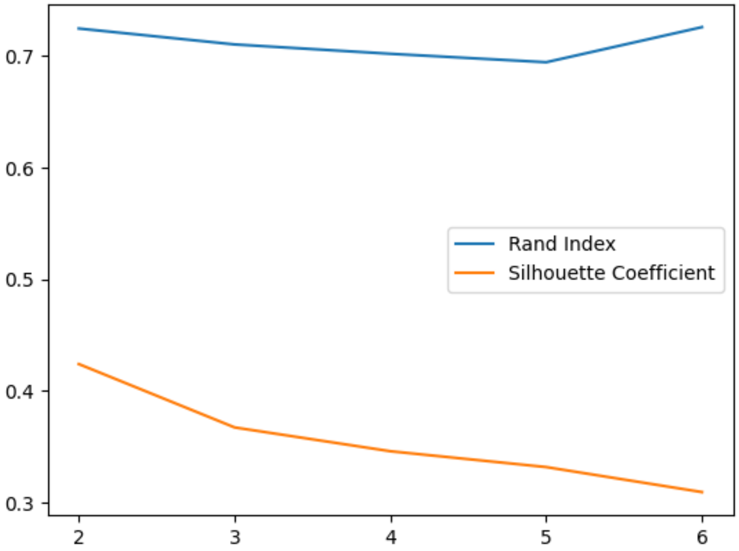


Figure 7: Silhouette Coefficient and Rand Index w.r.t. Number of Clusters

7.2 K-Means and GMM Model

Using Model II, the attributes of the provided word can be used to predict the distribution of the number of attempts for the original word. Hence we can make classifications on original data and new data generated from the trained model. The procedure for implementing K-Means Algorithm is shown as follows:

- Choose a specific K denoting the number of clusters, which was optimized to 3.
- Randomly put K feature vectors, named centroids, to the feature space.
- Compute the distance from each example x to each centroid c using the Euclidean distance.
- Label each example with the closest centroid ID.

- Then it becomes the optimization problem:

$$\min_r \sum_i^n \sum_k^K r_{ik}(x_i - c_k)^2, \text{ subject to } r \in \{0,1\}^{n\times K}, \sum_k^K r_{ik} = 1.$$

- Repeat the following steps:
  - For each centroid, calculate the average feature vector of the examples labeled. Assign the average feature vectors as the new centroid locations.
  - Recompute the distance from each example to the centroid, modify the assignments, and repeat until they are stable.
- Calculate the expectation of the distribution of the number of trials on each centroid.
- Assign the centroid with the largest expectation as “D”, medium as “M”, and the smallest as “E”

Compared with K-Means Model, the GMM Model gives a soft margin clustering on the samples. Figure 8 displays each data’s clustering result, using k-Means and GMM algorithms. From the figure, it is observed that both algorithms fit the original data well. However, the two difficulty levels share nearly the same tendency and value when the generated data is fit by the GMM method. Therefore, the K-Means algorithm better includes the whole structure. Hence K-means is used for predicting the difficulty levels.

7.3 Model Accuracy

Using Assumption 5we can measure the difficulty by the average of the trials to guess the solution. The more difficult the word, the more trials are needed. The classification result is shown in Figure9.

Table 6 shows the correspondence among colors of curves, average steps, and difficulty levels.

Difficulty	**Average Steps**	**Color**
Easy	3.8168856375037326	blue
Medium	4.256003684452924	green
Difficult	4.754900616860864	yellow

Table 6: Colors and Average Steps of Each Difficulty

Table 7 shows the Silhouette Coefficient and Rand Index. The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is $(b - a)/\max(a, b)$

Since Silhouette Coefficient ranges from -1 to 1, where 1 is the beautiful clustering, our model does not classify it well. However, it is acceptable because there are only six attempts. It is difficult for players to guess the correct answer in the first two trials, and four trials is a mode number. Besides, most players can guess it within six times. Therefore, the sample data of the three difficulty levels generally have the same trend, and the sample size is not large enough, which results in a relatively small Silhouette Coefficient.

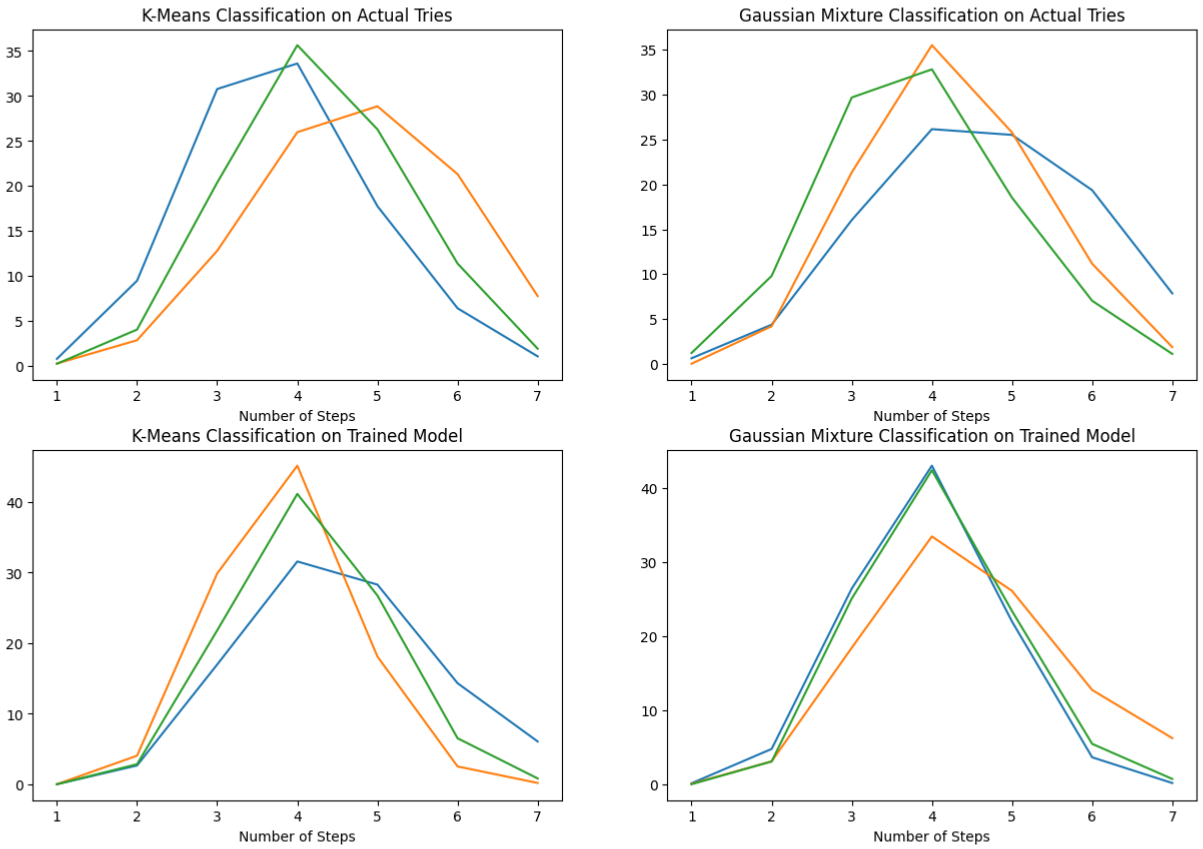


Figure 8: K-Means and GMM Comparison

Parameter	Value
Silhouette Coefficient	0.3673287439085649
Rand Index	0.710057422075598

Table 7: Model Accuracy measured by Silhouette Coefficient and Rand Index

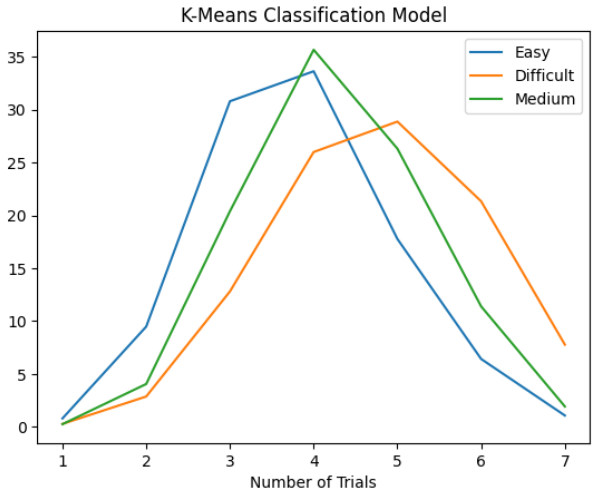


Figure 9: Difficulty of Three Centroids in the K-Means Model

8 Conclusion

8.1 Solution

8.1.1 Task 1

According to our Mathematical Time Series Wordle Prediction Model, we can predict the number of the reported result on March 1, 2023, which is 14377. Several word attributes affect the percentage of scores played in Hard Mode. For instance, with the decrease in word frequency, the rate of finishing the game with fewer tries will decrease. Also, particular prefixes and suffixes like ”ab-” or ”-st” will reduce the difficulty of the game, which leads to an increase in the percentage of finishing the game with fewer tries. Some high-frequency letters like ”a”, ”e” in words also influence the distribution of report results. However, the pattern of those letters has different effects on the results. For instance, ”mummy” is a problematic word in the game, although it has a relatively high word frequency.

8.1.2 Task 2

Our model's prediction shows that the proportion of players who can solve the Wordle problem in 1 trial is always less than 0.1%. However, the reported result shows that the proportion of many words' 1 trial success can reach 1% or even larger. We consider this bias caused by our assumption 2 that all players are honest, which may not be valid. Thus, our model may need to be more accurate in predicting on both 1 try and 7 or more trials. Moreover, as time passes, the left players are considered more experienced, which means their skills on Wordle may lead to smaller intermediate steps to guess the solution. However, it is a feature that is hard to quantify. We only consider a two-period comparison. The players might be cleverer in the future. The prediction results of "EERIE" is shown in Table 8:

	1 try	**2 tries**	**3 tries**	**4 tries**	**5 tries**	**6 tries**	**7 or more tries**
Result	0	1.9%	17.1%	40.5%	31.5%	7.5%	1.2%
STD	0.7	0.51	1.25	2.75	2.94	1.25	0.62

Table 8: Prediction of EERIE Report Distribution

8.1.3 Task 3

The best cluster number is 3, optimized using Rand Index and Silhouette Coefficient. Then we classify the difficulty into three levels: D, MD, and E. After comparing the accuracy of the K-Means and GMM models. We conclude that the K-Means model better fits the sample data. Using the cluster centers from the K-Means model and the above prediction distribution of "ERRIE," we classify the difficulty of "ERRIE" as Medium Difficult. Our model has a moderate rand index, which falls from 0.65 to 0.8, which can predict the difficulty well. However, our model does not have a high Silhouette Coefficient which means that the instances in the cluster are compact, and part of the instances overlap to some extent.

8.1.4 Task 4

- Several words have an abnormally high one-try successful rate, which our model cannot fit well.

- There are words with high failure rates. For instance, in position M108, 48% of players can only solve the puzzle in up to six tries.
- The percentage of Hard Mode players continues to rise and holds steady at 10% in around the last 100 days of the report. This trend shows an increase in the proportion of experienced players.

8.2 Overall Evaluation

8.2.1 Strengths

For Model I, we use the natural world approach to fit the dataset, which considers there is a "stable user group" to support the total user volume. Also, we apply a constantly decreased linear model to represent the user churn as the software develops. For Model II, we develop a logical model to simulate the Wordle process for a specific word, which allows us to give a reasonable distribution for each word within the word range. We make some adjustments to the data and algorithm in consideration of realistic situations. We also view the reported distribution as the manifestation of difficulty and use it to do K-means Clustering, giving the words a particular class. After predicting the DTNT of every word that has already appeared in Wordle, we use the predicting data as a quantitative variable and the class given by K-means as the nominal level variable to operate Random Forest Classification to verify the accuracy of our model's prediction. The combination of two machine learning algorithms enhances the model's robustness. For Model III, we optimize the number of clusters to 3 to ensure higher accuracy and convenience. Also, our clustering results generate relatively fitting curves w.r.t the reporting data.

8.2.2 Weaknesses

However, for Model I, since we are using the mathematical method to estimate the results, we cannot fully cover the trends of the reporting results. For instance, the weekly fluctuation due to the Working Days and Rest Days, and the festivals effects are ignored in our model. For Model II, our simulation model's bias mainly comes from our assumptions 1 and 2. These two assumptions has eliminated a lot of uncertain variables for the model. Those behavioral variables are too complex to be quantified. For Model III, our clustering curves follow the same trend, which decreases the classification accuracy and inevitably influences our model. Meanwhile, after predicting the results, the clustering model cannot distinguish efficiently, increasing our classification's difficulty.

8.2.3 Future Work

- Due to the time limit, many potential attributes will influence the distribution of the dataset. Also, the limited sample size restricts our models' performance—more data is needed in the future.
- Since the gaming results depend primarily on the first several guesses, we can study deeper on this determined aspect to better predict the result distribution.
- Despite only analyzing the influence of word frequency, we can take information entropy into account to simulate the thinking process of the human brain.

9 A Letter to the Puzzle Editor

Dear Editor,

Following your instructions, we analyzed the data in the file you provided and the dictionary data from Kaggle. We constructed different mathematical models to answer the first three question sets. The first is a Mathematical Time Series Wordle Prediction Model, a combination of the Logarithmic Normal and Logistic Growth Model. The second one is a Monte Carlo Simulation Model for the Wordle process. The third one contains K-means Model and GMM Model. The detailed report is as follows.

For the first question set, as we drew the scatter graph of the number of reported players along the time series, we observed that the data's shape is close to the probability density function (PDF) of the logarithmic normal distribution. Then we write a function fitting code to establish a Logarithmic Normal Model to fit the initial data. By changing the parameter of the function curve, we received a result in Figure 10. This result is ideal in the first 200 days but obtained a significant decrease compared to the reported data. Thus, we introduced a new concept, the "stable users". We assume that the group of stable users follows a logistic growth and add a curve generated by the Logistic Growth Model. After fitting the new model, we received a more precise visual and numerical prediction result. Details are shown in Figure 11 below. The final result of our prediction is that we will receive data from 14377 players.

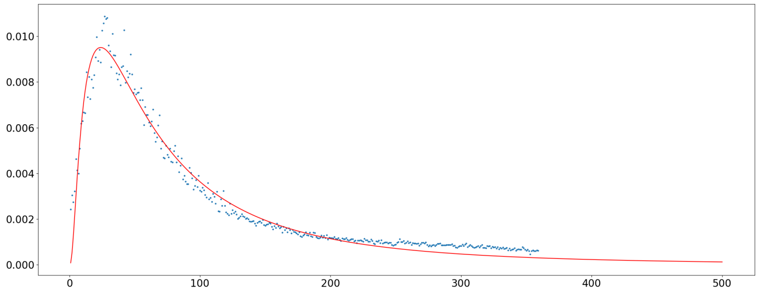


Figure 10: Fitting Result with Our Model

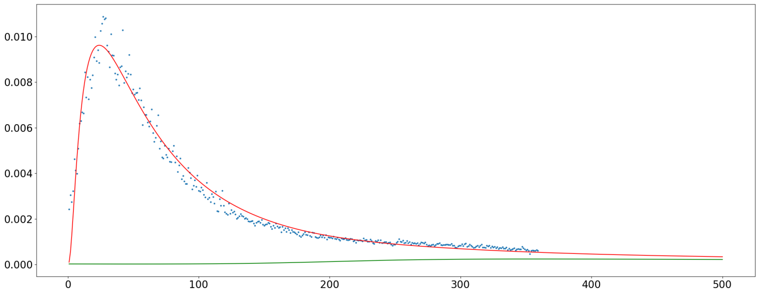


Figure 11: Fitting Result with Our Model

Several attributes of the word affect the percentage of scores reported that were played in Hard Mode. For instance, with the decrease in word frequency, the percentage of finishing the game with fewer tries will decrease. Also, for a specific prefix and suffix like "ab-" or "-st", it will reduce the difficulty of the game, which leads to an increase in the percentage of finishing the game with fewer tries. Moreover, high-frequency letters like "a", "e" in words also influence the distribution of report results. However, the pattern of those letters has different effects on the results. For instance, "mummy" is a problematic word in the game, although it has a relatively high word frequency.

For the second question set, we needed help building a model that only analyzes the data numerically and statistically.Then we wrote down the detailed logic of the Wordle process and established a Monte Carlo Simulation Model (MCSM) to simulate the game process precisely. Once we input a particular word and cycle index, it returns a percentage distribution of the number of trials. For the word EERIE that you required, we used the predicted result from the first model that there will be 14377 players' data on March 1st, 2023 as the cycle index. By running the MCSM, we received a result as Table 9.

	1 try	**2 tries**	**3 tries**	**4 tries**	**5 tries**	**6 tries**	**7 or more tries**
Result	0	1.9%	17.1%	40.5%	31.5%	7.5%	1.2%
STD	0.7	0.51	1.25	2.75	2.94	1.25	0.62

Table 9: Prediction of EERIE Report Distribution

To test the preciseness of our model, we use K-means Clustering to classify the reported data into three groups. Then we run a Random Forest Classification Model based on the prediction data, which contains all the words appearing in the given reported data. The results of the two classification models show significant homogeneity. Thus, we are confident in our prediction results. The uncertainties of our model may be the bias caused by our assumptions 1 and 2, which may not be accurate as players are not so efficient in playing the game as our MCSM do and might need to be more honest when they report their grades.

For the third question set, we found the optimal number of clusters equaled three, where both overfitting and underfitting were well avoided. Moreover, when clustering the distribution of the number of trials, using the K-Means model rather than Gaussian Mixture Model is better. Since there were overlaps in the number of trial distributions, the Silhouette Coefficient wouldn't be large enough, while the Rand Index remained high.

For the fourth question set, we proposed three interesting data set features. For instance, several words have an abnormally high one-try successful rate, which our model cannot fit well. And there are words with high failure rates. For instance, in position M108, 48% of players can only solve the puzzle in up to six tries. Also, the percentage of Hard Mode players continues to rise and holds steady at 10% in around the last 100 days of the report. This trend shows an increase in the proportion of experienced players.

The above is the entire content of our models. Should you need more information, we'll gladly discuss our investment strategy models in detail at the next meeting.

Sincerely,
Team 2322161

References

[1] https://www.kaggle.com/datasets/rtatman/english-word-frequency?resource=download.

[2] https://www.spsspro.com.