| Problem Chosen | 2023 MCM/ICM Summary Sheet | Team Control Number |
|---|---|---|
| C | | 2318770 |

# Predicting Wordle Results


**Summary**

Wordle is a highly popular crossword puzzle on the ***New York Times***. Through the five-letter words, the times players need to guess correctly and the number of players in hard mode, we analyze and predict the data of Wordle.

**For problem 1**, we preprocess the original data firstly, including correcting non-5-letter words, using **Cubic Spline Interpolation** and **Hermitian Interpolation** for abnormal value in reported results. Following the processed data, we apply the number of results in 2022 to the forecast model combining **ARIMA** and **GM(1,1)**, and find it present a descending trend. Then we predict that the number of reported results on March 1, 2023 will be between [**8067, 12406**]. Finally, we use **Ridge Regression** to explore the relationship between the percentage of people in hard mode and the attributes of word such as vowels, repeated letters, part of speech and word frequency. It shows that the regression coefficients corresponding to each attributes are less than 0.005, so we infer that hard mode percentage and word's attributes is **irrelevant**.

**For problem 2**, we use **One-hot**, **Bag of words**, **N-gram** in **NLP** respectively to encode words, applying which to **BP Neural Network** training to predict percentages of (1, 2, 3, 4, 5, 6, X). Then we find that the encoding method and word frequency will affect the accuracy of the model. The optimal number of iterations under the one-hot encoding method is 198, while under the N-gram is 545. Subsequently, for the given word EERIE, we predict that its percentages are **(0.84%, 5.50%, 10.63%, 18.35%, 28.48%, 26.13%, 10.05%)**. Finally, we test the accuracy of the model by calculating the average **Euclidean Distance** between the predicted and real value. The result shows that the average value of the Euclidean Distance of all words is , therefore the model is quite accurate.

**For problem 3**, we perform **Systematic Clustering** on 359 words, and determine the optimal number of clusters is 3 through the Elbow Rule, and then use $K = 3$ for **Kmeans** to classify words into easy, middle and hard types. Using XGBoost, Random Forest, and Adaboost as base classifiers, we establish an **Ensemble Learning** model of **Plurality Voting**, and the clustering labels are used to train the model. We analyze the attributes of different types and find that hard words have repeated letters, middle words start with vowels and easy words are prefixed with aspirated *ch*, *th*, etc. We use the model to determine the difficulty of EERIE and it shows that EERIE is of high difficulty. Finally, we make the sensitivity analysis on the ensemble learning model, and its classification accuracy is stable at around 85%, so the accuracy of the model is high.

**For problem 4**, We perform descriptive statistics on dataset and find that the number of players rose urgently in short term, and after March 2022, the number of players gradually decreased, while the percentage of hard mode increased by degree. At the same time, times of guesses obey the normal distribution, that is, most players need 3 to 5 tries to guess the word correctly.



**Keywords: Arima, Ridge Regression, NLP, BP Neural Network, Plurality Voting**

# Contents

# 1 Introduction

## 1.1 Background and Problem Statement

Wordle is a popular puzzle currently offered daily by the ***New York Times***. In this game, there are $5X6$ grids where player can enter words. Each line consists of 5 grids to form a word, and each time a word is entered, click ENTER to verify the result. Different results will change the background color of the grid, and each time a word is input, the result will be run once. Player need to judge whether to re-enter the word according to the prompt of the color of the square until the 6 chances are exhausted or the answer is guessed.

Considering the background of the question and the given data set, we decide to focus on the following questions:

- Develop a model to explain the change in the number of reported results daily and use the model to predict the results on March 1, 2023. In addition, analysis whether and how the attributes of the word will affect the percentage of scores reported that were played in Hard Mode.
- Develop a model to predict the associated percentages of $(1, 2, 3, 4, 5, 6, X)$ for a future date and give our prediction for the word **EERIE** on March 1, 2023. Also, we should identify the uncertainties that are associated with our model and predictions.
- Develop a model to classify solution words by difficulty and discuss the accuracy of the classification model. Then, use our model to determine the difficulty level for the world **EERIE** and explain the attributes of the world word in each classification.
- Summarize our results and write a one or two-page letter to the Puzzle Editor of the ***New York Times***.

## 1.2 Our Works

Firstly, we combine Arima and GM(1,1) to make a forecast for the interval of number of reported results on March 1, 2022. Due to the differences in the predicted results of the two prediction models, we take the smaller value as the minimum value of the interval, and the larger value as the maximum value of the interval so as to obtain a result interval of the reported number. At the same time, we explore the relationship between the attributes of words, including the number of vowels, repeated letters in words, word frequency, word part of speech, and the proportion of people in hard mode.

Secondly, we develop the BP Neural Network Model to predict the distribution of the reported results, and before applying the model, we encode words into data through One-hot, Bag-of-world and N-gram respectively. The neural network model is then applied to predict the associated percentages of the word ERRIE on March 1, 2022 (1, 2, 3, 4, 5, 6, X). Among them, the encoding method and word frequency will affect the prediction of the distribution. Finally, we get the accuracy of our model through the goodness of fit.

Thirdly, we perform Systematic Clustering on 359 words, and determine the optimal number of clusters through the Elbow Rule, based on which we apply Kmeans to classify words into different types. In order to improve the accuracy of the model, we use XGBoost, Random Forest,

and Adaboost as base classifiers to establish a Plurality Voting Ensemble Learning Model, using the clustering label results to train the model. Subsequently, we analyze the attributes of different types and use the ensemble learning model to determine the difficulty of EERIE. Finally, we perform a sensitivity analysis to discuss the accuracy of our model.

At last, we conduct descriptive statistics on the dataset, analyzing the changes in the number of reported results, number in hard mode over different time and the distribution of times in which players guess the correct word.

The flow chart of our idea is shown in figure below:

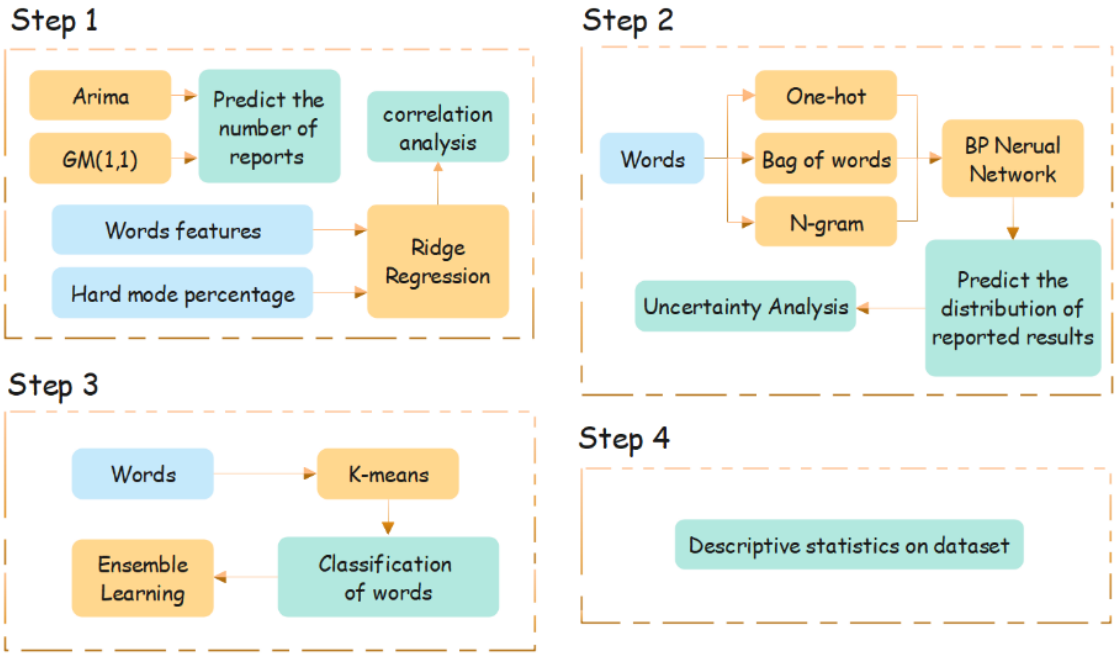


Figure 1: *Flow chart*

# 2 Assumptions and Notations

## 2.1 Assumptions

To simplify our problems, we make the following basic assumptions, each of which is adequately justified.

- The player can't get any information about the correct answer of the day before choosing hard mode.
- The difficulty of a word is only related to the word attribute, not the date of its occurrence.
- The frequency of words is exactly in accordance with Zipf's law.
- All data are authentic.

## 2.2 Notations

Table 1: *Notations.*

| Symbols | Definition |
|---|---|
| $K$ | number of clusters |
| $\bar{d}$ | the average of Euclidean distance |
| $V_{\text{one}}(i)$ | the $i-th$ letter's $one-hot$ encoded vector |
| $V_{\text{N}}(i)$ | the $i-th$ word's $N-gram$ encoded vector |
| $V_{\text{b}}(i)$ | the $i-th$ word's Bag of words encoded vector |
| $p_i$ | the percentage of players solving the puzzle in $i-th$ guess |

# 3 Data Preprocessing

- **Word Correction:**

By traversing the words in data file, we observed that there were three days of words that were not five-letter words, so we searched for the real words of these three days on the New York Times and corrected the words in data file. The results are shown in the table below.

Table 2: *Correction of the wrong words*

| Date | Before | After |
|---|---|---|
| 2022-04-29 | tash | trash |
| 2022-11-26 | clen | clean |
| 2022-12-16 | rprobe | probe |

- **Outlier Processing:**

We observed that the Number of reported results was abnormal on 2022-11-30. In order to avoid the impact on subsequent analysis, we used the Number of reported results in November 2022 to perform cubic spline interpolation and three Hermitian interpolation on 2022-11-30, then took the average of the values obtained by the two as the Number of reported results of 2022-11-30. The result is shown in the table below.

Table 3: *Processing of the abnormal data*

| Date | Word | reported results | Hermite | Cubic Spline | Average |
|---|---|---|---|---|---|
| 2022-11-30 | study | 2569 | 24181 | 25344 | 24763 |

- **Data Normalization:**

Due to rounding, the percentages of players solving the puzzle may not always sum to 100%. Therefore, we normalized the percentage of players solving the puzzle, where $i$ is the contest number and $j$ is the times of try.

$$b_{ij} = a_{ij} / \sqrt{\sum_{i=1}^{m} a_{ij}{}^2} (i = 1, \ldots m; j = 1, 2, \ldots, 7) \tag{1}$$

# 4 Model 1: Prediction for the number of reported results

## 4.1 Model Establishment

According to the data preprocessing and the report generated by MCM, we have continuous and nearly accurate data of the Number of reported results from January 7, 2022 to December 31, 2022, based on which, we predict the data range of future reported results.

In order to improve the accuracy of the results, we consider combining Arima(Autoregressive Integrated Moving Average Model)and GM(1,1) to predict the results. Since the prediction results of the two models on March 1, 2023 are independent data points, and the two prediction results are different, we take the smaller of the two as the minimum value of the result range, and the larger as the maximum value of the result range. That is, the predicted range of results on March 1, 2022 is expressed as:

$$[\min(\text{ Arima(result)}), GM(\text{result})), \max(\text{Arima(result)}, GM(\text{result}))] \tag{2}$$

Where Arima(result) is the prediction result of ARIMA, $GM$(result) is the prediction result of GM(1,1).

First, we use ARIMA(Autoregressive Integrated Moving Average Model)to predict the Number of results reported on March 1, 2022.

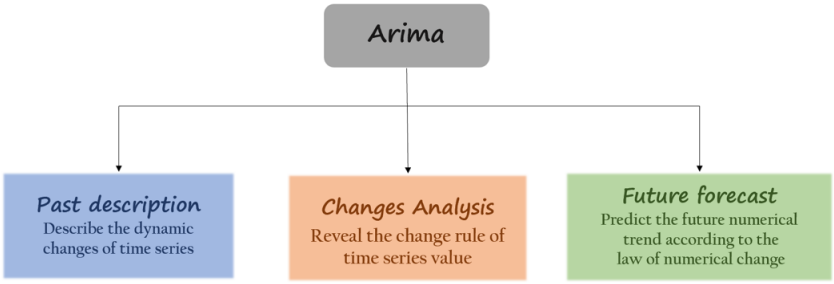


Figure 2: *Brief introduction of Arima*

ARIMA have the following three main parts:

**Autoregressive(AR)**: Analyze previous values in data and make assumptions. The number of results per day in 2022 is considered as an "evolutionary variable" in the data set. This model is established by regressing the number of results submitted per day according to its own lag value (i.e. a priori value). The formula of AR model is as follow:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_p y_{t-p} + \epsilon_t \tag{3}$$

**Integrated(I)**: Since we need stable data when applying difference calculation, the data need to be comprehensive. That is, we describe the difference of the original data to allow the time series to become stable.

**Moving average(MA)**: Moving average model focuses on the accumulation of error terms in autoregressive model. The order of MA model is recorded as q value. The formula is as follows:

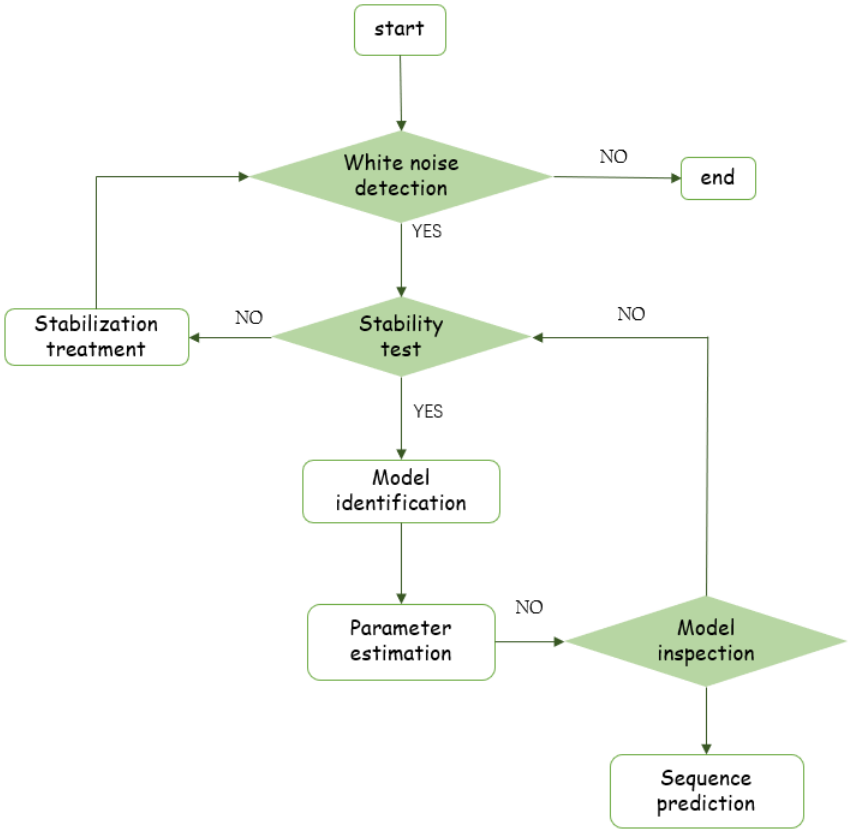


Figure 3: *Algorithm flow chart of Arima*

Then, we forecast the results of March 1, 2023 through **GM (1,1)**, the formula as follow:

$$\hat{x}^{(1)}(k+1) = \left(x^{(0)}(1) - \frac{\hat{b}}{\hat{a}}\right) e^{-ak} + \frac{\hat{b}}{\hat{a}} \tag{4}$$

## 4.2 Model Sovling and Results

## 4.3 Predicted interval range

- **Prediction through Arima Model**

We get the results by Arima Model firstly. Comparing time series forecast data results with raw data, we get the comparison chart as follows. It can be inferred from the figure that the time series forecast results are very close to the true value, therefore, the prediction results are very accurate.

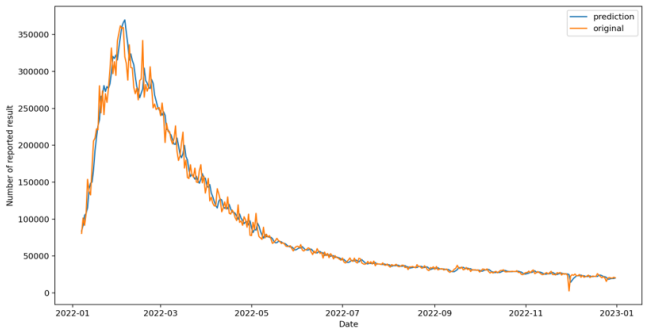


Figure 4: *Prediction results through Arima*

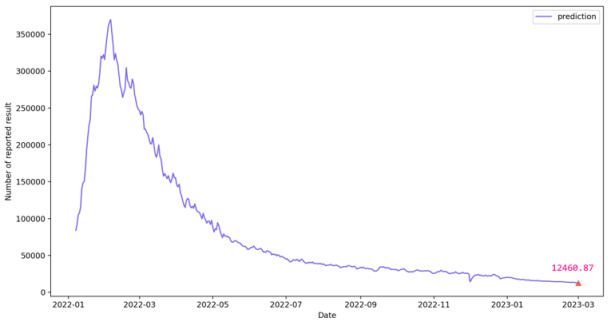


Figure 5: *Prediction from Jan 2022 to Mar 2023 through Arima*

Arima Model predicts that the number of reported result is 12460.87 on March 1, 2023, rounded to 12461.

- **Prediction through GM(1,1)**

According to the GM(1,1) algorithm and the data from January 2022 to March 2022 given by Wordle, we get the prediction equation as follows:

$$y = -33616864.6e^{-0.009366538(t-44568)} + 33697494.6 \tag{5}$$

In order to show the degree of model fitting and model accuracy more clearly, we draw a comparison between the GM(1,1) predicted sequence and the actual sequence from January 2022 to December 31, 2022. Advance the time prediction to March 2023, and the number of reported result on March 1, 2023 is predicted to be 8066.26, which is rounded to 8067 person-times.GM(1,1). The prediction curve for number of reported result on March 1, 2023 is as follows:

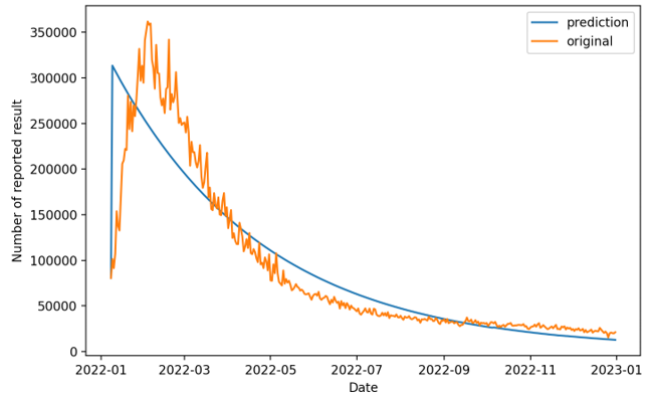


Figure 6: *Prediction results through GM(1,1)*

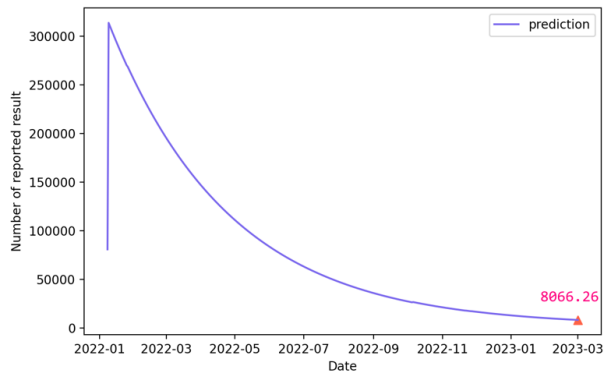


Figure 7: *Prediction from Jan 2022 to Mar 2023 through GM(1,1)*

- **Predicted interval range**

In order to explore and obtain the prediction interval better, we compared the two sets of prediction results of ARIMA and GM(1,1) visually. Then we obtained the following comparison chart. The final prediction interval corresponds to the area between the blue and orange curves.

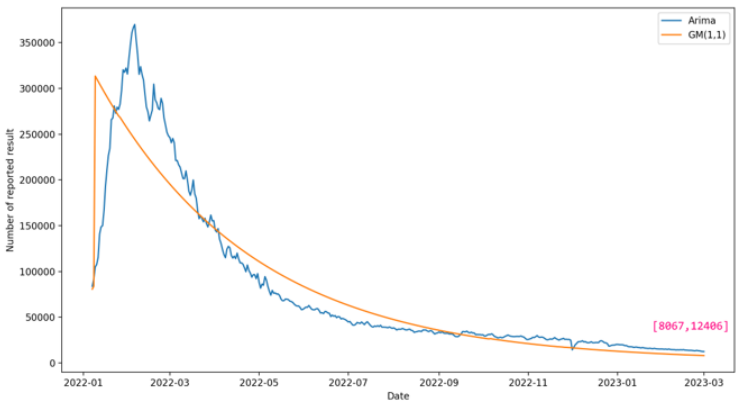


Figure 8: *Difference between ARIMA and GM(1,1)*

The GM(1,1) equation belongs to the prediction of a smooth curve. It can be seen that the prediction result of the GM(1,1) curve shows a relatively rapid decline trend while the prediction result of Arima declines slowly. The final result of GM(1,1) on March 1,2022 is lower, while the Arima model's result is higher, so we consider **[8067, 12461]** as the prediction result interval. Moreover, we believe that the GM(1,1) results are relatively ideal while the time series prediction results are closer to the real value of the data, so the results should be closer to the upper bound of the interval.

# 5 Model 2: Effect of attributes in hard mode

## 5.1 Ridge Regression Model Establition

Based on all the data in 2022, we explored whether various attributes of a word affect the percentage of people in the data who choose the hard mode. We consider the following four word attributes: vowels, repetitions, and word frequency.

- **Vowel ratio** $vol$**:** Consider the percentage of vowels in five letters, for example, the word apple, which contains the vowels $a$ and $e$, has two vowels, therefor, $v_{vol_w} = \frac{2}{5}$.
- **Number of repeated letters** $times$**:** Five-letter words often have three cases where the letter is not repeated, the letter is repeated twice, and the letter is repeated three times. We use the word given by wordle in 2022 as the Thesaurus to calculate the ratio of no repeated letters, repeated letters twice, and three repeated letters to the total number of words. We use the probability of repetition as a measure of the number of repeated letters. For example, The letter $m$ occurs three times in the word $Mummy$, therefore, for the word $Mummy$, $times = 3$.
- **Word frequency** $freq$**:** According to Zipf's law [1], we roughly classify word frequency. Words are divided into three categories, namely high-frequency, medium-frequency and low-frequency words. At the same time, we quantify attributes into data. If the word is a high frequency word, then $freq = 2$. If it is a medium frequency word, then $freq = 1$. If it is a low frequency word, then $freq = 0$.

- **Part of speech** *par***:** Based on the Thesaurus given by Wordle 2022, We have four parts of speech words including nouns, verbs, adjectives and adverbs. For polysemous words, we use the most commonly used attributes of the word as the part of speech of the word. If the word is a noun, $par = 1$. If it is a verb, $par = 2$. If it is a adjectives$par = 3$. If it is a adverbs$par = 4$.

After quantifying the data for the four attributes, we explored their relationship to the percentage of people choosing the difficult mode by Ridge regression. We take the attributes of the words as the independent variable, and select the percentage of people in difficult mode as the dependent variable, fiting the Ridge regression parametric equation.

Ridge regression,which is also known as L2-Norm, is an improvement to linear regression. On the basis of it, an L2 regularization term is added. By adding a penalty term for adding bias to the estimate, we get better estimation. At the same time, L2 solves the problem of model overfitting and enhances the fitting of existing data point data. The loss function of ridge regression adds a regularization term $\lambda\|\theta\|^2$ on the basis of linear regression, supposing the number of data features is $n$, the number of samples is $m$, then we got the loss function

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}\left(\left(h_\theta\left(x_i\right) - y_i\right)^2 + \lambda\sum_{i=1}^{n}\theta_i^2\right) \tag{6}$$

$$h_e(x) = \theta_1 \cdot \text{ vowel } + \theta_2 \cdot \text{ times } + \theta_3 \cdot \text{ freq } + \theta_4 \cdot \text{ par} \tag{7}$$

## 5.2 Ridge Regression Model Sovling and Results

Firstly, we quantify the word attributes to the form of data, and some of the data quantification results are as follows

Table 4: *Data quantification results*

| **Word** | **Vowel** | **Repeated Num** | **Word frequency** | **Part of speech** |
|---|---|---|---|---|
| manly | 1 | 0.721448 | 0 | 3 |
| molar | 2 | 0.721448 | 1 | 1 |
| havoc | 2 | 0.721448 | 0 | 1 |
| impel | 2 | 0.721448 | 0 | 2 |

Then we use Ridge Regression to establish the relationship between the proportion of the number of people who choose the difficult mode and the word attributes, and the coefficients of Ridge Regression corresponding to each attributes are obtained through the model solution as follows:

Table 5: *Coefficients of each attributes*

| **vowel** | **repeat_num** | **frequency** | **speech** | **bias** |
|---|---|---|---|---|
| 0.00139691 | -0.00110645 | -0.00181398 | -0.00016018 | 0.076 |

According to the ridge regression prediction results, we can see that the proportion of the number of people in hard mode has tiny relationship with the vowel proportion of the word, the number of repeated letters, the word frequency and the part of speech of the words.

- **Positive or Negative Correlation**

The proportion of people who choose the hard mode is slightly positively correlated with the vowels, that is, the more vowels, the more people choose the hard mode. And it is slightly negatively correlated with the other three attributes, that is, the more repeated letters, the higher the word frequency, and the more inclined the part of speech is to adjectives and adverbs, the number of people in hard mode will be slightly reduced.

- **Degree of relevance**

According to the regression equation of ridge regression, the ridge regression coefficients corresponding to the four attributes are all less than 0.005, which is at a relatively low level. Therefore, we believe that the attributes of a word have little to do with the percentage of whether the player chooses to play the game in hard mode or not.

# 6 Model 3: BP Neural Network Prediction Based on NLP

## 6.1 Corpus preprocessing

Based on the model assumptions, we consider that the only factor affecting the distribution of the reported result is the word itself, not the date and contest number. However, words cannot be directly used as training sets for traditional model fitting, so encoding conversion is required. Therefore, we introduced three encoding methods, One-hot, Bag of words, and N-gram, which are common in the corpus preprocessing part of NLP, to encode and convert words respectively.

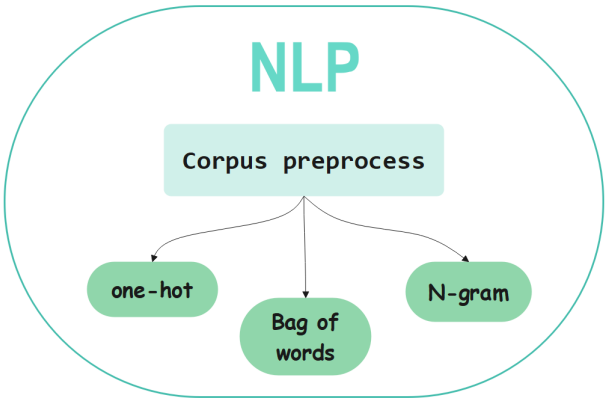


Figure 9: *Three encoding methods*

- **One-hot:** One-hot uses letters as the main part to encode, and converts each letter into a vector. For each vector, only one dimension can be the value of 1, while others are all 0.

Since a letter has 26 types, we map the letters a to z to 26 dimensions respectively to form a 26-dimensional vector. Therefore, in this article, for a word with the length of 5, it can be described as 5 26-dimensional vector. An example of expressing part of the letter with one-hot is given below:

$$\begin{aligned} V_{\text{one}}(a) &= [1_a, 0, 0, \ldots, 0, 0] \\ V_{\text{ore}}(b) &= [0, 1_b, 0, \ldots, 0, 0] \\ V_{\text{ore}}(c) &= [0, 0, 1_c, \ldots, 0, 0] \\ &\cdots\cdots \\ V_{\text{one}}(y) &= [0, 0, 0, \ldots, 1_y, 0] \\ V_{\text{ore}}(z) &= [0, 0, 0, \ldots, 0, 1_z] \end{aligned} \tag{8}$$

- **Bag of words:** Based on one-hot encoding, Bag of words uses word as the main part to encode. We merge all the vectors vertically to obtain one 26-dimensional vector, and the value of each dimension in one vector represents the times of occurrences of the corresponding letter in the word. Taking the word mummy as an example, since the letter m appears 3 times in the word, and u and y appear 1 time each, so the Bag of words encoding of the word is expressed as below:

$$V_b(mummy) = [\ldots 0, 3_m, 0, \ldots, 1_v, 0, 0, 0, 1_v, 0] \tag{9}$$

- **N-gram:** Based on Bag of words, N-gram uses the affixes composed of N consecutive letters in a word as the main part to encode the affixes. Taking N=2 as example, for the word truth, the encoding method of N-gram is:

$$V_N(truth) = \{V_b(tr), V_b(ru), V_b(ut), V_b(th)\} \tag{10}$$

## 6.2 BP Neural Network

### 6.2.1 Principle of model

After corpus preprocessing, we get encoded data which can be directly used for model training. In order to predict the distribution of the reported results, we use the encoded data as input, and use the distribution of the reported results as output to perform BP neural network fitting, then use it to predict the future words' reported distribution.

**Step1 Forward propagation** The forward propagation process is as follows

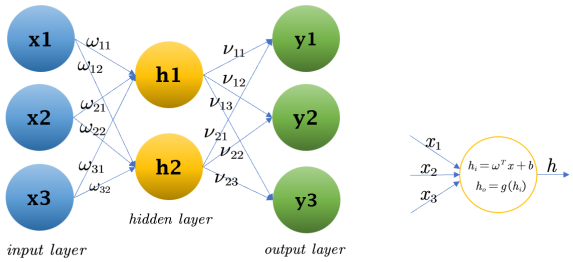


Figure 10: *Forward propagation process*

The input layer is the transformed encoding, and the output layer is the distribution percentage corresponding to the word. The weight is the connection between two nodes and the weight is randomly initialized and revised continuously in the future.

**Step2 Back propagation** The parameters of the weight matrix are optimized through backpropagation, and the iterative weight is calculated through the error between the value $y_0$ output by the network and the real value $y$, and the coefficient $w$is continuously corrected.

**Step3** Repeat the above two steps until the error is acceptable or the iteration times reaches the maximum value. After each forward propagation and back propagation, the weights in the network will be updated. Finally the optimal weight matrix $w$ will be obtained.

### 6.2.2 Goodness of fit

We take the encoding of 359 words in dataset as input, taking 80% as the training set, and 20% as the testing set. Then we get the neural network with optimal weights by model training. The (the Goodness of fit) of each encoding is as below:

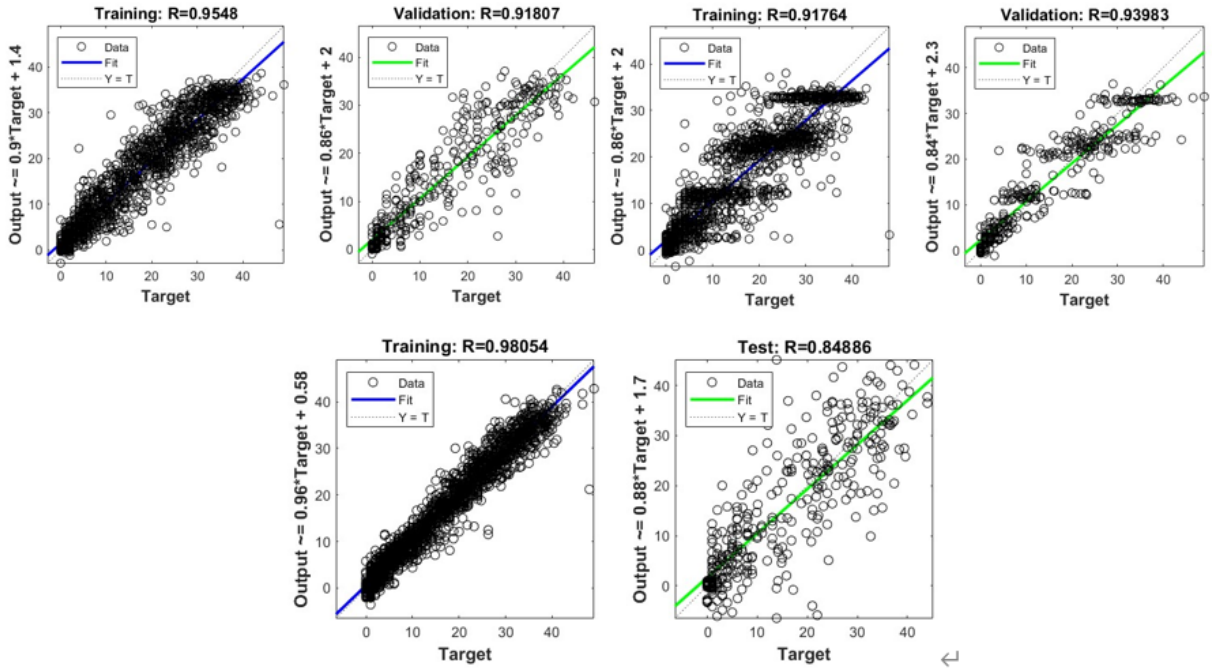


Figure 11: *The Goodness of fit of each encoding*

Although one-hot encoding requires a large number of vectors, the word features it reflects are also more specific. N-gram takes the impact of word affixes into account on our prediction, so it is better on features. Therefore, the overall is at a high level.

The Bag of words encoding merges multidimensional vectors vertically. Although the input features are simplified, the process of merging vectors loses the order information of the letters themselves. Therefore, even if the of the training set is at a high level of 0.98, the of the testing set is only 0.849, which means there is still room for improvement.

## 6.3 Uncertainties of model

- **Word encoding**

Different encoding methods of words will lead to different model training accuracy. The following figure shows the iterative process of one-hot's and N-gram's MSE respectively. It is obviously that the one-hot encoding reaches the MSE of 31.54 in only the eighth iteration when fitting the model, while the N-gram encoding method achieved the MSE of 22.86 at the 42nd iteration.

Even though one-hot encoding requires fewer iterations, its MSE is slightly higher than that of N-gram encoding. This is because one-hot reflects comprehensive information such as the order of letters in a word and the times of its occurrences, which means its characteristics are obvious. So the iterations required for the model to train features is less, while N-gram is more inclined to reflect deeper information such as continuous syllables, prefixes and suffixes in words (such as th, sh, ch and others), the feature quantity is large so feature training requires more iterations and has a smaller MSE. Thus the different encoding of words will lead to the uncertainties of the model's prediction.

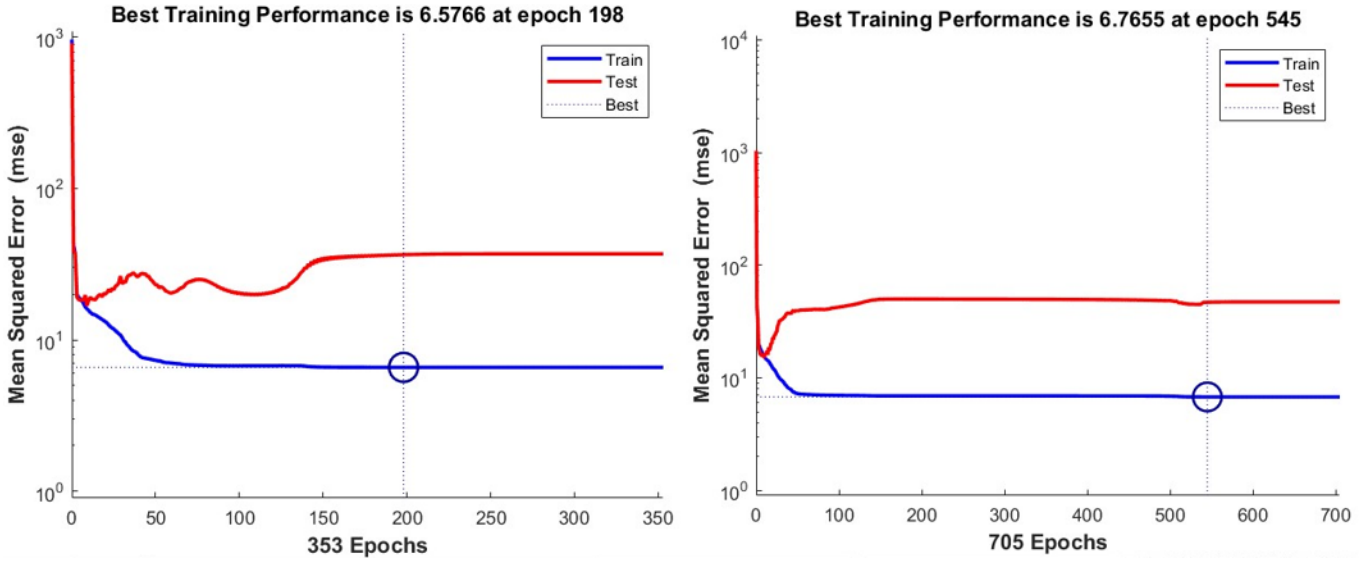


Figure 12: *Iterations in each meathods*

- **Word frequency**

In addition to encoding ways, the frequency of daily use of the words can also affect the accuracy of the model's predictions. For some high-frequency words that appear in daily life, even if the word itself has complex structural features, its high-frequency usage will make it easier for people to associate the word and complete the game with fewer tries.

## 6.4 The prediction of EERIE

We use the NLP-based BP neural network model to predict the percentage distribution of the word EERIE on March 1, 2023, and the final predicted distribution of reported results is shown in the figure below:

Table 6: *Distribution of reported results*

| 1 try | 2 tries | 3 tries | 4 tries | 5 tries | 6 tries | 7 or more |
|---|---|---|---|---|---|---|
| 0.84% | 5.50% | 10.63% | 18.35% | 28.48% | 26.13% | 10.05% |

According to the prediction results, we know that the percentages of trying 5 times, 6 times and more than 6 times are 28.48%, 26.13% and 10.05% respectively, all at a relatively high level. The word EERIE consists of three repeated letters E, which is the same as the word MUMMY in the dataset, which contains three repeated letters M. For the word MUMMY, the percentages of people trying 5 times, 6 times and more than 6 times are 27%, 37% and 18% respectively, which are also at a high level, similar to the predictive features of the word EERIE. Therefore, we consider that the model has high credibility for the distribution prediction of the word EERIE.

## 6.5 Accuracy Test Based on Average Euclidean Distance

In order to test the prediction accuracy, we calculate the average Euclidean distance between the real distribution of the words in dataset and the predicted distribution provided by the neural network, in order to measure the error between the prediction result and the real result. For the average Euclidean distance between the predicted distribution and the true distribution, we calculate it by the following ways:

$$d = \frac{\sqrt{\sum_{i=1}^{7} (p_i - q_i)^2}}{7} \tag{11}$$

Where $d$ is the Euclidean distance corresponding to the real value and predicted value of each word, $p_i$ is the real percentage of players solving the puzzle in different guesses, and $qi$ is the predicted percentage of players solving the puzzle in different guesses. Then we calculated the average Euclidean distance of words as $d = 0.9602$ , which indicates that our prediction are very close to reality and the prediction model has a high accuracy.

# 7 Model 4: Kmeans-Based Ensemble Learning Word Difficulty Classification

## 7.1 Kmeans Clustering Model

- **Model Establishment**

The advantage of the Kmeans algorithm is that the algorithm is simple and fast, and it's relatively efficient when dealing with large data sets. Among them, the algorithm flow chart is shown below:

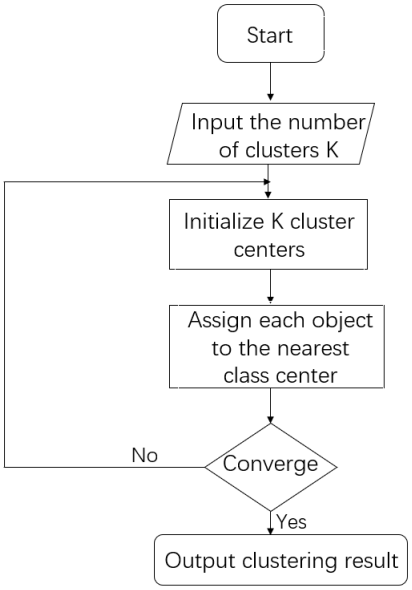


Figure 13: *The flow chart of Kmeans*

We use SPSS for systematic clustering and get the corresponding distortion coefficient $J$, through the elbow rule, estimates the optimal number of clusters $K$. The elbow rule is to roughly estimate the optimal number of clusters through graphics. The principle is to define the distortion coefficient of each class as the sum of the squares of the distance between the center of gravity of the class and its internal points. Then make the total distortion coefficient as the vertical axis and the number of clustered categories as the horizontal axis in the line chart. When the variation of the degree of distortion is significantly reduced, the corresponding number of categories K is the optimal number of clusters, through which reducing the influence of subjectivity on the classification results.

- **Model Solving**

We use systematic clustering to conduct systematic clustering analysis on 359 words based on the times of trying, and the line chart of clustering coefficient can be obtained as shown below. According to the elbow rule, when $K = 3$ or $K = 4$, the degree of distortion tends to change gently, here we make K=3, that is, the difficulty of words is divided into three levels.

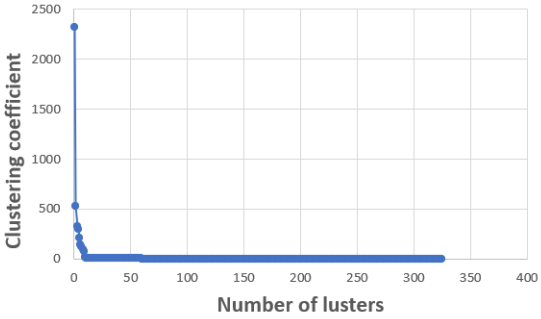


Figure 14: *Number of clusters*

Then we use PCA to reduce the dimension of the trying times, and classify the original 7 times of trying with the fourth times as the cut-off point. The sum of percentages less than 4 times is used as a simple level feature, the sum of percentages more than 4 times is used as a hard level. feature, and a percentage equal to 4 times as a middle level feature. The relationship between the degree of difficulty and the times of trying is shown in the table below. Words with a high percentage in the sum of trying times less than 4 are easy words, words with a relatively average percentage are medium difficulty words, and words with a high percentage in the sum of trying times more than 4 are difficult words.

Table 7: *Different types*

| | ”4” | ”=4” | ”4” |
|---|---|---|---|
| Easy | **42.57**% | 34% | 23.77% |
| Hard | 15.66% | 28% | **57.06**% |
| Middle | 26.81% | **35**% | 37.75% |

## 7.2 Attributes of the words in three classified levels

We selected typical words from each of the three levels to generate a wordle. Purple words in the wordle are difficult words, green words are medium difficulty words, and orange words are easy words. By analyzing the common features of the same level words, it can be seen that there are usually repeated letters in words with high difficulty, such as excel, mummy, foggy, etc. Among words with medium difficulty, many words are adjectives or verbs and start with with vowels, such as awful, ample, exist, etc. While simple words are mostly nouns and start with aspirated sounds such as ch, th, etc, and the pronunciation has basically changed, such as chief, chute, charm, etc.


Figure 15: *Word Cloud*

## 7.3 Ensemble Learning

In the previous article, we used the Kmeans to initially classify the difficulty of words in dataset, but this classification is more subjective. In order to further improve the classification efficiency, based on the Kmeans classification results, we use ensemble learning to highly fit the words and their classifications, thereby improving the classification accuracy of all words in Wordle.

Ensemble learning classifies words by building and combining multiple base classifiers, and its accuracy is usually significantly improved compared with its base classifiers. In this paper, XGBoost, random forest and AdaBoost are used as three base classifiers, and the classification result of Kmeans is used as the training label to further improve the classification model.

XGBoost makes classification decisions by concatenating multiple underlying decision trees, and the concatenation of classifiers also enables XGBoost to iteratively reduce the classification error, until the error of all the decision trees is reduced to an ideal situation.

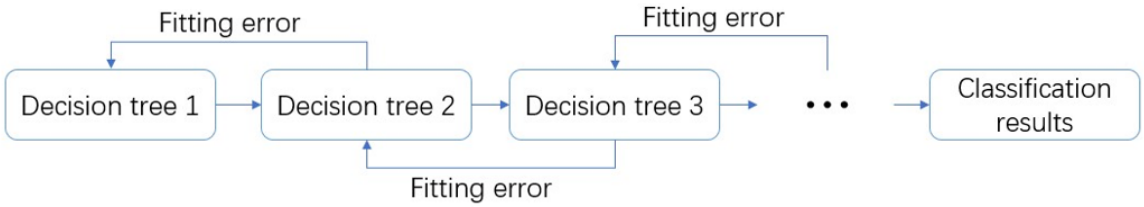


Figure 16: *XGBoost's flow chart*

The random forest is composed of multiple bottom-level decision trees that are not related to each other. For the target value to be classified, each decision tree completes the classification task for the target in parallel, gets multiple classification results, and then confirms the final result based on the majority voting principle.

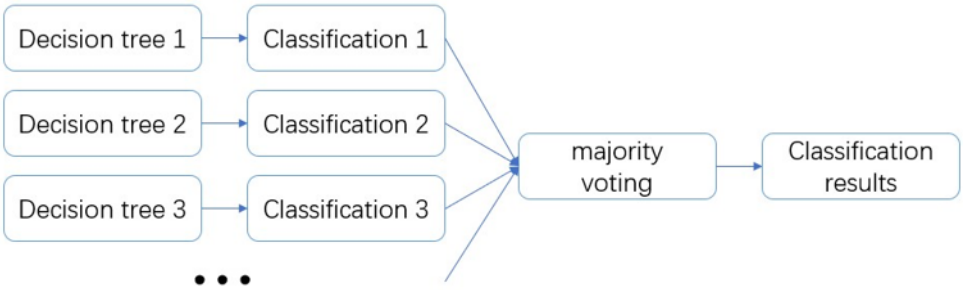


Figure 17: *Random Forest's flow chart*

For samples with the same weight at the beginning, AdaBoost focuses on increasing the weight of samples correctly classified by the base classifier in the previous round, and reducing the weight of misclassified samples, thereby increasing the probability of successful classification in the next iteration. The classification results are solved by a series of base classifiers in a divide-and-conquer way. The final AdaBoost classifier is a linear combination of the corresponding weights of each base classifier.

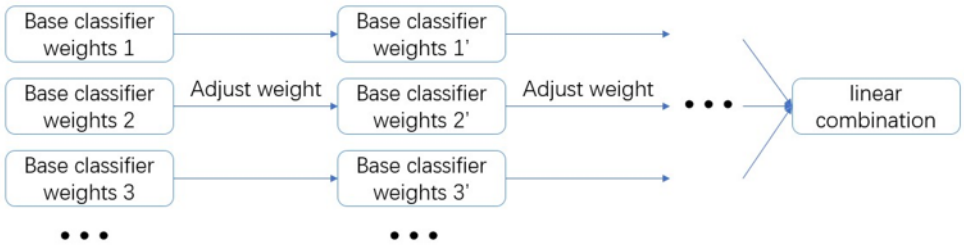


Figure 18: *AdaBoost's flow chart*

After getting the classification results of the three base classifiers, we use Plurality Voting to confirm the final classification results of the word. That is, if a classification result has the most votes, then the final classification result is that classification. In this way, we built an ensemble classifier model which is applicable for all words in Wordle.

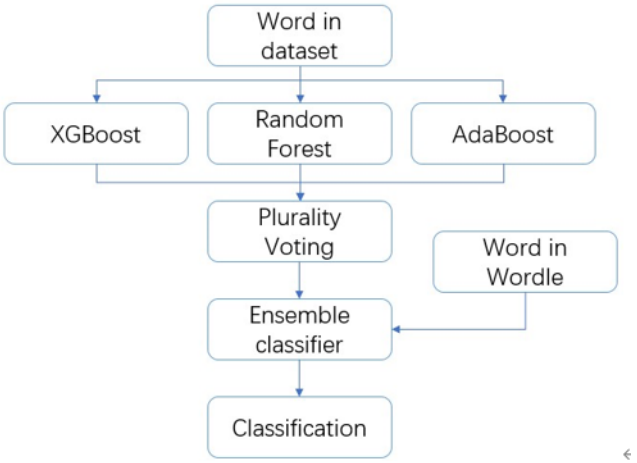


Figure 19: *Plurality Voting's flow chart*

## 7.4 EERIE Difficulty Classification

- **Classification result:** By training classifiers to get the difficulty of the word EERIE, we found that the classification results of the three base classifiers XGBoost, Random Forest, and AdaBoost are all hard. According to the principle of Plurality Voting, the category of hard type has the most votes, so the word should be classified as hard.

- **Classification probability:** In order to further confirm the classification details, we output the probability predictions of different classifications by ensemble learning as shown in the following table.

Table 8: *EERIE Difficulty prediction*

| easy | hard | middle |
|---|---|---|
| 7.53% | **64.64%** | 27.84% |

According to the ensemble learning, the word EERIE has a 7.53% probability of easy, a 64.64% probability of difficult, and a 27.84% probability of medium. Therefore, the final word EERIE is classified as hard.

And by analyzing the word EERIE:

- **Repeated letter**The word EERIE contains 3 Es, which belong to the characteristics that should be possessed by hard classification.
- **Vowel letter** The word contains 4 vowels, which also belong to the characteristics that the medium classification should contain.
- **Affix**The word does not have th, ch and other affixes at the beginning of the word, so it does not have the characteristics that simple classification should have

The difficult classification features are more obvious than the medium classification features, so the classification results of EERIE are in line with expectations.

## 7.5 Model Accuracy

Through the sensitivity analysis of the ensemble learning model, we found that the higher the depth of the base tree, the greater the probability of overfitting, and the classification accuracy gradually decreased from 1.0 to about 0.75. However, the difference in iteration step, base classifier sampling ratio and tree type sampling ratio maintains the classification accuracy at a stable level of around 0.85, so we consider that the overall accuracy of the ensemble learning model is at high level.

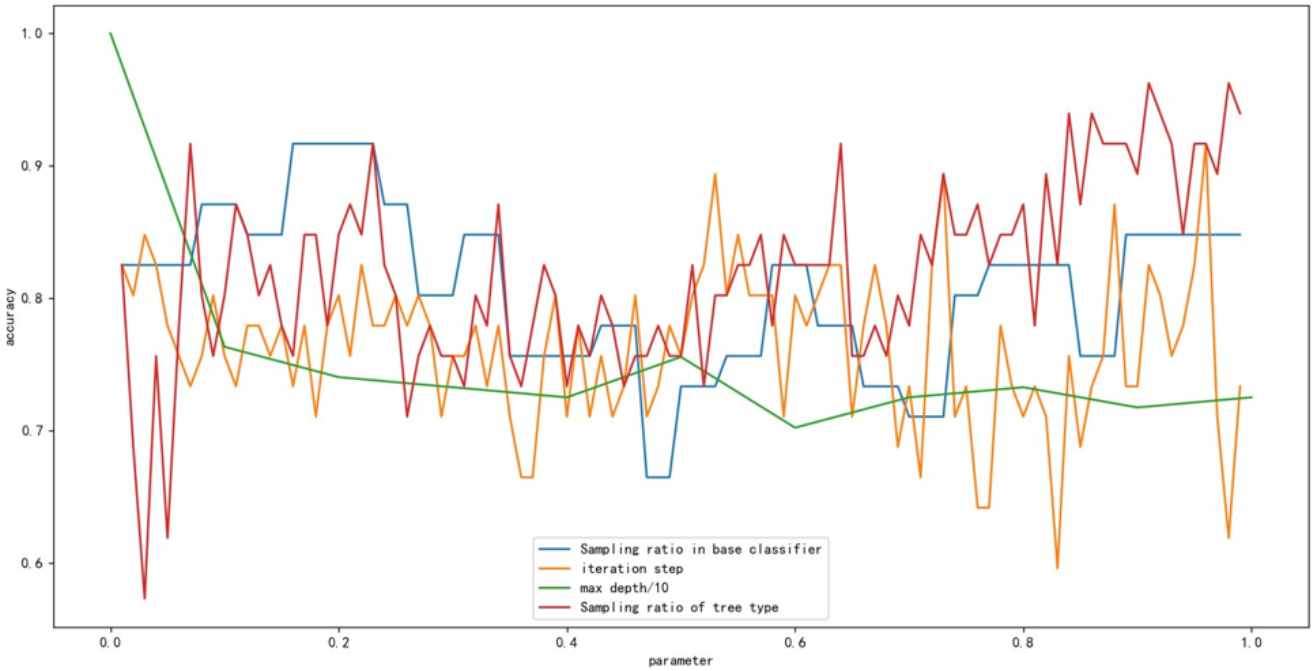


Figure 20: *sensitivity analysis*

# 8 Model 5: Descriptive Statistics

## 8.1 Changes in the number of players

It's easily to find out that the number of reported results can reflect how many people are playing this game. As shown in the figure, we can observe that at the beginning of the game's release, the number of players continued to grow. At this time, this game was quite popular among people. After March 2022, the number of players began to gradually decrease, in other words, players are losing their interest on this game. On the one hand, it is related to the decline in the popularity of this game, and on the other hand, it is also related to the daily occurrence of the global epidemic, the gradual restoration of social order and stability and people's high enthusiasm for outdoor sports after a long period of lockdown.

But at the same time, we can also find that the proportion of number in hard mode is constantly increasing, indicating that while the game is losing their players, experienced players are also accumulating, whose stickiness on this game is getting higher and higher, and most of them tend to choose hard mode.

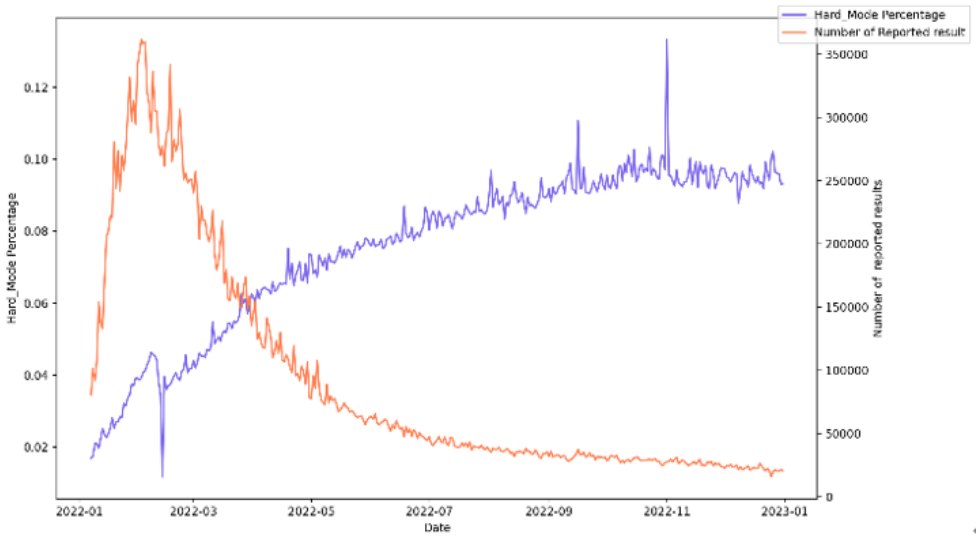


Figure 21: *Changes in reported result and hard mode*

## 8.2 Distribution of trying times

We average the percentage of players solving the puzzle in different guessing times, and it is approximately subject to a normal distribution, that is, for most players, they need to guess the correct word for four times, which is consistent with the actual situation. Because the probability of successfully guessing the correct word for the first time without any prompts is very rare to happen, players need to follow the game prompts and continue to guess and make mistakes, and most of this part requires 3 to 5 trying.

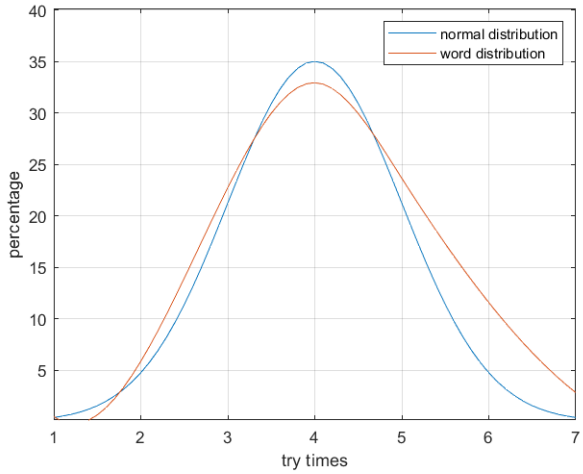


Figure 22: *Changes in reported result and hard mode*

# 9 Model Evaluation

## 9.1 Strengths

- **High sensitivity** The accuracy of the model obtained by adjusting the ensemble learning parameters is at a stable level of around 0.85, and the results of different classifications have obvious characteristics to explain.
- **High precision** The R of the neural network is high. And ensemble learning uses a variety of classifiers for joint classification. Compared with the simple classification method, the classification accuracy has been significantly improved.
- **Flexible solution** For words that cannot be directly used for model training, we quantify them into coded data through corpus preprocessing in NLP, which is convenient for subsequent analysis and processing. It is also flexible and changeable.

## 9.2 Weaknesses

- Word feature extraction is incomplete. This article only uses part of the features of words to solve the model. For the deeper feature mining, it needs to use deep learning in NLP to complete.
- The principle of the emsenble learning classifier is complex which is similar to the black box, so it is difficult to analyze its internal classification logic.