Problem Chosen	2023	Team Control Number
C	MCM/ICM Summary Sheet	2319848

Cracking Wordle: Predicting Wordle Results

Wordle, a popular online word-guessing game, has attracted significant attention in the recent years. Players attempt to guess a five-letter word within six attempts and receive feedback after each guess. The game's popularity has generated vast amounts of data, providing opportunities for statistical analysis and modeling. In this paper, we will present a comprehensive study of the Wordle game, focusing on predicting the number of reported results on a future date, the associated percentage of scores, and categorizing the difficulty of a word.

To predict future results, we have established three models: **Prophet, Multilayer Perceptron, and *K*-Means Clustering.** These models use characteristic values extracted from the word attributes to train and test their performance, and achieved a promising prediction effect, especially when in predicting the difficulty of guessing words, the accuracy reached nearly 90%. To ensure accuracy, we also designed an innovative algorithms called **Normal Distribution Principle Components Analysis.**

Our models have demonstrated superior performance, strong robustness, and adaptability to different situations, as well as decent generalization ability. We have provided a detailed explanation of the results and accuracy of our models, which can help readers better understand the relationship between words and results.

Furthermore, we have written a memorandum to the Puzzle Editor of the New York Times to offer our assistance in improving the user experience. Our models have potential value in enhancing the game and benefiting its large player base.

In conclusion, this paper offers a mature and professional analysis of predicting Wordle results. We hope that our research and models can serve as a valuable resource for future investigations and practical applications.

Key Words: Prophet, Multilayer Perceptron, *K*-Means Clustering, Principle Components Analysis, Wordle, Prediction

Contents

1 Introduction

1.1 Problem Background

Wordle, an online daily puzzle game, went viral on social media platforms of all kinds. The goal of Wordle is to guess a secret five letter word in no more than six attempts. After each guess, the commonalities between the secret word and the guessed word are revealed. With the available information, the player narrows down the list of possible solutions. According to the Wordle instructions on the New York Times website, the color of the tiles will change when players submit their words. A yellow tile indicates that the letter in that tile is in the word, but in the incorrect place. A green tile indicates that the letter in that tile is present in the right spot inside the word. A grey tile indicates that the letter in that tile is completely not included in the word. Figure 1 is the directions of Wordle posted to the New York Times website.

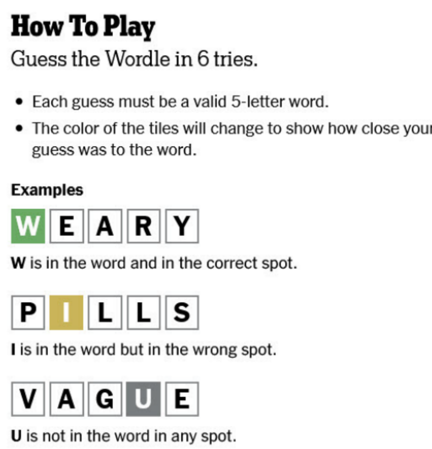

Figure 1: Directions of Wordle

Players can choose between normal mode and ”Hard Mode.” Wordle’s Hard Mode makes the game more challenging by demanding that, once a player has guessed a proper letter in a word (the tile becomes yellow or green), he or she must utilize that letter for further guesses. Figure 1 depicts a game played in Hard Mode. When playing in “Hard Mode”, players are required to use all the information available to them with each guess, while in easy mode, this constraint is not present. As a result, players can exploit this lack of restriction, especially when they have already solved four of the five letters in the secret word, leaving multiple possibilities for the final letter.

MCM has compiled a file of daily results for January 7, 2022, through December 31, 2022, which contains data such as the contest date, word of the day, number of participants reporting scores, and the percentage of hard mode players who guessed the word in one to six tries or couldn't solve it (indicated by X). Figure 2 showcases the results for the word TRITE on July 20, 2022, gathered by mining Twitter. While the percentages in Figure 2 sum up to 100%, this may not always be the case due to rounding.

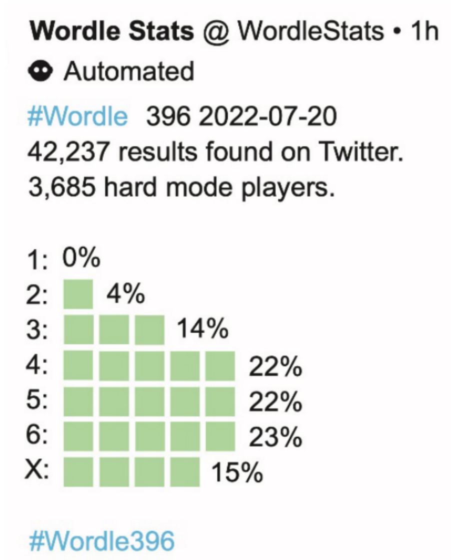

Figure 2: Results for the Word TRITE

1.2 Restatement of the Problem

Our team has been asked by the New York Times to analyze the results in a provided file and answer the following questions. The project consists of four main components: develop a model to explain the daily variation in the number of reported results and use it to create a prediction interval for March 1, 2023; create a model to predict the distribution of reported results for a given solution word on a future date and discuss the uncertainties associated with the model and predictions; develop and summarize a model to classify solution words by difficulty and determine the difficulty of a specific word, EERIE, while discussing the accuracy of the model; and list and describe other interesting features of the data set. Once completed, we will write a letter to the Puzzle Editor of the New York Times summarizing our analysis results.

2 Data Pre-processing

Firstly, we noticed that there were some obvious issues with the data. We cleaned the data by removing outliers, including:

(1) After visualizing the data, we noticed a distinct issue with the Number of reported results on November 30^{th}, 2022. Since we considered that Numbers in hard mode are basically equal to the Number of reported results, we took the weighted average of the previous and next day's values: Let this day's reported results = x, therefore, $\frac{x-23739}{2405-2316} = \frac{22628-x}{2200-2405}$, getting the result $x = 24591$, which can be used to replace 2569, the anomalous data.

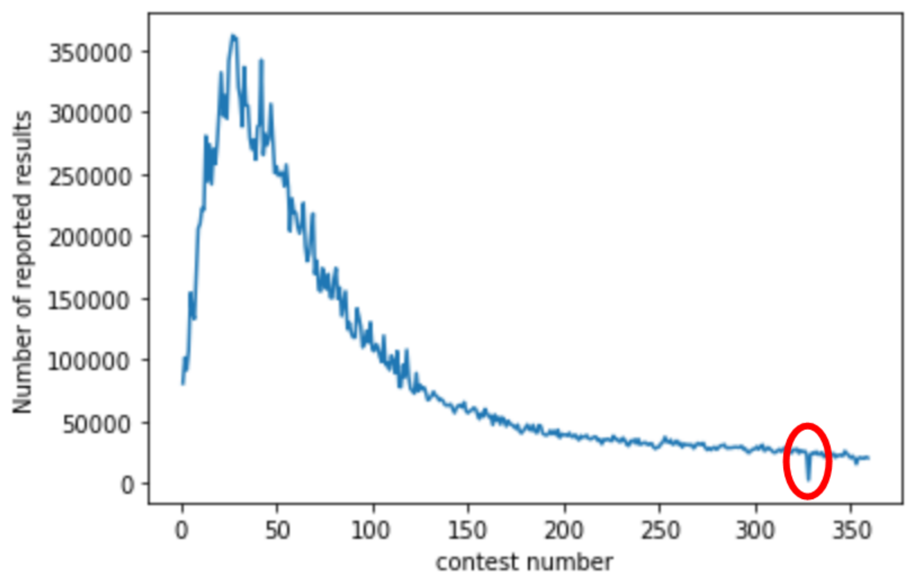

Figure 3: Abnormal Data

(2) We also noticed that the distribution vector $[1, 2, 18, 44, 26, 26, 9]$ for the word "nymph" on March 27^{th}, 2022 added up to 126%, which was deviated way too much from the theoretical value of 100%. Due to the fact that we find it hard to speculate the source of this issue, we used normalization $[x_1, x_2, \cdots, x_6] \rightarrow [\frac{x_1}{X}, \frac{x_2}{X}, \cdots, \frac{x_6}{X}]$ (here $X = x_1 + x_2 + ... + x_6$) to avoid this error, ensuring that the 1-norm of each word's distribution vector was equal to 1.

(3) We found some non-standard word lengths in the data, such as 314^{th}, 525^{th}, 545^{th} words(tash, clen and rprobe), and some non-existent words, such as 473^{th} word (marxh). We removed these anomalies and analyzed the remaining data.

3 Quantity Prediction Model and Feature Extraction

3.1 Model Selection

We calculated the average number of results in different days of week, the influence of different *days of the week* on the *Number of Reported Results* variable is evidently noteworthy.

Table 1: Average number of results in different days of week

Days of week	Average number of people reporting scores
Sunday	88308
Monday	90320
Tuesday	92754
Wednesday	94276
Thursday	91749
Friday	91341
Saturday	88160

The average number of submitted results has been calculated for both holiday and non-holiday periods, with the resulting data displayed in the figure.

Table 2: Average number of results in holiday and non-holiday

Days of week	Average number of people reporting scores
holiday	94335
non-holiday	89430

It is undeniable that the observed impact of both days of the week[1] and holiday status on the number of reported results is significant. Accordingly, the Prophet algorithm has been selected for forecasting, as it takes these variables into consideration.

3.2 Model Introduction

Prophet[2] is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. Its performance is optimized when used with time series data characterized by strong seasonal effects and multiple seasons of historical data. Additionally, prophet is known for its robustness in handling outliers, shifts in trends, and missing data.

The primary strengths of the Prophet algorithm lie in its ease of use and accuracy. It boasts several built-in features, such as automated detection and adjustment of seasonality, as well as handling of missing data. The Prophet model has demonstrated high accuracy in practical applications, forecasting complex time series patterns including holiday effects and gradually changing trends.

Another key advantage of Prophet is its open-source nature and ease of use. Developed in Python, it can be easily integrated with other data analysis tools and libraries, such as Pandas and Scikit-learn. Furthermore, the model offers a wide range of customizable parameters, which can be adjusted to improve prediction accuracy for different data sets. As such, the Prophet model is particularly well-suited for time series-based forecasting of reported results.

3.3 Model Application

Forecasting the reported results for a specific date is a typical time series prediction problem. To this end, we first created and fitted a Prophet model to the data after it was loaded.

The resulting forecast time series is then obtained using the $predict()$ method, and saved in a forecast data frame. Using Pandas' $loc[]$ method, we can filter for a specific date, and output the prediction value and confidence interval. Here, the output prediction value is given by $prediction['yhat']$, and the upper and lower bounds of the confidence interval are given by $prediction['yhat\_lower', 'yhat\_upper']$.

This process generates a visual representation of the trends, seasonality, and holiday effects in the predicted results, allowing us to make predictions for a specific day's reported results and evaluate the accuracy of the prediction using the confidence interval and visualization.

3.4 Model Evaluation

In the early stage of the popularity of the Wordle game, the number of reported results grew rapidly in the short term, reached a peak, and then declined at a relatively fast rate before gradually stabilizing. However, the rapid growth in the initial period does not conform to the overall trend of the data, and it may introduce a significant amount of error. To mitigate this potential issue, we selected only 300 data points from March 4^{th}, 2022 to December 31^{st}, 2022 for model training and testing[3].

We divided the 300 data points into a training set, consisting of the first 80% of the data, and a testing set, which consisted of the remaining 20%, to evaluate the performance of our model. We then visualized the predicted values and the true values of the testing set, as shown in the Figure 4.

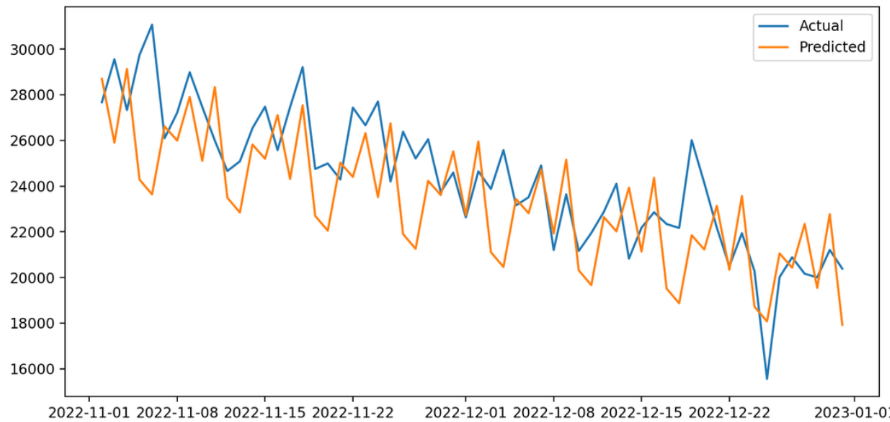

Figure 4: Predicted Value and True Value of Test Set

Based on the Prophet model we have created and trained, the predicted number of reported results on March 1, 2023 is estimated to be 15,090, with a confidence interval of $yhat\_lower$: 1,964 and $yhat\_upper$: 29,371, as shown in the Figure 5.

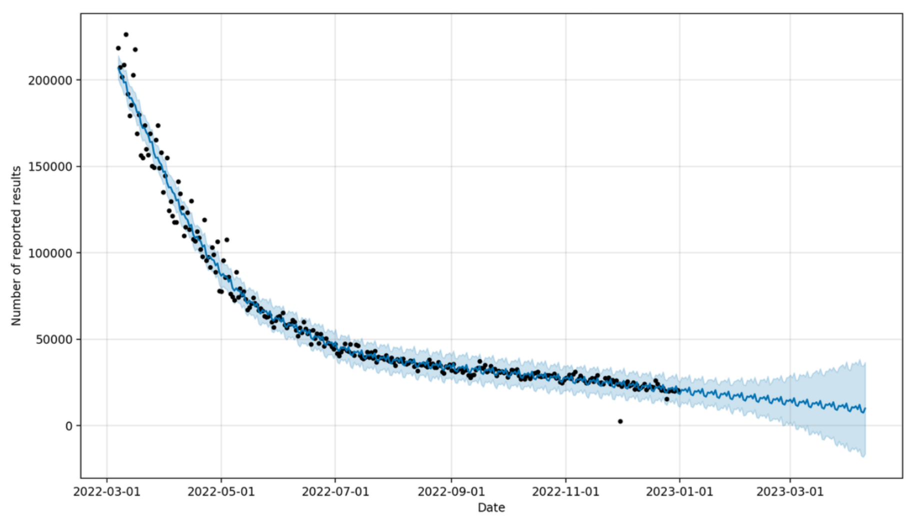

Figure 5: Prophet Model Prediction Chart

3.5 Impact Factors

The following discussion will pertain to the attributes of words that may influence the frequency of guessing. To restate the problem, given the distribution $vector = (x_1, x_2, \cdots, x_7)$, $x_1 + x_2 + \cdots + x_7 = 1$, we aim to investigate whether its label attribute can explain the vector. Since the label attribute is 1-dimensional while the vector is 7-dimensional with six degrees of freedom, we seek to reduce the dimensionality of the 7-dimensional vector. Upon inspection, we observe that each vector is approximately normally distributed,as shown in Figure 6, and hence, we consider reducing the 7-dimensional data to the 2-dimensional one (μ, σ).

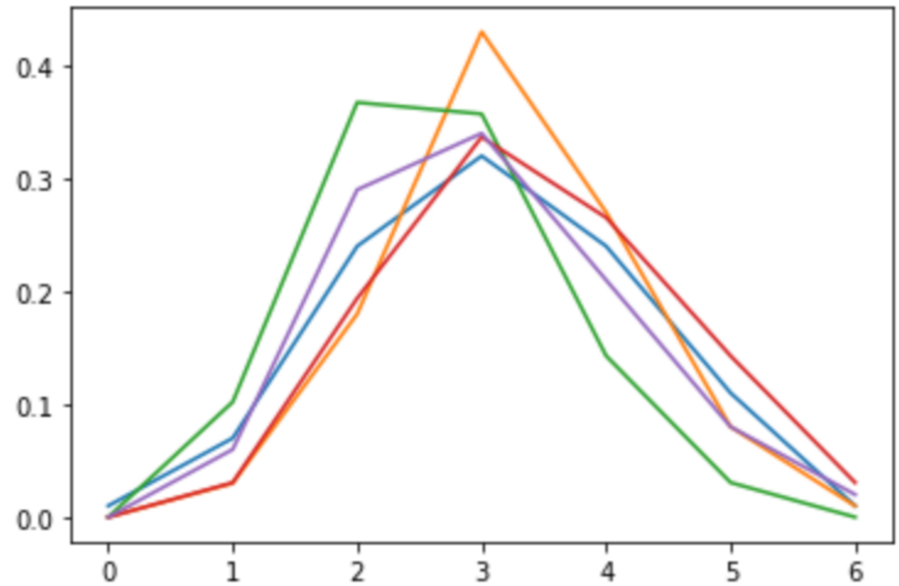

Figure 6: Percentage Distribution of Guess Times

Based on the 2-dimensional data, we can know that $\mu = E(vector) = \sum_{i=1}^{5} ix_i$ represents the average number of guesses made by players. It is evident that a larger value of μ implies a harder word to be guessed, while $\sigma = \sqrt{D(vector)} = \sqrt{\sum_{i=1}^{5} i(x_i - \mu)^2}$ indicates the degree of variability in the number of guesses. In order to find the correlation between $(\mu, label)$ and $(\sigma, label)$ Using the characteristic features of the distribution vector represented by the expected value μ and standard deviation σ we employed a dimensionality reduction technique called "ND-PCA" (Normal Distribution Principle Components Analysis). We used the Pearson correlation coefficient to determine whether two variables are correlated.

Correspondingly, We designed the labels of the following word attributes to examine whether they could explain the vector.

Let these five-letter word as $word = \overline{l_1 l_2 l_3 l_4 l_5}$.

(1) The Frequency of Word Usage

$$label = -\log f(word)$$

After querying the frequency of use for all five-letter words and recording the frequency of each word as $f(word)$, with $f(word)$ its range $[3\times10^{-8}, 2\times10^{-3}]$, we noticed that the differences between the frequencies are too large. Therefore, we considered using the logarithm of $f(word)$ as the label instead of using $f(word)$ directly. The Pearson correlation coefficients under this transformation are shown below.

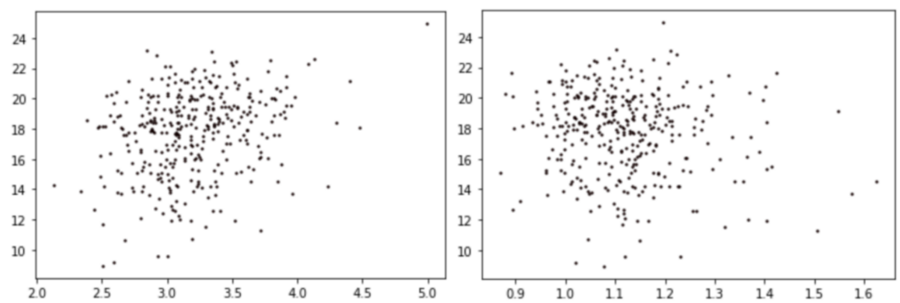

Figure 7: Label: Frequency of Word Usage

Table 3: Pearson coefficient and P-value

Pearson coefficient	P-value	Pearson coefficient	P-value
0.2708	2.107×10^{-7}	-0.1085	4.069×10^{-2}

As shown in Figure 7 and Table 3, the left half side displays the scatter diagram, Pearson coefficient and P-value of $(\mu, label)$, and the right half side displays the scatter diagram, Pearson coefficient and P-value of $(\sigma, label)$.

The figure shows the Pearson correlation coefficient between two sets of data, and the P-value is defined as follows. We know that when discussing whether two variables are related, we must first discuss the significance level of their linear correlation. Ignoring the P-value and only discussing the magnitude of the correlation coefficient is meaningless, as the correlation between the two may be caused by chance. Therefore, we need to determine the significance level of the correlation between the two variables.

We use the hypothesis testing method (R is the Pearson coefficient):

Null hypothesis $H_0 : R = 0$, there is no linear relationship between the two variables.

Alternative hypothesis $H_1 : R \neq 0$, there is a linear relationship between the two variables.

According to the hypothesis testing method, under the condition that H_0 is established, calculate the probability value (P-value) that the two variables are not related. If this P-value is small, it means that the probability that the two variables are not related is small, and we can reject the null hypothesis and accept the alternative hypothesis. We need a threshold here. Usually, the threshold is set at 5% (this threshold is also called the significance level). If $p < 0.05$, it means that we can reject the null hypothesis and accept the alternative hypothesis, that is, there is a significant linear relationship between the two variables.

Therefore, even if the correlation coefficient is large when the P-value is far greater than 0.05, we cannot say that there is a clear correlation between the two variables; and generally, we need to discuss the size of the correlation coefficient after the P-value meets the requirements. At this time, the P-value is very small, indicating that there is a linear relationship between the two variables.

For this data set, the P-value is very small (much less than 0.05), and the Pearson correlation coefficient falls within $(0.1, 0.3)$, indicating that there is a weak linear correlation both within $(\mu, label)$ and $(\sigma, label)$, and this attribute has a certain explanatory power.

(2) The Word ASCII Code Average

$$label = \frac{\sum_{i=1}^{5} ASCII(l_i)}{5}$$

Let the ASCII value of *letter* denoted by $ASCII(letter)$.

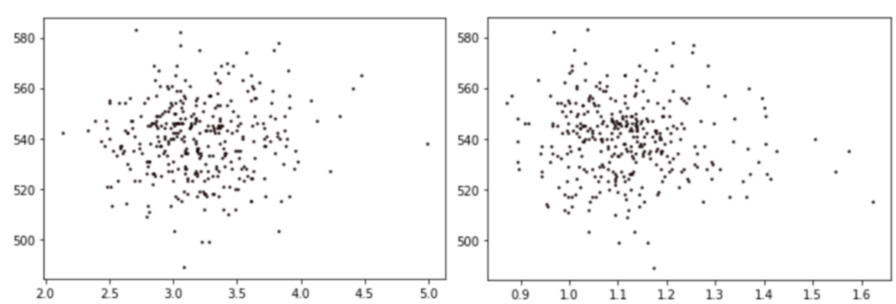

Figure 8: Label: The Word ASCII Code Average

As shown in Table 4, both the P-value is fairly high, and the Pearson correlation coefficient is low, indicating that the attribute does not significantly affect the percentage of people with different scores in hard mode.

Table 4: Pearson coefficient and P-value

Pearson coefficient	P-value	Pearson coefficient	P-value
0.0507	0.3398	-0.0744	0.1610

(3) Whether the Word Has Repeated Letters

$$label \in \{0, 1\}$$

if *word* has repeated letters (e.g. hello has repeated letter 'l'), then $label = 1$, else $label = 0$.

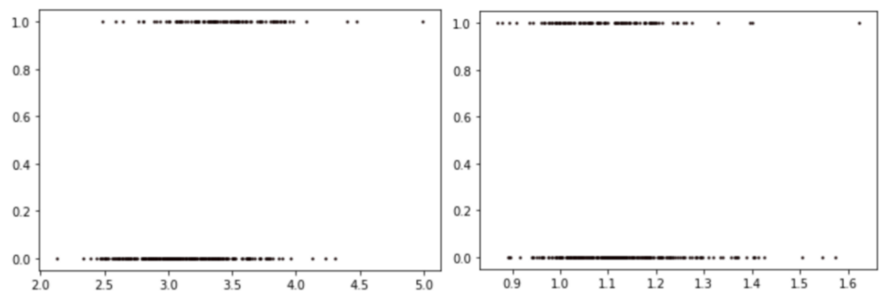

Figure 9: Label: Whether the Word Has Repeated Letters

Table 5: Pearson coefficient and P-value

Pearson coefficient	P-value	Pearson coefficient	P-value
0.3812	9.267×10^{-14}	-0.1096	0.0388

As shown in Table 5, the P-value is very small (much less than 0.05), and the Pearson correlation coefficient of $(\mu, label)$ is 0.3812, indicating that there is a weak linear correlation within $(\mu, label)$, the Pearson correlation coefficient of $(\sigma, label)$ is -0.1096, indicating that there is a weak linear correlation within $(\sigma, label)$.

(4) Sum of the Frequency of Using Each Letter in the Word

$$label = \sum_{i=1}^{5} Letter\_frequency(l_i)$$

Let the Letter frequency of *letter* denoted by $Letter\_frequency(letter)$.

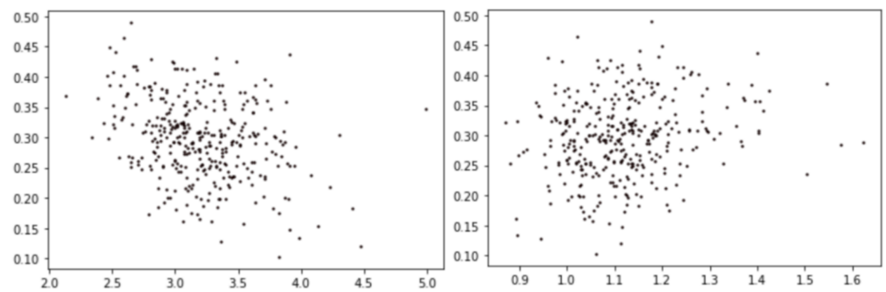

Figure 10: Label: Sum of the Frequency of Using Each Letter in the Word

Table 6: Pearson coefficient and P-value

Pearson coefficient	P-value	Pearson coefficient	P-value
-0.3136	1.447×10^{-10}	0.1997	1.485×10^{-4}

As shown in Table 6, the P-value is very small (much less than 0.05), and the Pearson correlation coefficient of $(\mu, label)$ is -0.3136, indicating that there is a weak linear correlation within $(\mu, label)$, the Pearson correlation coefficient of $(\sigma, label)$ is 0.1997, indicating that there is a weak linear correlation within $(\sigma, label)$.

(5) Product of Frequency of Using Each Word

$$label = \prod_{i=1}^{5} Letter\_frequency(l_i)$$

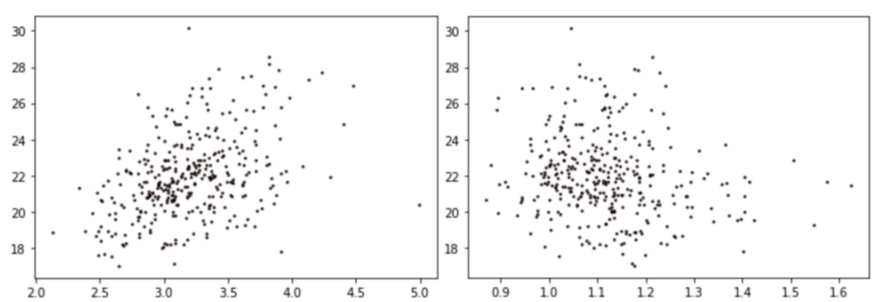

Figure 11: Label: Product of Frequency of Using Each Word

As shown in Table 7, the P-value is very small (much less than 0.05), and the Pearson correlation coefficient of $(\mu, label)$ is 0.4095, indicating that there is a weak linear correlation within $(\mu, label)$, the Pearson correlation coefficient

Table 7: Pearson coefficient and P-value

Pearson coefficient	P-value	Pearson coefficient	P-value
0.4095	7.898×10^{-16}	-0.1655	1.727×10^{-3}

of $(\sigma, label)$ is -0.1655, indicating that there is a weak linear correlation within $(\sigma, label)$.

4 Distribution Prediction Model

4.1 Model Selection

In order to predict the distribution of the results in the report, we need to extract some features of the words and analyze them.

We attempt to extract the attributes of the words and explore their relationship with the distribution of the results. One idea is to use each attribute as an input in a neural network and obtain seven outputs representing how many guesses it takes to successfully guess a particular word, and then train the model on the dataset to obtain a model that can ideally predict the distribution of the results.

Considering the multi-input and multi-output nature of the problem, we use a multilayer perceptron artificial neural network regression algorithm to predict the distribution of the results.

Furthermore, in the initial consideration of the distribution of the results, it is not difficult to find that it is a set of nonlinear relationships, and a multilayer perceptron can learn nonlinear models, which can better fit nonlinear datasets.

4.2 Model Introduction

The multilayer perceptron (MLP)[4] is a type of feedforward artificial neural network model that maps multiple input datasets to a single output dataset. The MLP introduces one or more hidden layers between the input and output layers on top of the single-layer neural network.

In regression problems, the MLP generates predictions by approximating the mapping relationship between input data and target variables through non-linear transformations across multiple layers and gradually decreasing errors. Compared to simple linear models, MLP has higher non-linear modeling capabilities and can handle more complex datasets, making it a very suitable neural network model for predicting multidimensional result distributions in this problem.

When designing an MLP model, it is necessary to choose suitable network structures, activation functions, loss functions, and optimizers to maximize the performance of the model. Commonly used activation functions include Sigmoid, ReLU, and tanh, while common loss functions include mean squared error (MSE), mean absolute error (MAE)[5], and commonly used optimizers include stochastic gradient descent (SGD), Adam, and RMSProp.

4.3 Model Application

In order to predict the proportion of guessing times for each word in the report's result distribution, we cannot limit ourselves to time series prediction models. Instead, we need to analyze and extract more characteristics of the words, and the model's input should contain more content. Therefore, we used the multilayer perceptron (MLP) artificial neural network regression algorithm to solve this prediction problem.

The MLP Regressor class in the Python Scikit-learn library can be used to build and train the MLP model. We extracted the following features to use as input variables for the MLP artificial neural network regression algorithm: the sum of the class information entropy of the frequency of each letter in the word, the sum of the frequency of each letter in the word, the number of vowel letters in the word, the frequency of occurrence of the word, and whether the word contains duplicate letters.

After training the model, we evaluated it using data from the test set, normalized the data, and calculated the mean squared error (MSE) to measure the predictive performance of the model.

4.4 Model Evaluation

In the model, we used ReLU[6] activation function and Adam optimization algorithm, with 100, 50, and 25 neurons in the hidden layers and 10,000 iterations. 80% of the data was used for training and 20% for testing.

The uncertain factors considered by our model include:

- The sum of the class information entropy of the frequency of each letter in the word.
- The sum of the frequency of each letter in the word.

- The number of vowels in the word.
- Frequency of the word.
- The word contains duplicate letters.

Using the MLP model we created and trained, we predicted the result distribution of EERIE for the given problem to be $[1, 8, 28, 33, 20, 9, 3]$. The mean squared error (MSE) between the test set's true data and the predicted data is approximately 0.002.

5 Classification Model

5.1 Model Selection

We aim to develop a model M that assigns difficulty levels to each word. We divide the difficulty of each word into three levels $\{0, 1, 2\}$, where 0 represents words of moderate difficulty, 1 represents words of high difficulty, and 2 represents words of low difficulty. Given a dataset of 356 7-dimensional vectors, our goal is to cluster them into three groups. To achieve this, we consider using the K-Means algorithm[7]. However, since it is difficult to work with 7-dimensional vectors and visualize their results, we will use PCA to reduce the dimensionality of the vectors and map them into a 3-dimensional space.

5.2 Model Introduction

5.2.1 K-Means Clustering

K-Means Clustering is a clustering algorithm that can partition a large amount of data into a small number of clusters based on the similarity of their data features. The objective of the K-Means algorithm is to divide n d-dimensional data points into K clusters, such that the within-cluster sum of squares is minimized. Since there are numerous possible clustering outcomes, it is not practical to expect any specific clustering algorithm to always achieve the best solution. Therefore, the K-Means algorithm seeks a "local optimal solution," rather than a global optimal solution.

5.2.2 Principle Components Analysis

PCA[8], which stands for Principal Component Analysis, is a method of feature dimensionality reduction. It has a wide range of applications in eliminating data noise, redundancy, and other aspects. PCA analyzes and finds the primary components of data features and uses these primary components to represent the original data. On the one hand, this can deepen the understanding of the data itself, and on the other hand, the simplified data has the characteristics of less noise and ease of processing and computation when used for downstream tasks.

PCA requires the preservation of the overall variance structure of high-dimensional data to the maximum extent possible.

5.3 Model Application

First, we use PCA to reduce 356 7-dimensional data to 3-dimensional data, and then perform 3-clustering[9] on these 356 3-dimensional data to obtain 3 clusters.

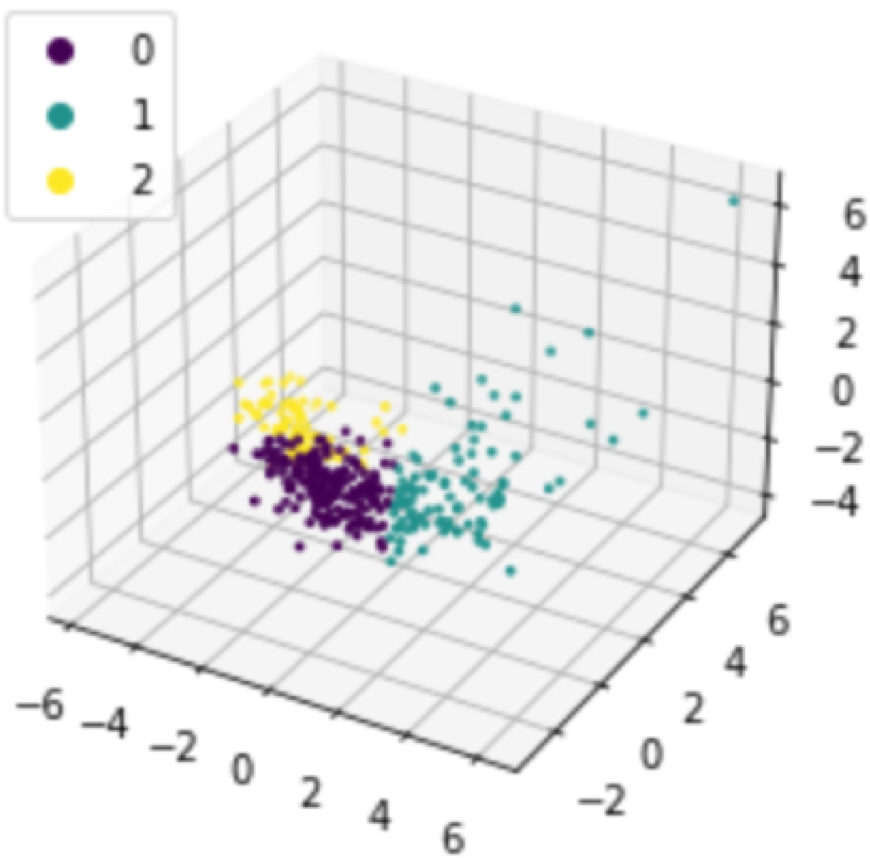

Figure 12: Label: 3-Means Clustering after PCA

Next, we calculate the centroid coordinates and the average number of guesses for each cluster[10], which is μ (defined in Section 3.5), and based on the significant differences in the average number of guesses, we label the three clusters as *medium*, *hard*, and *easy* respectively, as shown in Table 8.

Table 8: Pearson coefficient and P-value

Cluster	Centroid Coordinates	Average Number of Guesses	Difficulty
Cluster 0	$[-0.19472049, -0.86728776, 0.10475941]$	4.146	Medium
Cluster 1	$[2.44173911, 0.79714845, -0.12933217]$	4.692	Difficult
Cluster 2	$[-2.42092617, 1.08159674, -0.0917448]$	3.738	Easy

5.4 Model Evaluation

For each set of 7-dimensional data, we denote its mean as μ (defined in Section 3.5). We would mark the *data* as:

$$\begin{cases} 2(Easy) & \text{if } \mu \leq \frac{3.738+4.146}{2} \\ 0(Medium) & \text{if } \frac{3.738+4.146}{2} < \mu < \frac{4.146+4.692}{2} \\ 1(Hard) & \text{if } \mu \geq \frac{4.146+4.692}{2} \end{cases}$$

With a total of 356 sets of data, we found that 310 of them are correctly labeled after comparison. Therefore, the accuracy of our model is $\frac{310}{356} \approx 87.1\%$, which is fairly accurate.

We might as well reconsider the word EERIE. In the Wordle Puzzle game, the word EERIE could be considered a medium to difficult level word. It has a length of 5 letters, which reduces the number of attempts needed before guessing, and hence lowers its difficulty level. However, it contains two letter "E"s, which could make it more difficult to guess as the guesser needs to determine the position of the two "E"s. Additionally, EERIE is not one of the most common words, which could also increase its difficulty level. Thus, it could be considered a medium to difficult level word depending on the specific rules of the game and the guessing skill of the player. Besides, through manual labelling, the accuracy of our model is about 0.871.

6 Interesting Features of the Dataset

(1) Upon analyzing the number of daily submissions to Wordle from January 7^{th}, 2022 to December 31st, 2022, it is evident that the game's popularity experienced

a sharp increase beginning on January 7^{th}, 2022. The game's peak popularity occurred from February 2^{nd} to February 4^{th}, 2022, after which the game's popularity gradually diminished. This trend is reminiscent of many real-life phenomena, such as social issues that garner widespread attention upon initial exposure, with popularity rapidly increasing until it reaches a peak after a certain period of time, and subsequently declining.

(2) The number of individuals submitting results on Wordle is greatly influenced by two factors, namely, the particular day of the week and whether or not it is a holiday. Based on our statistical analysis, higher numbers of users are reported on holidays and on weekdays from Monday to Friday. From a macro perspective, there is a substantial increase in reported results due to people's inclination toward novelty and following trends. The number of reported results during holidays is also higher as individuals have more leisure time to engage with the Wordle Puzzle game. Interestingly, the number of reported results during workdays surpasses those on weekends. This can be attributed to the abundance of fragmented time available during workdays, which is ideal for playing the Wordle Puzzle game.

(3) An investigation was conducted on the daily submissions in the hard mode of Wordle, from January 7^{th}, 2022 to December 31^{st}, 2022. Overall, the trend was similar to that of all modes. However, one doubtful point was observed on February 13^{th}, 2022, which appeared to be an outlier. It is highly probable that this data point was due to input error. If the average of the data from the day before and the day after February 13^{th} were used to replace the data point, the trend would become much clearer.

(4) The proportion of individuals selecting the hard mode has been consistently increasing. Graphing the number of individuals who select the hard mode on the vertical axis, it is readily apparent that the proportion of individuals selecting the hard mode has been steadily increasing over time, gradually approaching a steady state. The trend is demonstrated in Figure 13. This phenomenon may be attributed to an increasing familiarity with the game, prompting more individuals to attempt higher difficulty modes.

(5) We have identified several extremely difficult-to-guess words with a high proportion of "7 or more tries" for successful completion. These words include "parer" with a 48% success rate, "foyer" (26%), "catch" (23%), and "watch" (20%) among others. Notably, "parer" stands out as significantly more difficult to guess than

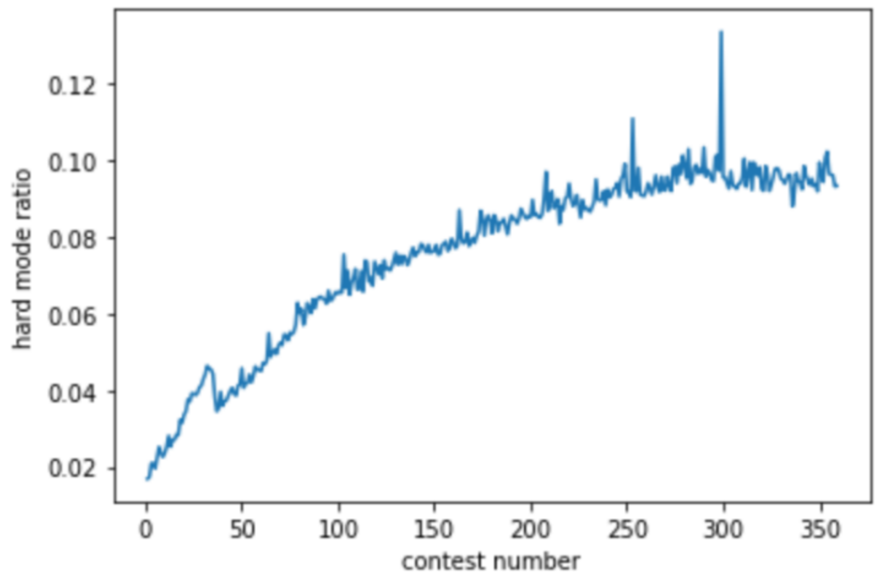

Figure 13: Hard Mode Ratio - Contest Number

the other 355 words in the dataset, possibly due to its extremely low usage frequency of only 3×10^{-8}, which is the lowest of all the words. In the case of the other difficult-to-guess words, we found that they typically contain repeated letters, which may make them more challenging.

7 Merits and Demerits of Our Models

7.1 Prophet Model

Merits

(1) The introduction of seasonal and holiday factors. As compared with other predictive models, it could be better fit for the trend changes, seasonality, holidays, and unexpected events in the time series. These features are precisely what make Wordle game unique, making its predictions more accurate.

(2) Minimal dependence on data continuity. It can effectively handle missing values in the data, and this is particularly important in Wordle games with random word selection and potential missing data.

(3) Fast and efficient. The curve fitting process is significantly faster than traditional training models, allowing for quicker data iteration.

Demerits

(1) Heavily rely on historical data. It may limit the accuracy of the prediction results due to the limited scope of the given data set.

(2) Lack the ability to remove abnormal data. It may potentially ignore special events and abnormal situations in Wordle games. This can affect the accuracy of the Prophet model's predictions when certain anomalies occur in the data set.

(3) Relatively simple and may not have the expressive power to capture complex patterns. As a result, the model may under-fit and struggle to learn complex patterns, leading to a decrease in prediction accuracy.

7.2 MLP Model

Merits

(1) Can handle high-dimensional data. It can manage the relationships between high-dimensional data and non-linear relationships, making it well-suited for capturing the complex relationships between the results' percentages in Wordle game.

(2) High degree of flexibility. It allows for customized adjustments to the number of neurons in the hidden layer and activation algorithms. These adjustments can be made based on the specific characteristics of Wordle game data to improve the accuracy of the model's predictions.

Demerits

(1) High data requirements. The training process requires a large amount of data and computing resources, so a significant amount of data pre-processing and feature extraction are needed before training.

(2) Be prone to the problem of local optimal solution. The MLP model is prone to local optimal solutions and appropriate optimization algorithms are needed to avoid this situation.

7.3 K-Means Clustering Model

Merits

(1) Automatic analysis of data characteristics. The K-means clustering model can automatically discover patterns and rules in data based on its characteristics.

(2) High computational efficiency. It can handle large-scale data, which is important for random word selection and potential missing data in the Wordle game.

(3) Good interpretability of the model. It is convenient to analyze the difficulty level of words based on the clustering results.

Demerits

(1) High data requirements. The K-means clustering model is sensitive to noise and outliers in the data, so appropriate data pre-processing and outlier handling are required.

(2) Difficulty in determining the number of clusters k. The K-means clustering model requires manual setting of the number of clusters k, which requires prior knowledge of the characteristics of Wordle game data or trial and optimization.

(3) Strong sensitivity to initialization. The K-means clustering model is very sensitive to the initial point selection, and selecting different initial points may lead to different clustering results.

8 Brief Summary

This paper presents an analysis of data from the popular game Wordle, and the development of three models to address various questions related to the game. For the first question, we utilized the Prophet algorithm to model the variation in the number of reported results and identified several attributes that affect the percentage of scores reported in Hard Mode. These attributes include the frequency of word usage, whether the word has repeating letters, the sum of the frequencies of letters in the word, and the product of the frequencies of letters in the word.

For the second question, a MLP neural network regression model was employed to predict the distribution of reported results for a given solution word on a future date. A specific example was provided for the word EERIE on March 1^{st}, 2023, and uncertainties associated with the model and predictions were discussed.

For the third question, we adopted the K-means clustering algorithm to classify solution words by difficulty and identified the attributes associated with each classification. The difficulty of the word EERIE was classified as medium, and the accuracy of the model was evaluated to be approximately 87.1% through manual labelling.

Overall, these models provide insights into different aspects of the Wordle game and can be used to improve its management and operation.

9 Memorandum

To: The Puzzle Editor of the New York Times

From: Team #2319848

Subject: Cracking Wordle: Predicting Wordle Results

Date: February 20, 2023

Dear Editor:

Here is Team #2319848. We are writing to express our admiration for the fascinating Wordle Puzzle game and to offer our assistance in predicting the game's results. We have developed several models that can accurately predict the number of reported results on a given day and the distribution of results for any specific word. All you need to do is provide us with the date or word you wish to make a prediction for.

We have carefully tested our models using the provided data file, and are proud to report an accuracy rate of almost 90%. We believe that our predictions could be a valuable resource for you and help you improve user experience.

Now let me introduce the results.

Number of reported results prediction

Based on our Prophet model, the number of reported results on March 1, 2023 could be 15090. Overall, the number of reported results increased rapidly at the beginning, and after reaching the peak, it keeps decreasing until now. By analysing a part of the period, we discovered that it is related to the week and holidays as well, which means the number will be a little bit higher if it is a holiday or not a weekend. So, here is the explanation of the variation. From a macro perspective, the number of reported results will increase rapidly because of people's favor for new things and a kind of fashion trend. During holidays, the number will be higher because people have more spare time to play the Wordle Puzzle game. Interestingly, during work days the number will be higher than during weekends. That's because there is a lot of fragmentation time between work, which is perfect for playing Wordle Puzzle games.

We explored many possible attributes and found duplicate letters, frequency of each letter, and frequency of the word will affect the percentage of scores reported that were played in Hard Mode. For example, using duplicate letters in a word does indeed increase the difficulty level of guessing the word, as it can mislead the guesser into guessing letters that do not appear in the word. And ASCII will make no sense, because people will not guess the word by considering ASCII.

The prediction of associated percentages of (1, 2, 3, 4, 5, 6, X) for a future date

Duplicate letters, frequency of each letter, and frequency of the word are associated with our model and predictions. By our model, the result of the word EERIE on March 1^{st}, 2023 could be [0, 1, 27, 33, 21, 8, 3]. After careful inspection and analysis, the MSE is about 0.002, which means our model is highly accurate.

Classify solution words by difficulty

Our model enables you to input any word for prediction. The model will then provide you with the word's attributes and its corresponding difficulty level, which can be categorized as Easy, Medium, or Difficult. Using the model, the word EERIE is categorized as Medium level. According to the analysis in the paper, the category of EERIE is very reasonable. Besides, through manual labeling, the accuracy of our model is about 0.87. This means our model can categorize the difficulty level of words relatively well.

Interesting features

Once a social topic is exposed, its popularity will rise rapidly, reach its peak after a period of time, and then gradually decline. Meanwhile, the proportion of people choosing the Hard mode in the game is constantly increasing.

That's all the results we want to share with you. We are confident that our models have discovered many interesting features of this game accurately. If you have any questions about the results or our models, I would be interested in discussing them with you.

References

[1] Andreas Abels, Elias Pitschmann, and Daniel Schmand. Prophet-inequalities over time. *arXiv preprint arXiv:2211.10471*, 2022.

[2] Hugo Ticona-Salluca, Fred Torres-Cruz, and Ernesto Nayer Tumi-Figueroa. Machine learning applied to peruvian vegetables imports. *arXiv preprint arXiv:2301.03587*, 2023.

[3] Rameshwar Garg, Shriya Barpanda, et al. Machine learning algorithms for time series analysis and forecasting. *arXiv preprint arXiv:2211.14387*, 2022.

[4] Hassan Ramchoun, Youssef Ghanou, Mohamed Ettaouil, and Mohammed Amine Janati Idrissi. Multilayer perceptron: Architecture optimization and training. 2016.

[5] Y-S Park and S Lek. Artificial neural networks: multilayer perceptron for ecological modeling. In *Developments in environmental modelling*, volume 28, pages 123–140. Elsevier, 2016.

[6] Meha Desai and Manan Shah. An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (mlp) and convolutional neural network (cnn). *Clinical eHealth*, 4:1–11, 2021.

[7] Kristina P Sinaga and Miin-Shen Yang. Unsupervised k-means clustering algorithm. *IEEE access*, 8:80716–80727, 2020.

[8] Lindsay I Smith. A tutorial on principal components analysis. 2002.

[9] Ismail Bin Mohamad and Dauda Usman. Standardization and its effects on k-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, 6(17):3299–3303, 2013.

[10] Carly L Clayman, Satish M Srinivasan, and Raghvinder S Sangwan. K-means clustering and principal components analysis of microarray data of l1000 landmark genes. *Procedia Computer Science*, 168:97–104, 2020.