PUZZLE GAME: PREDICTION AND CLASSIFICATION OF WORDLE SOLUTION WORDS [*]

Haidong Xin[1], **Fang Wu**[1], **Zhitong Zhou**[1], **Shujuan Wang**[2*]
Harbin Engineering University
[1]College of Computer Science and Technology
[2]School of Mathematical Sciences
{xhd0728, wufangcs, zhouzhitong, wangshujuan}@hrbeu.edu.cn

ABSTRACT

In MCM/ICM 2023, we proposed a new result prediction model for the popular game Wordle launched by The New York Times. We first preprocessed the raw data and then established a prediction model based on ARIMA to predict the number of report results as of March 1, 2023. We selected word usage frequency, word information entropy, and the number of repeated letters contained in the word as the attributes of the word, and conducted a correlation analysis between these three attributes and the percentage of seven attempts. We also established a regression model based on the XGBoost algorithm, predicted the distribution of reported results, and predicted the correlation percentage of "EERIE". In addition, we also constructed a word classification model that classified words into "simple", "moderate", and "difficult", and explored the relationship between the three attributes and the classification results. Finally, we calculated the percentage of players in the dataset who needed 3 or more attempts for each word. The appendix provides relevant information and problems to be solved for the mathematical modeling competition.

***K*eywords** ARIMA · XGBoost · K-Means · Decision Trees

1 Introduction

1.1 Background

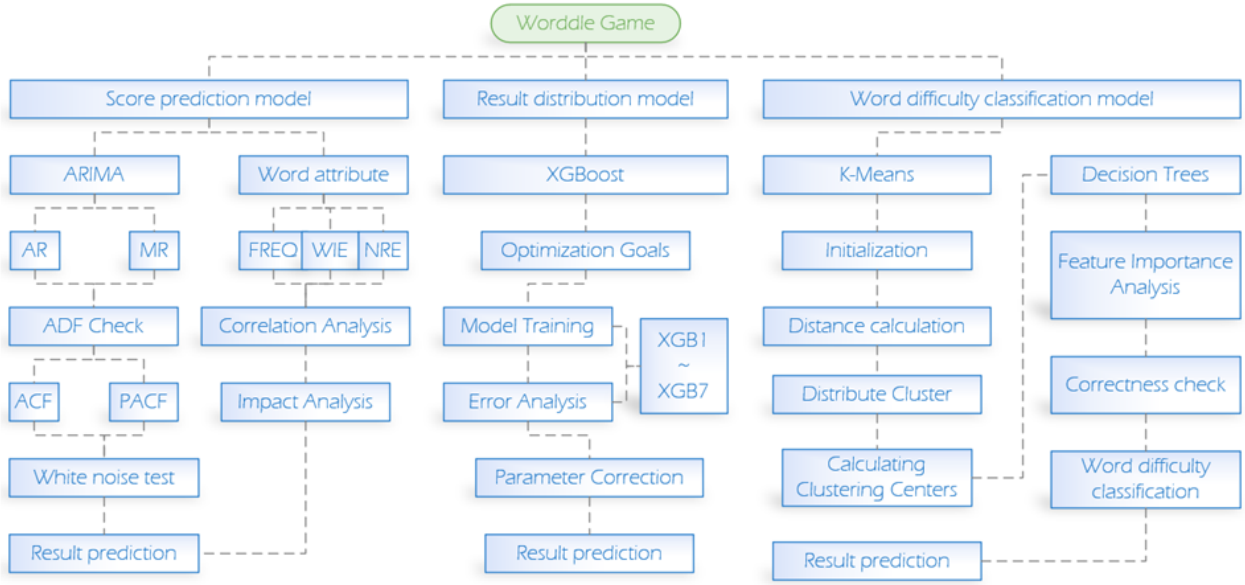
Wordle, a popular puzzle game[1], is now being offered as a daily challenge by The New York Times. Players need to guess a word consisting of five letters in six or fewer guesses, and receive feedback after each guess. In order to participate in the contest, each guess must be an English word.

When you submit a word, the color of the letter tiles will change. A yellow tile indicates that the letter appears in the word, but in the incorrect position. A green tile means that the letter appears in the word and is in the correct position. A gray tile means that the letter is not in the word. Players can play the game in either regular mode or "hard mode," which makes the game more difficult by requiring that once the player finds the correct letter in a word (yellow or green tiles), those letters must be used in subsequent guesses.

1.2 Our Works

We developed an ARIMA-based prediction model for the number of reported results, which is able to predict the number of reported results at a future date. The frequency of word usage (FREQ), the information entropy of a word (WIE), and the number of repeated letters contained in a word (NRE) were selected as attributes of the word, and a regression model based on XGBoost for predicting the distribution of results was developed, which was able to predict the percentage of tries corresponding to a word in a future period. Finally, a classification model based on k-means and

[*] indicates corresponding author.



Symbols	Definition
FREQ	Frequency of word occurrences
WIE	Word information entropy
NRE	Number of repeated letters in a word
p_i	Probability of occurrence of the i-th letter in the corpus
$L(t)$	XGBoost objective function
n	Sample size
I	Loss function
$f_t(x_i)$	Newly added functions for each iteration
γ	Decision Tree Complexity Parameters
λ	Parameters of leaf node weights in decision trees

Table 1: Symbol Description

decision trees for the difficulty of solving words was developed, which was able to classify a certain solution word as easy, medium or difficult according to its attributes, and the accuracy of the model classification was 77

2 Symbol Table and Assumptions

2.1 Symbol Table

We will explain some important symbols in Table1

2.2 Assumptions

2.2.1 The data is reliable enough

For the calculation of the information entropy and frequency of occurrence of words, we used data from Mathematica and Google books, and we selected all corpus documents from 2020-2022 as the data source, assuming that the data volume is large enough and reliable enough.

2.2.2 Each word in the answer lexicon has the same probability of being a solved word

When each word has the same probability of occurrence, it ensures that the data is universal and not influenced by human factors.

Date	Word	Number of reported results	Number in hard mode
2022-04-29	**tash**	106652	7001
2022-11-26	**clen**	26381	2424
2022-11-30	study	**2569**	2405
2022-12-16	**rprobe**	22853	2160

Table 2: Data preprocessing

2.2.3 The same word does not appear frequently in a short period of time

People are impressed by the solved words, and if a solved word appears frequently, it may lead to fewer guesses and reduce the authenticity of the data.

2.2.4 Each player is independent of each other and will not exchange information about solving words

Each player solves the words independently and no related information or answers are passed on. This ensures that the number of tries made by each player is the most realistic and objective.

3 Data Preprocessing

The check revealed that there was no missing data in the given data, so only outlier processing was required. In order to ensure the reasonableness and reliability of the data, outlier checking and outlier processing are performed on the given data.

In the three days of 04-29, 11-26 and 12-16, the words recorded are not words composed of five letters, so we choose to delete the data of these three days.

The number of people in difficult mode in the record of 11-30 is 2569, which is very different from the corresponding data of its nearby records, so we choose the average value of the corresponding number of people in the left and right 3 days as the number of people in difficult mode on this day.

After summing up the percentages of 1 tries to 7 tries more for each day, we found that the sum of many data is not 100%, even the sum of 3-27 is 126%. In this paper, we remove this data from 3-27 and normalize the associated percentages of (1, 2, 3, 4, 5, 6, X) in the rest of the records, and the table2 shows some of the processed data.

4 Number of Report Results Prediction Model

4.1 Introduction of ARIMA Model

Autoregressive Integer Moving Average, or ARIMA for short, is a statistical analysis model that uses time series data to predict future trends. The basic idea of ARIMA[2] is to treat the data series formed by forecasting over a period of time as a random series that can be approximated by a model that describes this series. Once this sequence is determined, the model can predict future values based on the past and present values of the time series. In this model, we try to predict the number of players reported on March 1, 2023 based on the number of participants in the game officially reported for each day from January 7, 2022 to December 31, 2022.

The ARIMA model consists of an autoregressive (AR) model and a moving average (MA) model. the AR model describes the relationship between current and lagged values and uses historical data to predict future values. the MA model uses a linear combination of past residual terms to look at future residuals. the ARIMA forecasting model can be written as follows:

$$\hat{p}^t = p_0 + \sum_{j=1}^{p} \gamma_j p^{t-j} + \sum_{j=1}^{q} \theta_j \varepsilon^{t-j} \tag{1}$$

where p is the order of Auto Regressive Model (AR), q is the order of Moving Average Model (AM), ε^t is the Error term between time t and $t-1$, γ_j and θ_j are the fitting coefficients, p_0 is constant term.

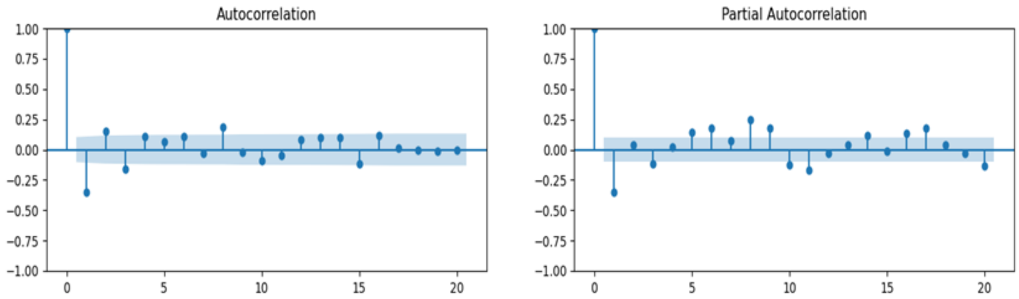


Figure 1: ACF and PACF tests

4.1.1 Model Building

We use Augmented Dickey-fuller unit root test to test for smoothness. If the p-value obtained from the ADF test is less than 0.05, the time series is stable. If it is unstable, the unstable series is transformed into a stable series using the difference method. The p-value after ADF test for the original data was calculated to be 0.25, and the p-value after first-order differencing was 0.02, and the time series tended to be stable after first-order differencing.

4.1.2 ACF and PACF Function – p, q Selection

Respectively, ACF (autocorrelation function) and PACF (partial autocorrelation function) are both functions to evaluate the linear relationship between historical data and the current value. The formula of ACF is

$$ACF(q) = \frac{Cov(X_j, X_{j-q}}{Var(X_0)} = \frac{\frac{1}{n-q}\sum\limits_{j=q+1}^{n}(x_j - \overline{x})(x_{j-q} - \overline{x})}{\frac{1}{n}\sum\limits_{j=1}^{n}(x_j - \overline{x})^2} \tag{2}$$

for the order q and the time series $\{x_1, x_2, \cdots, x_n\}$, The formula of PACF is so sophisticated that we will not list it in this article. Further information about PACF can be referred to [3]. The results of ACF and PACF are as fig 1, we choose $p = 0$, $q = 1$.

4.1.3 White Noise Test

From the computational analysis, we derived the model for reporting the scores as ARIMA (0,1,1). The p-value obtained from the white noise test of this model was close to 0 and passed the test.

4.1.4 Predicted Result Interval

We predict that the number of reported results on March 1, 2022 is 21005. Observing the fig 2, we find that after 2022-07-07, the predicted value curve is more consistent with the true value curve and both are smoother, so we calculate the prediction interval for March 1 based on the difference between the predicted and true values for the two months of November and December 2022. The forecast interval for the results reported on March 1 is given by the following formula, where Len denotes the fluctuation value of the predicted value, n denotes the total number of days in November and December after data pre-processing, r_i denotes the true value of the reported number, and p_i is the predicted value.

After calculation, $Len = 3.18\%$, so the interval where the result is located is $21005\pm$, i.e. $[20337, 21673]$.

In the graph above, the number of reported results changes over time, first increasing in the short term, then decreasing immediately, and finally stabilizing. We believe that Wordle, as a fun and challenging game, was widely sought after at the beginning of the period, so the number of participants increased rapidly and people were happy to share their results online. Later on, probably due to the excellent difficulty increasing or the loss of players' curiosity, the number of participants sharing gradually decreased, eventually leaving what people call the ashes of the game.

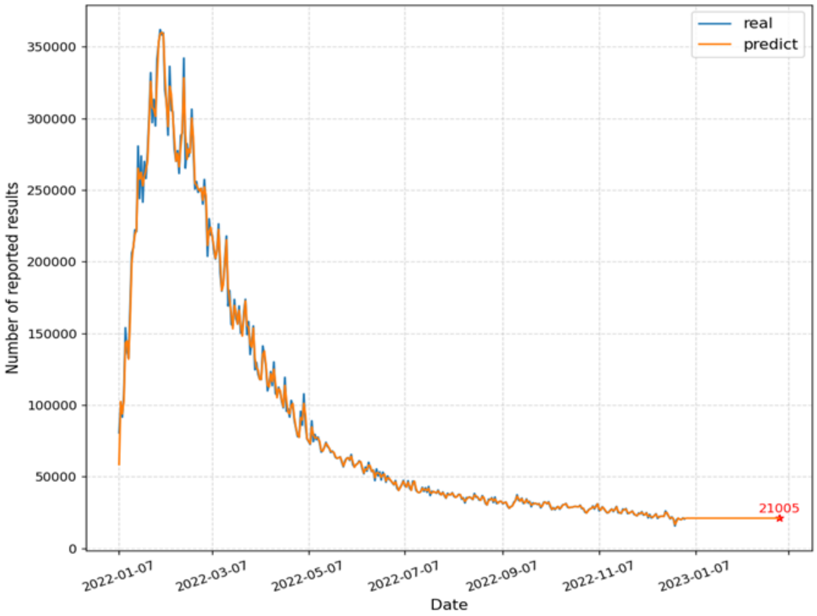


Figure 2: ARIMA predicted target values

FREQ	Frequency of word occurrences
WIE	Word information entropy
NRE	Number of repeated letters in a word

Table 3: Word attribute selection

$$Len = \frac{\sum_{i=1}^{n} |r_i - p_i|}{n} \tag{3}$$

4.2 Analysis of Word Attributes

4.2.1 Attribute Selection

English words have several attributes, such as word commonness, word composition, lexical properties, meaning, word form, word collocation, etc. For Wordle, a crossword puzzle, the word attributes of the answer may have an impact on the number of guesses a player makes. By searching the literature[4], we use the following table 3 to reflect the characteristics of a word.

FREQ The frequency of a word in the whole corpus is usually calculated by dividing the number of occurrences of the word in the corpus by the total number of all words in the corpus. In general, the more frequently words are used, the more familiar they are to people, and the easier it is for people to fill in the Wordle with words they are familiar with. When the frequency of a solution word is higher, it is easier to be guessed and the number of guesses is smaller.

We used data from Mathematica and Google books, and selected all corpus literature from 2020-2022 as the data source to count and calculate the frequency of use corresponding to all 5-letter words.

WIE Information entropy is a concept in information theory that is used to measure the uncertainty of a random variable or the uncertainty of information[5]. The frequency of the 26 English sub letters used to form words is different, such as the letters s, e, and t are more frequently used to form words compared to z, q and j. Therefore, in the Wordle game, when people fill in the letters in each position, they tend to fill in the letters that are used more frequently. According to the literature[6], based on the above, we introduce the concept of word information entropy to represent the composition

Word	WIE	Word	WIE
cater	1.26797168945536	happy	0.707190601600693
kauri	1.10682328140043	foggy	0.612490027812214

Table 4: Partial word information entropy calculation results

of a word. Word information entropy is the entropy value of the probability distribution of the letters composing a word to appear in the whole word bank, which can also be understood as the expectation of the amount of information represented by each letter in a word, and its calculation[7] is shown below.

$$WIE = -p_1 log_2(p_1) - p_2 log_2(p_2) - \cdots - p_n log_2(p_n) \tag{4}$$

n is the length of the word and p_i denotes the probability that the i-th letter appears in the whole word corpus. The larger value of information entropy indicates the higher complexity of the letter combination and the corresponding increase in the difficulty of the word. With the above formula, we can calculate the information entropy of all words of length 5. The results of information entropy calculation for some words are shown as table 4.

NRE In the process of Wordle, every time a word is filled in, feedback will be given and the grid where the letter is located will show three colors: green, yellow and gray; for green letters, the player will still fill in the letter in the same position in the next fill; for gray letters, which do not appear in the word, the player will not choose these letters in the next fill. However, when the result is yellow, the player is more likely to consider changing the position of the letter than to think about whether the letter will appear again. Therefore, the number of repeated letters in the word also affects the difficulty of the problem.

4.2.2 Correlation Analysis

We normalized the processed data from 1 try to 7 or more tries, FREQ, WIE and NRE, and then performed correlation analysis, and finally obtained the correlation heat map as fig 3.

We divided the above three attribute values into two categories according to their high and low values respectively, and calculated the distribution of the proportion of player tries in each category, as shown in following fig 4, fig 5, and fig 6.

From the above figure, it can be seen that the distribution under high FREQ moves in the direction of lower number of tries overall compared with that at low FREQ; for the two variables WIE and NRE, the distribution under the high-level values moves in the direction of higher number of tries overall compared with low level values.

Thus, it shows that the three attributes of FREQ, WIE and NRE will have some influence on the percentage of scores, in which the higher the value of FREQ, the lower the number of tries, which is positively correlated; the higher the value of WIE or NRE, the higher the number of tries, which is negatively correlated, which is also reflected in the heat map.

5 Results Distribution Model

XGBoost is a machine learning model based on the Gradient Boosting Decision Tree (GBDT) algorithm that can be used in regression problems. In a regression problem, XGBoost improves the prediction performance of the model by training multiple regression trees to fit the values of the target variables and by combining the prediction results of multiple regression trees. It has high prediction accuracy: XGBoost is a decision tree-based model with high accuracy and robustness. It can effectively handle large-scale, high-dimensional datasets with very good prediction performance.

5.1 Introduction of XGBoost Model

5.1.1 Optimization Goals

Suppose the model has k decision trees, as

$$\hat{y_i} = \sum_{i=1}^{k} f_k(x_i), f_k \in F \tag{5}$$

For each number of tries in 1 try to 7 or more tries, x_i in the above equation is the percentage of the i-th word at that number of tries, y_i is the predicted value calculated after the mapping relation f_k, and F is the set of all mapping relations, and the optimization objective and loss function are as follows.

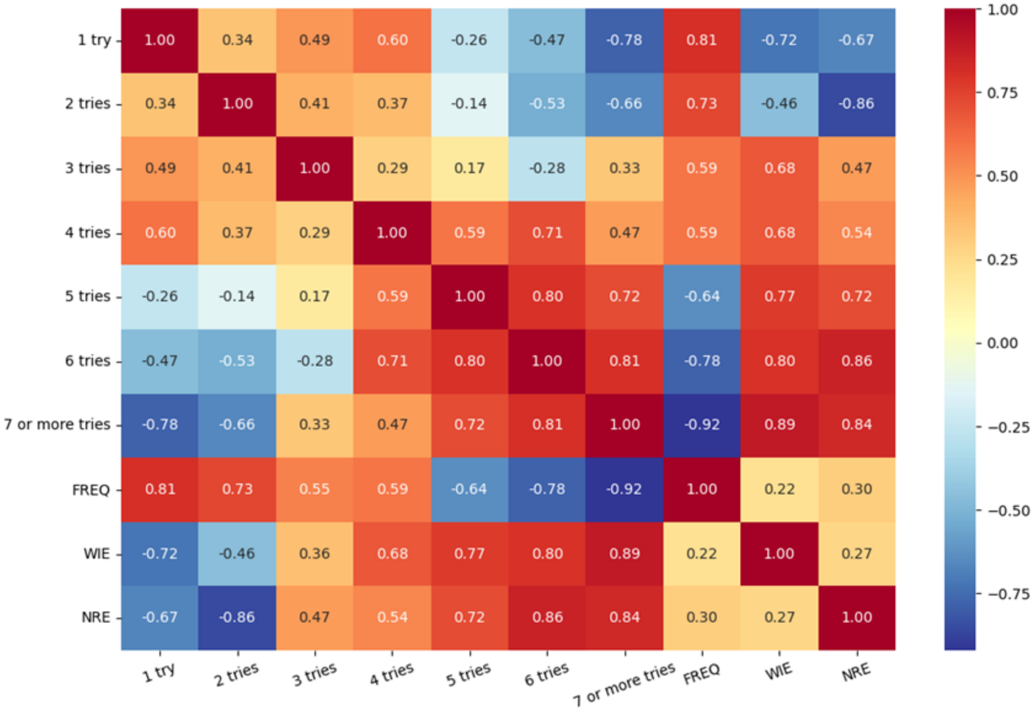


Figure 3: Word attribute relevance heat map

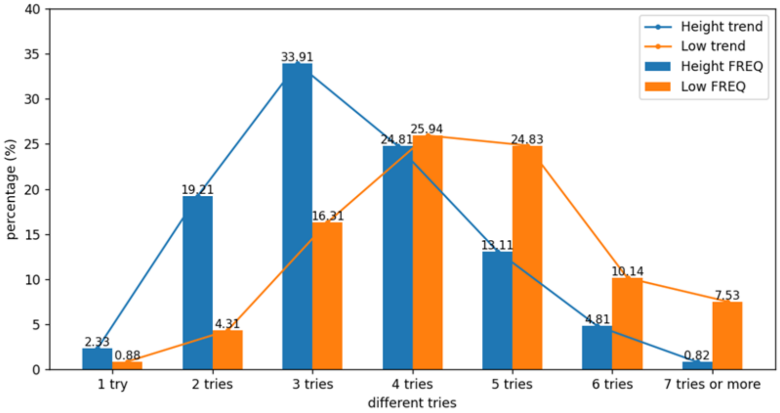


Figure 4: Comparison of High FREQ and Low FREQ

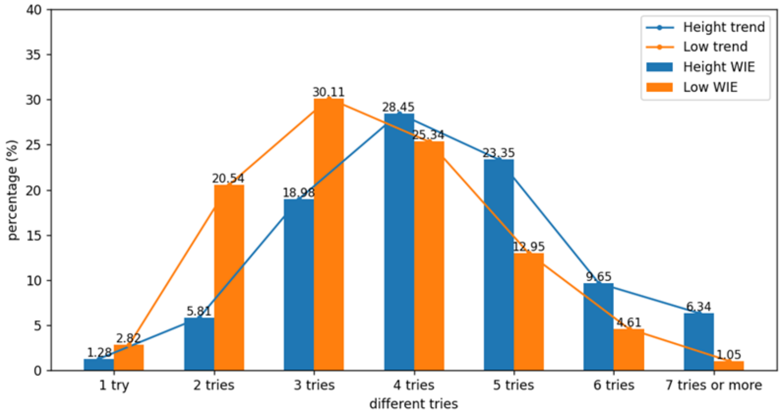


Figure 5: Comparison of High WIE and Low WIE

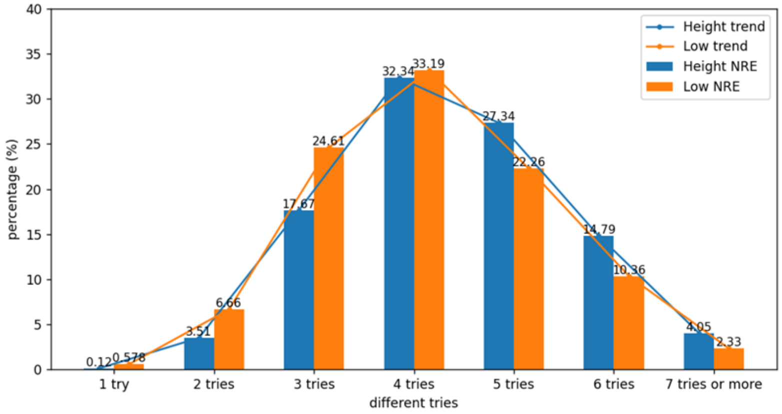


Figure 6: Comparison of High NRE and Low NRE

$$L(t) - \sum_{i=1}^{n} l[y_i, \hat{y}(t-1)_i + f_i(x_i)] + \Omega(f_t) \tag{6}$$

In the above equation, $L(t)$ is the objective function at the t-th iteration, n is the number of samples, I is the loss function, $\hat{y}(t-1)$ is the predicted value of the model at the t-1-th iteration, $f_t(x_i)$ is the newly added function, and $\Omega(f_t)$ is the canonical term.

The canonical terms γ and λ are the two hyper parameters in XGBoost, where γ is the complexity parameter and λ is the penalty factor of the leaf weight ω. The values of γ and λ determine the complexity of the model, and prevent over fitting[8].

Next, a Taylor expansion is performed and the constant term is removed, where g_i and h_i are the first-order and second-order derivatives of the loss function with respect to $\hat{y}(t-1)$, respectively. The expressions are shown as follows.

$$g_i = \frac{\partial l(y_i, \hat{y}_i(t-1))}{\partial \hat{y}_i(t-1)} \tag{7}$$

$$h_i = \frac{\partial^2 l(y_i, \hat{y}_i(t-1))}{\partial (\hat{y}_i(t-1))^2} \tag{8}$$

properties	FREQ	WIE	NRE
value	0.000002437871	1.4797732853992995	3

Table 5: Calculation of "EERIE" word properties

At this point, the values of the loss function of each sample are summed, and each sample falls into a leaf node. If the same leaf node sample is recombined, the objective function can be rewritten as a quadratic function about the fraction of leaf nodes, and the values of the optimal ω_j and the objective function can be obtained directly using the vertex formula.

$$\omega_j = -\frac{G_i}{H_i + \lambda} \tag{9}$$

where H_i and G_i are respectively as follows

$$G_i = \sum_{i \in I_j} g_i \tag{10}$$

$$H_j = \sum_{i \in I_j} h_i \tag{11}$$

Bringing this equation into the objective function, we get

$$L(t) = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \lambda T \tag{12}$$

At this point, we can train the model by minimizing the cost and predict the percentage of the distribution based on the training results for each number of tries.

5.1.2 Model Training

By using Python's xgboost module, seven regression models are built for each of the three selected attributes for seven different tries, and these seven training models are named XGB1 - XGB7. and a random 30% of the data is selected as the test set for model validation.

5.2 Analysis of Model Results

After training and testing the 7 models separately, we found that XGB1 could not predict the size of 1 tries more accurately no matter how we adjusted the model parameters; by studying the distribution of 1 tries in the whole dataset, we found that the dataset of 1 try caused some difficulty in prediction, so we deleted the XGB1 model and used the average value of 1 try in the dataset of 0.5 as We predict the value of 1 try for all words, and the images of the true and predicted values for the remaining six models XGB2 - XGB7 test sets are as follows.

The gray dashed line in the figure indicates that for a certain number of tries, 85% of the data set is concentrated in the interval indicated by the two gray dashed lines, and the 85% Interval Marker provides a good indication of the concentration of the data. We find that as the number of tries increases, the concentration of the data first increases and then decreases. Due to the high volatility of the data as a whole, we consider those predictions that are within 3% of the true value as correct predictions, and thus tally the accuracy of these six model predictions in fig 7.

Fig 8 shows the model prediction accuracy of the six models XGB2 - XGB7. As the number of tries increases, the model prediction accuracy first decreases and then increases because for 4 tries, 5 tries, and 6 tries, their data are more concentrated and it is difficult for the model to predict the data accurately within 3%, so their accuracy is lower. After calculating the mean value of the accuracy of these six models, the overall accuracy is 82.1%.

For the question asking to predict the distribution of players reported for the word "EERIE" on March 1, 2023, the attributes of EERIE were calculated based on the definition of the word attributes above as shown in the table 5.

Substituting the values into the XGBoost regression analysis model, the reported distribution of EERIE on March 1, 2023 was calculated as table 6

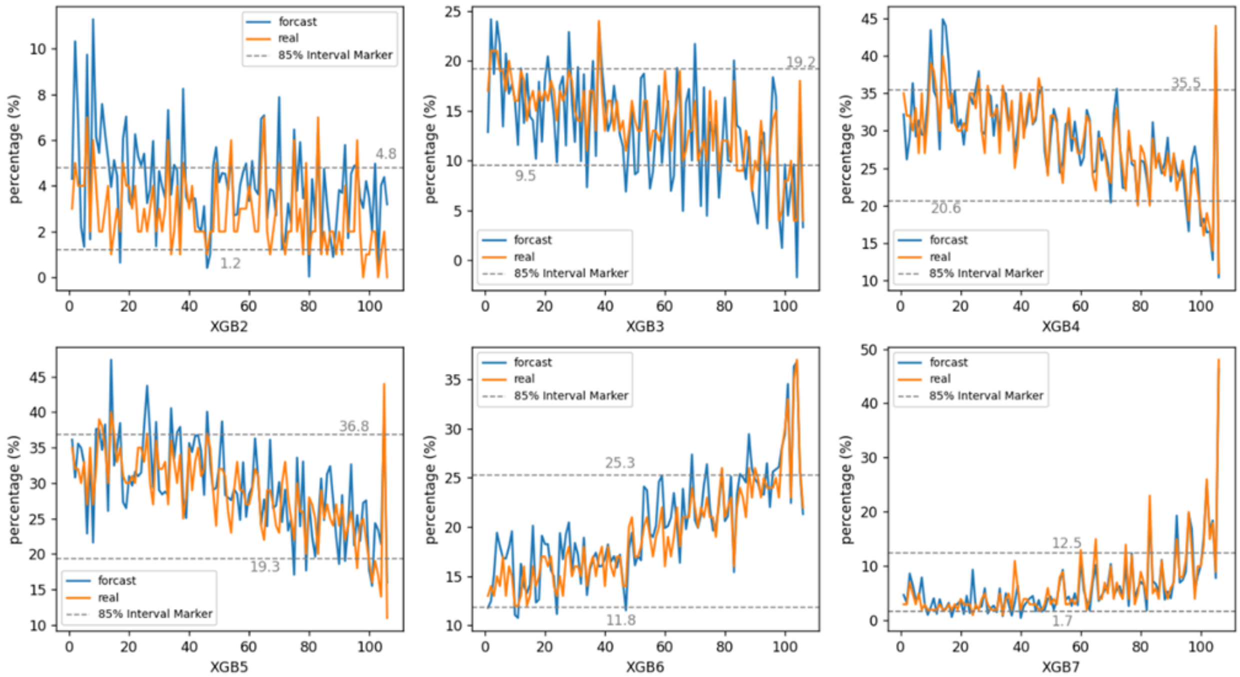


Figure 7: Training results of XGB2-XGB7

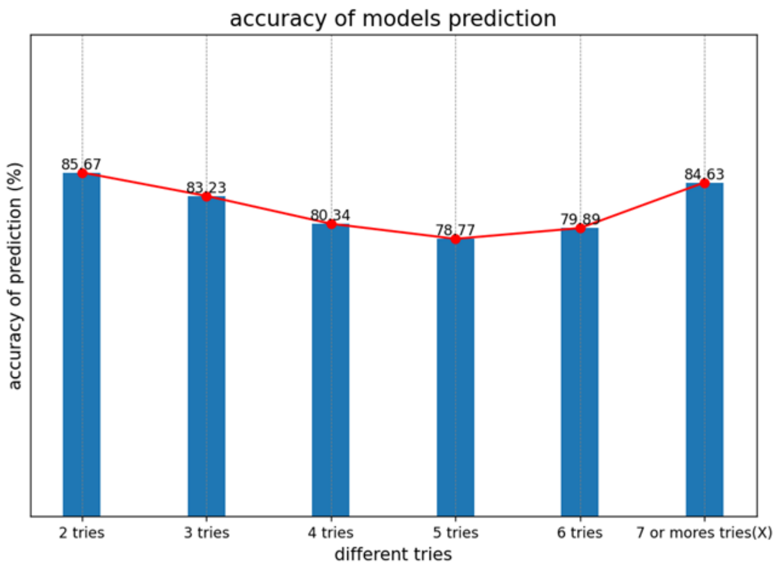


Figure 8: Accuracy of models' prediction

try times	1try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries
value	0.5	2.3	13.8	21.7	29.4	22.3	10

Table 6: The distribution of players corresponding to "EERIE"

```
K-means algorithm steps
$k$ samples are randomly selected as initial cluster class centers
repeat:
    for each sample in the dataset calculate its distance to $k$ cluster class centers
    assigning to the cluster with the smallest distance in the class corresponding to the class center
    for each cluster class, recalculate its cluster class center position
end repeat stop until termination conditions are met
```

Table 7: K-means algorithm steps

	Clustering categories (mean ± standard deviation)				
	Category 1(n=150)	Category 2(n=132)	Category 3(n=73)	F	P
1 try	0.267+0.459	0.795+1.061	0.288+0.456	20.535	0.000***
2 tries	4.033+1.759	9.333+4.077	2.877+1.907	166.258	0.000***
3 tries	20.327+3.481	30.689+3.815	12.808+4.068	589.176	0.000***
4 tries	35.673+3.773	33.697+3.814	25.986+4.511	151.460	0.000***
5 tries	26.34+3.085	17.879+3.123	28.863+5.564	266.781	0.000***
6 tries	11.427+2.955	6.477+2.256	21.329+4.226	561.346	0.000***
7 or more tries	1.933+1.162	1.091+0.937	7.781+6.915	108.121	0.000***

Table 8: Clustering results

Based on the model accuracy described above, we are at least 80% confident in the model's predictions.

6 Solution Word Classification Model

In this section, we build a classification model for solving words based on K-Means clustering algorithm and decision tree classification algorithm. By analyzing the percentage of tries, it was found that there was a classification phenomenon in the percentage trend. Therefore, we used the percentage of each attempt as the basis for classification and K-Means clustering was used to classify the word difficulty. To explore the relationship between word attributes and word classification, we used a decision tree algorithm for training and prediction, and verified the accuracy of the classification. Finally, the difficulty classification of EERIE was performed by the established word-solving classification model.

6.1 Difficulty Classification

In order to classify words in a more considerable way, we used the K-Means[9] clustering algorithm. K-Means algorithm is the most commonly used clustering algorithm, and the main idea is: Given K value and k initial class cluster centroids, each point (i.e., data record) is assigned to the class cluster represented by the nearest class cluster centroid.

After all points are assigned, the centroids of the class clusters are recalculated (averaged) based on all points within a class cluster, and then the iterative steps of assigning points and updating class cluster then iterate through the steps of assigning points and updating the centroids of the class clusters until the change in the centroids of the class clusters is small or the specified number of iterations is reached. The algorithm steps are as table 7.

The comparison graph of the number of clusters was obtained by cluster analysis, as shown in the figure. The horizontal coordinate is the number of clusters, and the vertical coordinate is the sum of squared distances from all samples to the center of clusters, i.e., the sum of squared errors, whose smaller values indicate better clustering. Observing this figure, the decreasing trend tends to level off when the number of clusters k=3. Therefore, the samples are classified into 3 classes according to difficulty by clustering. Fig 9 shows the result.

The specific classifications and clustering centers are shown below. The p-value obtained from the significance test is very close to 0, which indicates that there is a significant difference between the different classifications. Among them, the difficulty of words in category 1 was defined as moderate with 42.26%, the difficulty of words in category 2 was defined as easy with 37.18%, and the difficulty of words in category 3 was defined as difficult with 20.56%. We show results in table 8 and fig 10

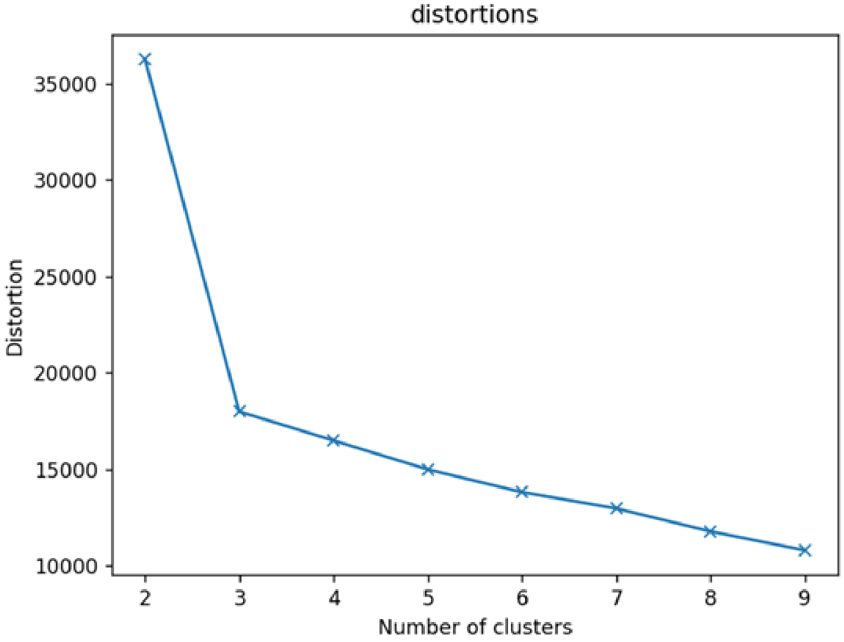


Figure 9: Results of Cluster Quantitative Analysis

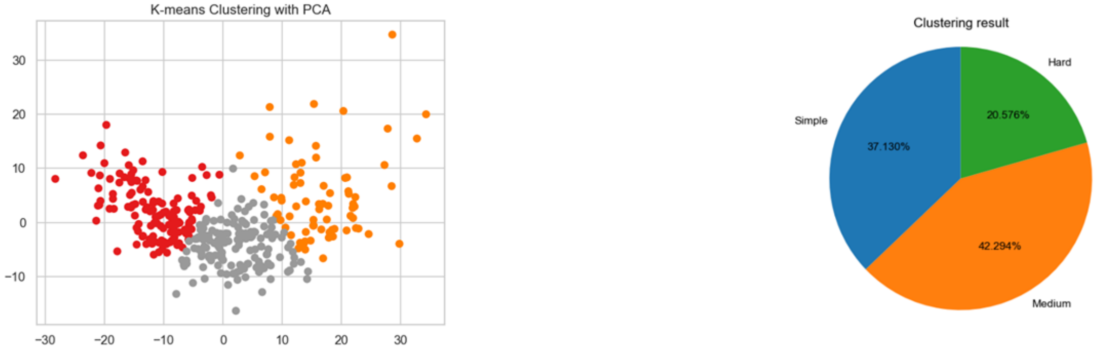


Figure 10: The k-means clustering results

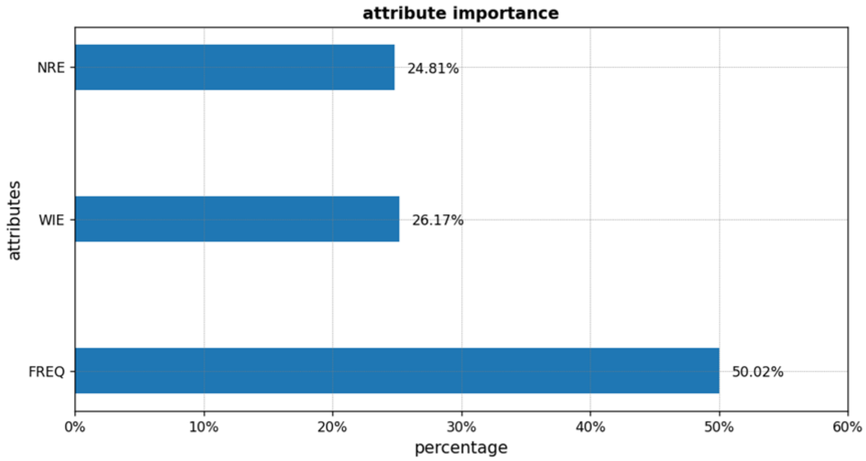


Figure 11: Analysis of the importance of factors

	Accuracy	Recall Rate	Precision	F1
Training Set	0.996	0.996	0.996	0.996
Test Set	0.776	0.776	0.777	0.773

Table 9: Model training results

6.2 Introduction of Decision Tree Model

The above-mentioned words have been classified into three categories according to their difficulty by K-Means clustering. In order to further explore the relationship between this classification method and word attributes, a model was built using the decision tree algorithm with each attribute as an indicator.

Decision trees[10] are a common class of machine learning methods that classify and summarize the attributes exhibited by the data in the training set and find an exact description and classification model for the exhibited attributes, by which the unpredictable future data can be classified.

Decision tree classification algorithm is an inductive learning method based on a given data sample. A top-down recursive approach is used to generate a tree structure given a dataset with known class labels. The decision tree classification algorithm first selects the descriptive attribute with the highest information gain as the branch attribute of the given data set, thus creating a node in the decision tree, and then creates branches according to the different values of the descriptive attribute, and then recursively invokes the above method for each sample subset in each branch to build each child node of the node.

The division stops when all data samples on a branch belong to the same class, forming a leaf node; or when the samples on a branch do not belong to the same class, but there are no remaining descriptive attributes to further divide the data set, and the leaf node is labeled with the class to which most of the samples belong. When classifying a data sample with unknown class label, the class label of the data sample is obtained by judging from the root node downward layer by layer until the leaf node. We show results in fig 11

In order to verify the accuracy of the classification model, we divided the data into training and prediction sets, and the results obtained are shown in the following table. From the data in the table, it can be concluded that the correct rate of classification using decision tree reached 77.6%, and most of the data were correctly predicted with good prediction, which also indicates that the accuracy of the classification model we established is good, and the relationship between the selected word attributes and word difficulty is relatively close. Detail are shown in table 9

6.3 Result of Word Classification

The distribution of the intervals for "EERIE" was predicted in Problem 2 as 0.5, 1.1, 3.2, 12.7, 23.8, 29.9, 28.8, and FREQ, WIE, and NRE for "EERIE" were **0.000002437871**, **1.4797732853992995** and **3**, respectively. The classification model yielded a difficulty for this decoding.

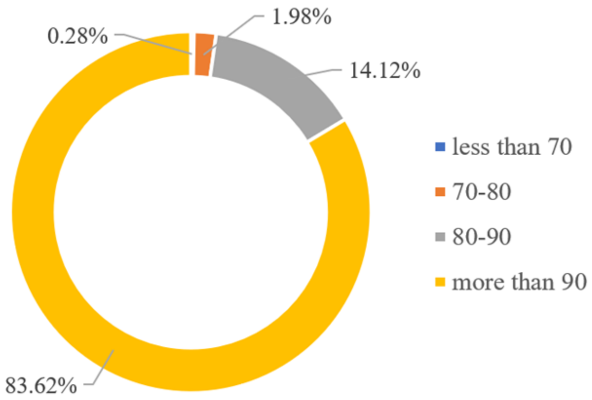


Figure 12: Percentage of players who guessed the correct answer more than 3 times

7 Data Features Mining

The percentage of players who needed 3 or more guesses for each word in the dataset to get it right is shown in the figure. For 83.9% of the words in the dataset, more than 90% of the players needed to guess 3 or more times to get it right, indicating that the words in this game are difficult and challenging for most players.

Fig 12 in the prediction model of the number of reported results shows the change of the number of reported results over time, first increasing in the short term, then decreasing immediately, and finally stabilizing. We believe that Wordle, as a fun and challenging game, was widely sought after in the beginning period, so the number of participants increased rapidly and people were happy to share their results online. Later on, probably due to the excellent difficulty increasing or the loss of players' curiosity, the number of participants sharing gradually decreased, eventually leaving what people call the game's ashes.

8 Sensitivity Analysis

In the ARIMA time series model, the values of each important parameter in the ARIMA model are derived by analyzing the data. In addition to p, d, and q, the coefficients before the independent variables in the equation of the time series model also have a great impact on the forecasts[11]. When the data changes, we expect the coefficients before the independent variables to change accordingly to accommodate the changes in the data and make the prediction results more accurate.

Therefore, we tried to change the values of the coefficients by changing them to 0.3,0.35, 0.4,0.45 and substituting them into the model to observe the changes in the predicted values of the reported score results on March 1, 2023, and the following are the results.

The graph shows that when the coefficient before the independent variable changes, the predicted value changes accordingly, so the sensitivity is better.

9 Advantages and Disadvantages

9.1 Score Prediction Model

- **Advantages:** the model relies on ARIMA time series, only endogenous variables are needed without resorting to other exogenous variables, and the model is a better fit.

- **Disadvantages:** less effective in extremely fast changes. Extremely fast score changes can affect the parameters of ARIMA and produce large errors. Also, ARIMA has limitations in prediction time and is conservative for long time predictions.

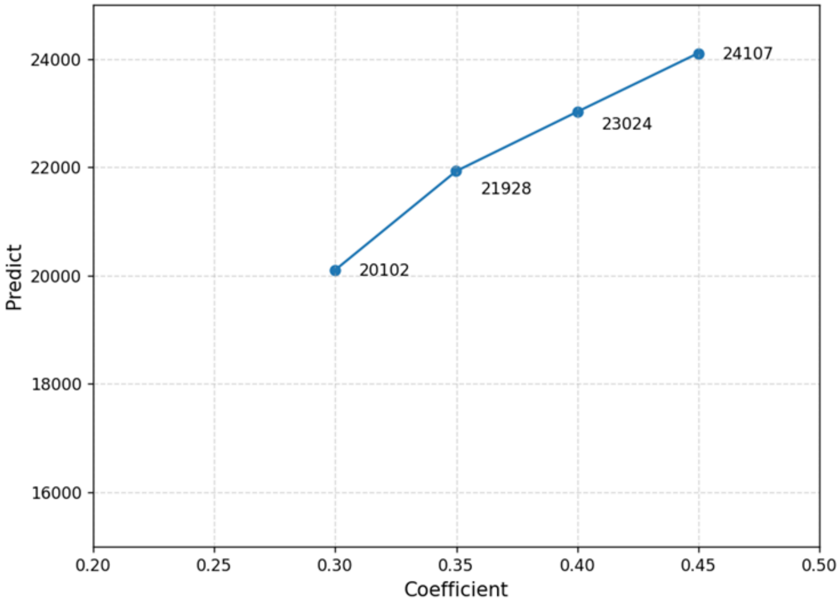


Figure 13: Adjustment of independent variable coefficients

9.2 Results Distribution Model

- **Benefits:** The model is based on XGBoost model with good mathematical theory support. XGBoost is a decision tree based model with high accuracy and robustness. It can effectively handle large-scale, high-dimensional datasets with very good predictive performance.
- **Disadvantages:** XGBoost can easily learn noise in the training data, which can lead to overfitting. XGBoost has many adjustable parameters, such as tree depth, learning rate, etc. If these parameters are not set correctly, it may lead to performance degradation.

9.3 Word Resolution Classification Model

- **Advantages:** The model combines K-Means clustering algorithm and decision tree algorithm, the classification results have objectivity, and the model can be verified by the prediction set for the correctness of the model, which is persuasive.
- **Disadvantages:** When there are more attribute indicators of words, the decision tree algorithm may have more errors and is prone to overfitting; when the amount of data is particularly large, it may have an impact on the clustering results and has uncertainty.

10 Report Letter

To: The Puzzle Editor of the New York Times.

From: Authors

Date: February 20, 2023

Dear The New York Times Wordle Puzzles Editor,

I am writing this letter to express my love and appreciation for the NYT Wordle puzzle game. I love this game and it is one of my must-play games every day.

I want to express my gratitude to you and your team for your efforts and creativity in making this game a very fun and challenging game. Wordle gives us great satisfaction because it not only allows me to exercise my vocabulary and thinking skills, but also makes me feel more confident and accomplished.

First, we built a prediction model for the reported results. We used ARIMA time series to analyze the historical reporting results and predict the reporting scores in the future period, which not only fit the changes of the historical data better, but also gave the predicted values in the future period. The results show that our model is more accurate, and the reported results on March 1, 2023 are expected to be between $[20337, 21673]$. We also selected the frequency of word usage, the information entropy of the word and the number of repeated letters in the word as attributes of the word and performed a correlation analysis with the percentages.

Then, we built a regression model for outcome prediction. Based on the XGBoost model, the distribution of the percentage of tries was predicted by the three attributes of words, and the prediction results were more accurate.

In addition, we also predicted the percentage distribution of the solved word "EERIE", which was 0.5, 2.3, 13.8, 21.7, 29.4, 22.3, and 10 in order.

Next, to classify words by difficulty and to investigate the relationship between the three attributes of words and the difficulty classification, we developed a model for declassifying words. Based on our model, the words were classified into three categories: easy, moderate, and difficult. Moreover, using this model, we also explored the importance of word attributes on classification and found that word frequency, information entropy, and letter repetition rate were in decreasing order of importance. In addition, the accuracy of the model was verified by its prediction of some data types with good accuracy.

Based on the results of our analysis, the following are some of our suggestions.

- New game modes can be developed appropriately to attract the number of participants in the game.
- Difficulty classification and percentage distribution of tries can be predicted in advance, and then players can be given appropriate hints according to the difficulty to have a better game experience.
- We can reduce the number of words in the lexicon that are very rare.

Of course, our model has many shortcomings and limitations, and there is still room for improvement. We hope our model is helpful to you, and we hope you and your team will continue to come up with more interesting and challenging puzzles for us players to enjoy.

Thank you again for your efforts and creativity, and I wish you and your team all the best!

References

[1] TAMP de Leeuw. What language can tell us about the elderly and their behaviour: An analysis of three language features subject to age-related change. B.S. thesis, 2017.

[2] Paul Newbold. Arima model building and the time series analysis approach to forecasting. *Journal of forecasting*, 2(1):23–35, 1983.

[3] Fred L Ramsey. Characterization of the partial autocorrelation function. *The Annals of Statistics*, pages 1296–1301, 1974.

[4] Simon De Deyne and Gert Storms. Word associations: Network and semantic properties. *Behavior research methods*, 40(1):213–231, 2008.

[5] Claude E Shannon. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64, 1951.

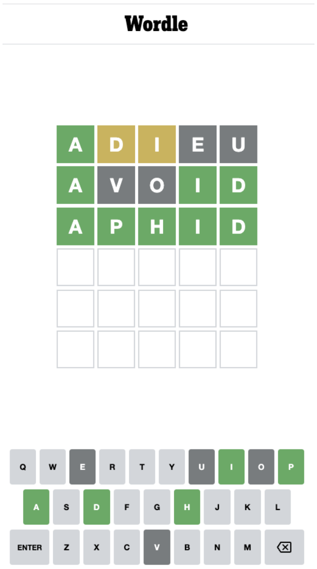
[6] Werner Ebeling, Thorsten Poschel, and Karl-Friedrich Albrecht. Entropy, transinformation and word distribution of information-carrying sequences. *International Journal of Bifurcation and Chaos*, 5(01):51–61, 1995.

[7] Werner Ebeling and Gregoire Nicolis. Word frequency and entropy of symbolic sequences: a dynamical perspective. *Chaos, Solitons & Fractals*, 2(6):635–650, 1992.

[8] Santhanam Ramraj, Nishant Uzir, R Sunil, and Shatadeep Banerjee. Experimenting xgboost algorithm for prediction and classification of different datasets. *International Journal of Control Theory and Applications*, 9(40):651–662, 2016.

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

Figure 14: Image: nytco.com[12]


Wordle

A D I E U

A V O I D

A P H I D

Figure 15: Example Solution of Wordle Puzzle from July 21, 2022[14]

[10] Yan-Yan Song and LU Ying. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130, 2015.

[11] Konstantinos Kalpakis, Dhiral Gada, and Vasundhara Puttagunta. Distance measures for effective clustering of arima time-series. In *Proceedings 2001 IEEE international conference on data mining*, pages 273–280. IEEE, 2001.

[12] The New York Times. Wordle logo. `https://nytco-assets.nytimes.com/2022/08/cropped-Screen-Shot-2022-08-24-at-8.49.39-AM.png`. Accessed: 2022-12-13.

[13] The New York Times. Wordle-the new york times, 2022. Accessed on December 13, 2022.

[14] The New York Times. Wordle-the new york times. *The New York Times*, 2022.

[15] Wordle stats. Twitter, July 2022. Accessed: 2022-07-20.

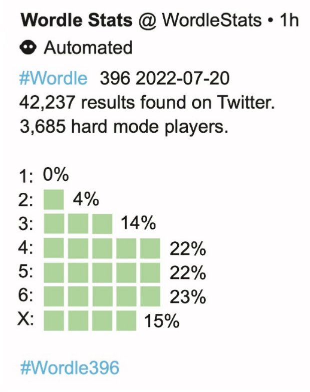
A Problem: Predicting Wordle Results

A.1 Background

Wordle is a popular puzzle currently offered daily by the New York Times. Players try to solve the puzzle by guessing a five-letter word in six tries or less, receiving feedback with every guess. For this version, each guess must be an actual word in English. Guesses that are not recognized as words by the contest are not allowed. Wordle continues to grow in popularity and versions of the game are now available in over 60 languages.

The New York Times website directions for Wordle state that the color of the tiles will change after you submit your word. A yellow tile indicates the letter in that tile is in the word, but it is 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 at all (see Attachment A.3.2)[13]. Figure 15 is an example solution where the correct result was found in three tries.

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. The example in Figure 15 was played in Hard Mode. Many (but not all) users report their scores on Twitter.



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 16: Distribution of the Reported Results for July 20, 2022 to Twitter [15]

For this problem, MCM has generated a file of daily results for January 7, 2022 through December 31, 2022 (see Attachment A.3.1). This file 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 in one try, two tries, three tries, four tries, five tries, six tries, or could not solve the puzzle (indicated by X). For example, in Figure 16 the word on July 20, 2022 was "TRITE" and the results were obtained by mining Twitter. Although the percentages in Figure 16 sum to 100%, in some cases this may not be true due to rounding.

A.2 Requirement

You have been asked by the New York Times to do an analysis of the results in this file to answer several questions.

- The number of reported results vary daily. Develop a model to explain this variation and use your model to create a prediction interval for the number of reported results on March 1, 2023. Do any attributes of the word affect the percentage of scores reported that were played in Hard Mode? If so, how? If not, why not?
- For a given future solution word on a future date, develop a model that allows you to predict the distribution of the reported results. In other words, to predict the associated percentages of (1, 2, 3, 4, 5, 6, X) for a future date. What uncertainties are associated with your model and predictions? Give a specific example of your prediction for the word EERIE on March 1, 2023. How confident are you in your model's prediction?
- Develop and summarize a model to classify solution words by difficulty. Identify the attributes of a given word that are associated with each classification. Using your model, how difficult is the word EERIE? Discuss the accuracy of your classification model.
- List and describe some other interesting features of this data set.

Finally, summarize your results in a one- to two-page letter to the Puzzle Editor of the New York Times.

A.3 Attachments

A.3.1 Data File

All information needed for this problem is given in the problem statement and the data file. You do not need to visit the New York Times website nor Twitter website. There is no additional information to be found on these sites.

Data File Entry Descriptions Date: The date in mm-dd-yyyy (month-day-year) format of a given Wordle puzzle.

Contest number: An index of the Wordle puzzles, beginning with 202 on January 7, 2022.

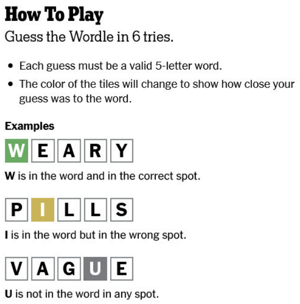
Word: The solution word players are trying to guess on the associated date and contest number.

Number of reported results: The total number scores that were recorded on Twitter that day.

Number in hard mode: The number of scores on Hard mode recorded on Twitter that day.

1 try: The percentage of players solving the puzzle in one guess.

2 tries: The percentage of players solving the puzzle in two guesses.



How To Play

Guess the Wordle in 6 tries.

- Each guess must be a valid 5-letter word.
- The color of the tiles will change to show how close your guess was to the word.

Examples

W E A R Y

W is in the word and in the correct spot.

P I L L S

I is in the word but in the wrong spot.

V A G U E

U is not in the word in any spot.

3 tries: The percentage of players solving the puzzle in three guesses.

4 tries: The percentage of players solving the puzzle in four guesses.

5 tries: The percentage of players solving the puzzle in five guesses.

6 tries: The percentage of players solving the puzzle in six guesses.

7 or more tries (X): The percentage of players that could not solve the puzzle in six or fewer tries. Note: the percentages may not always sum to 100% due to rounding.

A.3.2 Directions of Wordle posted to the New York Times website [13]

Glossary

New York Times: A daily newspaper based in New York City, New York, USA published in print and online.

Twitter: A social networking site that allows users to broadcast short posts of no more than 280 characters (increased from initial 140 characters).

Solve (the Wordle puzzle): Enter the correct letters in the correct order to form the Wordle word of the day.