Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2320339

Pan for Gold: What Can We Learn from Wordle?

Abstract

Wordle is a popular daily puzzle provided by the New York Times, which has created a global "word-guessing craze". In this paper, we analyze the statistics from Twitter to build a prediction model about"the number of reported results", "the distribution of the reported results" and other indicators, and evaluate and classify the difficulty of different words.

First, we analyzed the change trend of the number of reported results according to the data of attachment 1, and considered using functions to fit the change trend respectively. Select the **quadratic function** and **second-order exponential function** with good fitting effect and consistent with the actual change regulation to fit the change trend of the rising and falling stages respectively.In addition we also modeled the error. According to the integrated model, the prediction interval for the number of reported results on March 1, 2023 is **(16053,21565)**.

Next, to explore whether there are factors affect the percentage of scores reported that were played in Hard Mode, we proposed four attributes of words that may affect the results, and test the relationship between the undetermined attributes and the percentage of scores through partial **correlation analysis**. We conclude that **the attributes of the words themselves do not affect the results** of the dependent variable, and put forward relevant analysis and reasons for this.

Then, we built a model to predict the distribution of the number of guesses of a given word. Using the proposed four word attributes and time as the input characteristics of the samples, and the distribution column as the output, we build an **XGBoost** learning model for multi-input and multi-output. Using the model with better learning effect in the training set to predict the future results, we obtained the distribution of the number of guesses of the word "EERIE" on March 1, 2023. Through the analysis and evaluation of the model, we gave the uncertainty of the model and the reliability of the prediction results.

Further, for the classification of word difficulty, we first use **Factor Analysis (FA)** to construct a scoring mechanism for words based on "word use frequency", "letter use frequency" and other attributes, and give two main factors that affect the score. For the given data set, we draw the sample scatter diagram under two main factor dimensions, use the improved **K-means++ algorithm** to classify the difficulty of words, and based on this,we divide the given word "EERIE" into "difficult" categories (the highest level). Through the analysis of the classification results, we put forward the conclusion that **"word use frequency"** and **"letter use frequency"** mainly affect the classification results.

Finally, we performed a sensitivity analysis and a strengths and weaknesses analysis of the model and wrote a letter to the Puzzle Editor of the New York Times based on the above results.

Keywords: **Wordle, Model Fitting, Partial Correlation Analysis, XGBoost, Factor Analysis, K-means++**

Contents

1 Introduction **3**
- 1.1 Background . . . 3
- 1.2 Restatement of the Problem . . . 3
- 1.3 Our Work . . . 3

2 Assumptions and Justification **4**

3 Notations **5**

4 Model I Curve Fitting Model **5**
- 4.1 Preperation of Model I . . . 5
- 4.2 Establishment of Model I . . . 7
  - 4.2.1 Descent Process . . . 7
  - 4.2.2 Upward Process . . . 8
- 4.3 Application of Model I . . . 9
- 4.4 Result of Model I . . . 11

5 Model II Partial Correlation Analysis Model **11**
- 5.1 Preparation of Model II . . . 11
  - 5.1.1 Analysis of Independent Variables . . . 11
  - 5.1.2 Model Selection . . . 12
- 5.2 Establishment of Model II . . . 12
- 5.3 Evaluation of Model II . . . 13

6 Model III XGBoost Predictive Model **14**
- 6.1 Preparation of Model III . . . 14
- 6.2 Establishment of Model III . . . 14
- 6.3 Application of Model III . . . 15
- 6.4 Evaluation of Model III . . . 16

7 Model IV Factor Analysis and Clustering Model **16**
- 7.1 Establishment of Model IV . . . 16
  - 7.1.1 Factor analysis model . . . 17
  - 7.1.2 Cluster analysis model : K-means++ algorithm[7] . . . 19
- 7.2 Application and Evaluation of Model IV . . . 19
  - 7.2.1 Solution of Factor Analysis Model . . . 19
  - 7.2.2 Solution of Clustering Model . . . 20

8 Other Interesting Features **21**

9 Sensitivity Analysis and Model Validation **21**
- 9.1 Stability of Changes . . . 21
- 9.2 Influence of Feature Selection . . . 22
- 9.3 Influence of Factor Quantity Selection . . . 23

10 Strength and Weakness **23**
- 10.1 Strength . . . 23
- 10.2 Weakness . . . 23

Reference **24**

1 Introduction

1.1 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.

Wordle's rules are simple: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.

Due to the continuous popularity of wordle on social media, it is meaningful to consider the impact of time on the number of users, the determinants of word difficulty and the impact of word difficulty on the distribution of answer times.

1.2 Restatement of the Problem

Many (but not all) users report their scores on Twitter.For this problem, MCM has generated a file of daily results for January 7, 2022 through December 31, 2022. Answer the following questions according to the data contained in the attachment:

1) Develop a model to explain the variation of the number of reported results and use your model to create a prediction interval for the number of reported results on March 1, 2023. Expore how any attributes of the word affect the percentage of scores reported that were played in Hard Mode?

2) For a given future solution word on a future date, develop a model that allows you to predict the distribution of the reported results. Give a specific example of your prediction for the word EERIE on March 1, 2023. Evaluate and analyze the results.

3) Develop a model to classify solution words by difficulty. Identify the attributes of a given word that are associated with each classification. Using your model to evaluate the difficulty of the word EERIE and discuss the accuracy of your classification model.

4) 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.

1.3 Our Work

The main work of this paper is shown in **Figure 1**.

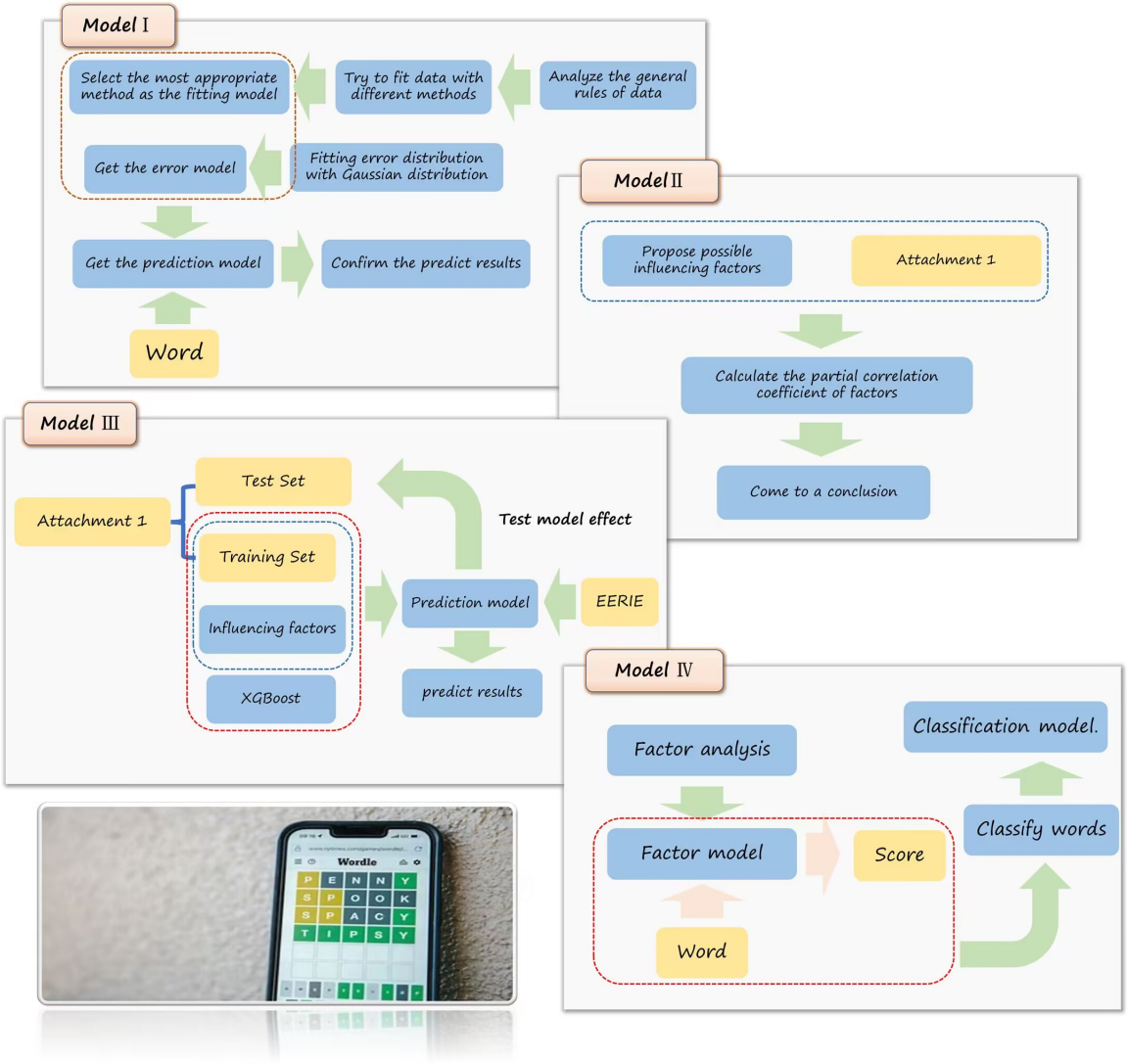


Figure 1: Overview of the article

2 Assumptions and Justification

- **Assumption 1: All players can only play the game once a day, and the result will either be a pass or a fail, and they cannot play the game again that day.**

  — **Justification:**If some players play multiple times a day, their previous experience will affect their accuracy in subsequent guesses, so that the data can not truly reflect the game level of the players.

- **Assumption 2:All players can only guess independently every day and cannot discuss with other players or search for answers.**

  — **Justification:**If some players answer the questions by cooperating with each other or checking the answers online, it will affect the detection of the players' real game level, and thus affect the subsequent model building and prediction.

- **Assumption 3:In order to make the model have good generalization ability and practical application ability, we assume that the data obtained can reflect the general situation of all players.**

  — **Justification:**In order to make the model have good generalization ability and

practical application, we assume that the data we get can reflect the general situation of all players.

- **Assumption 4:The actual number of reported results per day has some error compared to the theoretical variation regulation, and the error is random.**

  — **Justification:**To better model the error, we ignore the effect of some factors on the error and and approximately assume that the daily error is random.

- **Assumption 5:The daily word difficulty is random and has no obvious correlation with time change.**

  — **Justification:**In order to ensure the independence of the time factor and word attributes, and to better simulate the daily problem setting in reality, we make this assumption.

3 Notations

The parameters used in this article are listed in **Table 1**.

Table 1: Notation symbols used in this paper

Symbols	Description
c	The constest number
y	The number of reported results
ε	The error in the number of reported results
w	Frequency of word usage
l	Frequency of letter usage
r	Number of repeated letters
v	Number of vowel letters

4 Model I Curve Fitting Model

4.1 Preperation of Model I

Based on the data from Attachment 1, we first plotted the trend of the number of reported results from January 7, 2022 to December 31, 2022 on a line graph (the horizontal coordinate is the content number) as shown in **Figure 2**.

Based on preliminary observations, we found that the number of reported results showed an increasing trend over time, peaking on February 2, 2022 and then gradually decreasing.

In order to study the specific pattern of the number of reported results over time and to predict the future trend of the number of reported results, we consider using the existing data to construct a quantitative model and obtain the number of reported results over time by fitting the function. We consider using the existing data to construct a quantitative model to obtain the law of the number of reported results over time by function fitting. The flow of the model construction is shown in **Figure 3**[1].

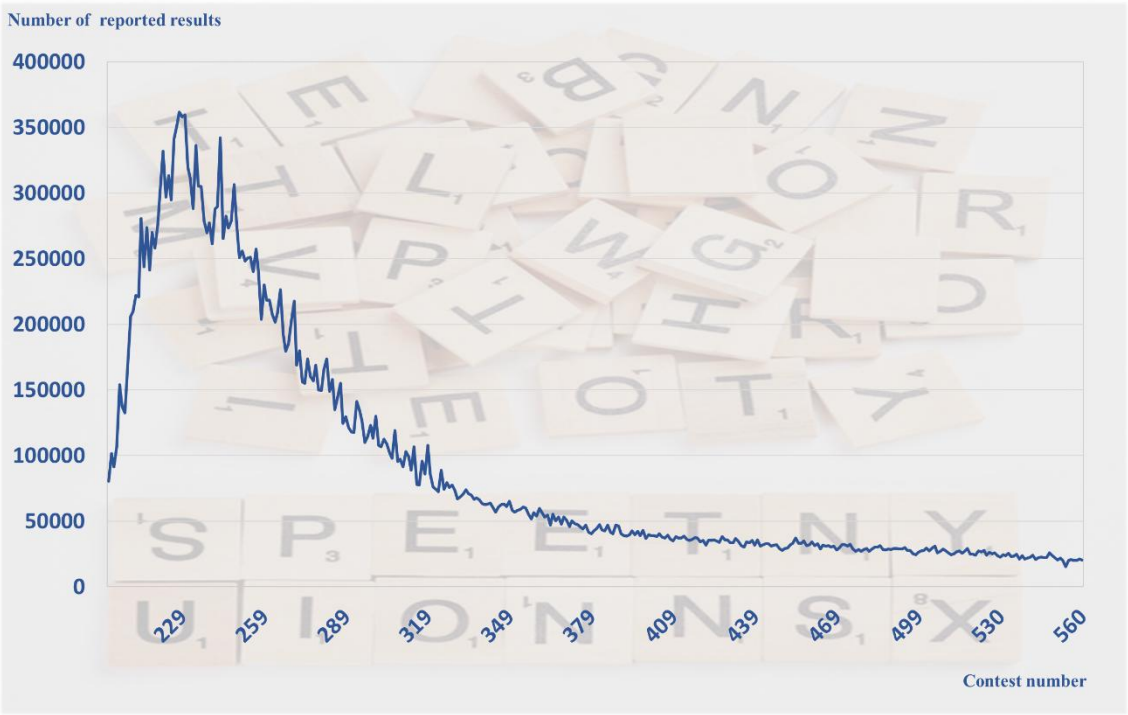


Figure 2: Change trend of the number of reported results

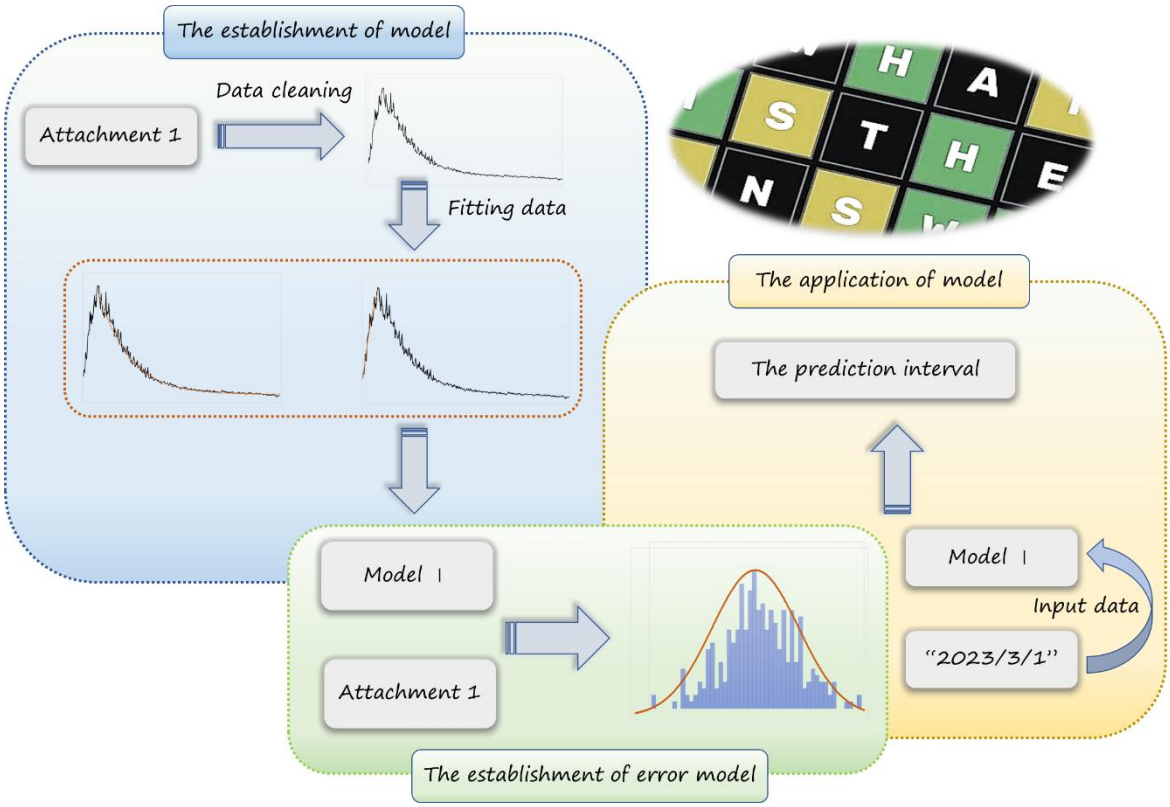


Figure 3: Construction flow chart of model I

4.2 Establishment of Model I

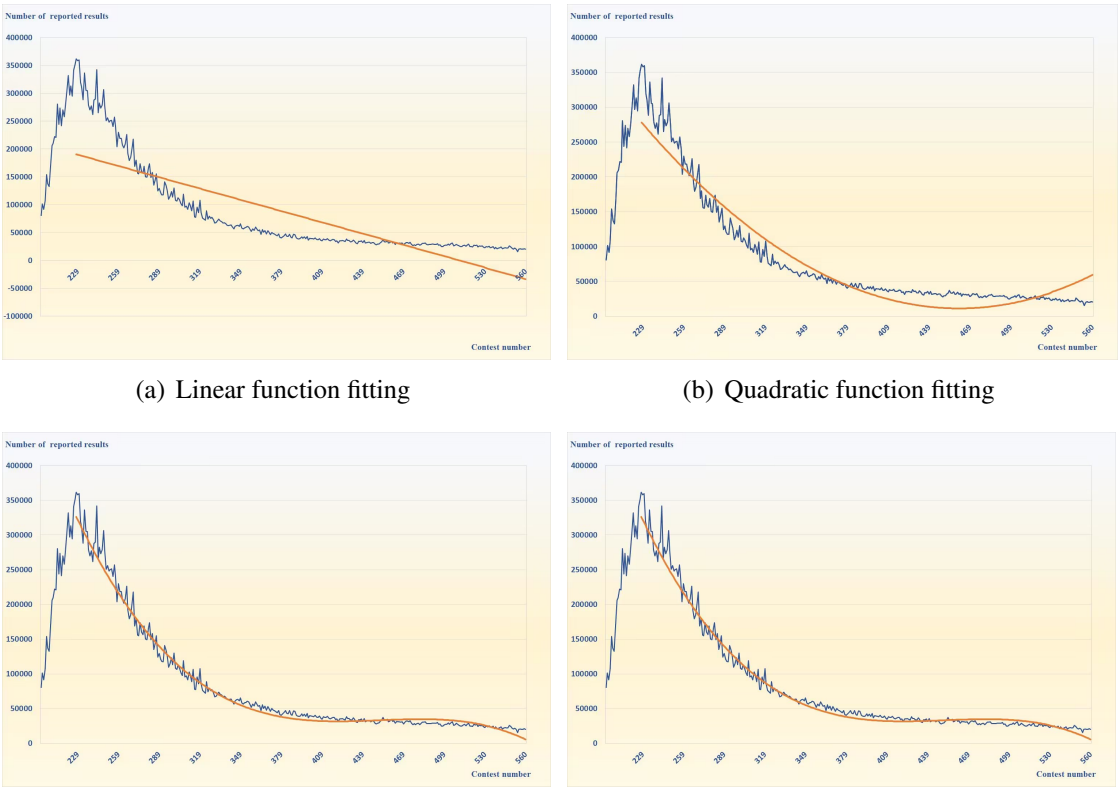
Our initial speculation is that at the early stage of the game's launch, due to the unique game mechanics and high playability, more and more new players flocked to the game and joined in the puzzle solving together under the role of official publicity and player sharing, so the game showed the trend of rapid growth of the number of reported results in the early stage.

After the game has been online for a period of time (content number≥228), the number of reported results no longer has room to increase because it has attracted enough potential players, and at the same time, as some players cannot get a sense of achievement from the game after repeated failed attempts and choose to retreat , another part of players may also lose their pursuit of success because of mastering the game skills and methods. After a period of steady increase, the number of reported results began to decline.

Through the above observation and analysis we know that the number of reported results since January 7, 2022 shows a trend of rising and then falling, so we take the inflection point (content number=228) as the dividing point for the corresponding function fitting. In principle, the fitted function should not differ too much from the distribution of sample points, and the function value should not be zero in the future. In combination with the above requirements, we respectively consider using polynomial function and exponential function to fit the data. In order to judge the fitting effect better, we use the correlation coefficient R^2 as the judging criterion, and the closer R^2 is to 1, the better the fitting effect of the model.

4.2.1 Descent Process

On the basis of the existing data, we use MATLAB to fit the data with 1∼4 degree polynomial function and 1∼2 order exponential function respectively, and the fitting effect is shown in the **Figure 4** and **Figure 5**.



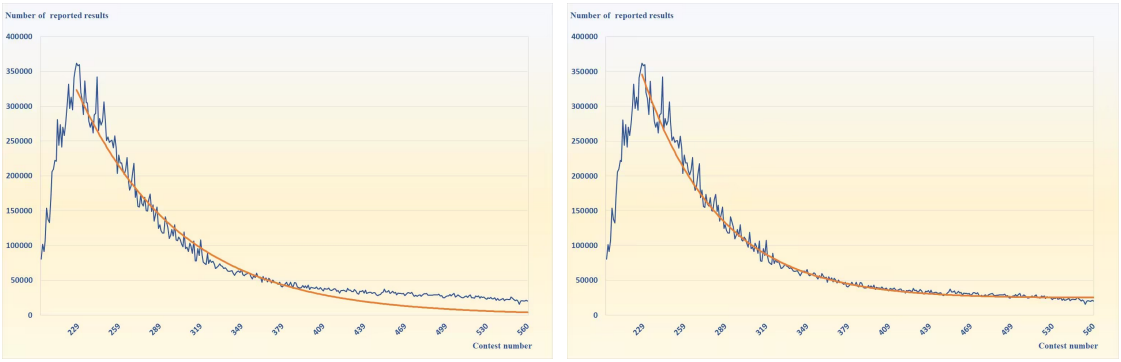
(a) Linear function fitting

(b) Quadratic function fitting

(c) Cubic function fitting

(d) Quartic function fitting

Figure 4: Polynomial function fitting diagram



(a) First-order exponential function fitting (b) Second-order exponential function fitting

Figure 5: Exponential function fitting

The correlation coefficients calculated by MATLAB for the above fitting methods are shown in the following **Table 2**.

Table 2: Comparison of fitting effects

Fitting method	Functional form	R^2
Linear function	$y = ax + b$	0.6654
Quadratic function	$y = a_1x^2 + a_2x + a_3$	0.9252
Cubic function	$y = a_1x^3 + a_2x^2 + a_3x + a_4$	0.9846
Quartic function	$y = a_1x^4 + a_2x^3 + a_3x^2 + a_4x + a_5$	0.9884
First-order exponential function	$y = a\exp(bx)$	0,9625
Second-order exponential function	$y = a\exp(bx) + c\exp(dx)$	0.9870

According to the comparison of the results in the table, the fitting effect of cubic, quartic and second-order exponential functions is good, but considering that the cubic and quartic polynomial functions obtained by fitting cannot meet the requirements of small change and greater than 0, we choose the second-order exponential function as the trend function of the number of reported results changing with the number of questions, the specific expression is

$$y = 5.418 \times 10^5 \exp(-0.01749x) + 1.946 \times 10^4 \exp(-0.0006212x) \tag{1}$$

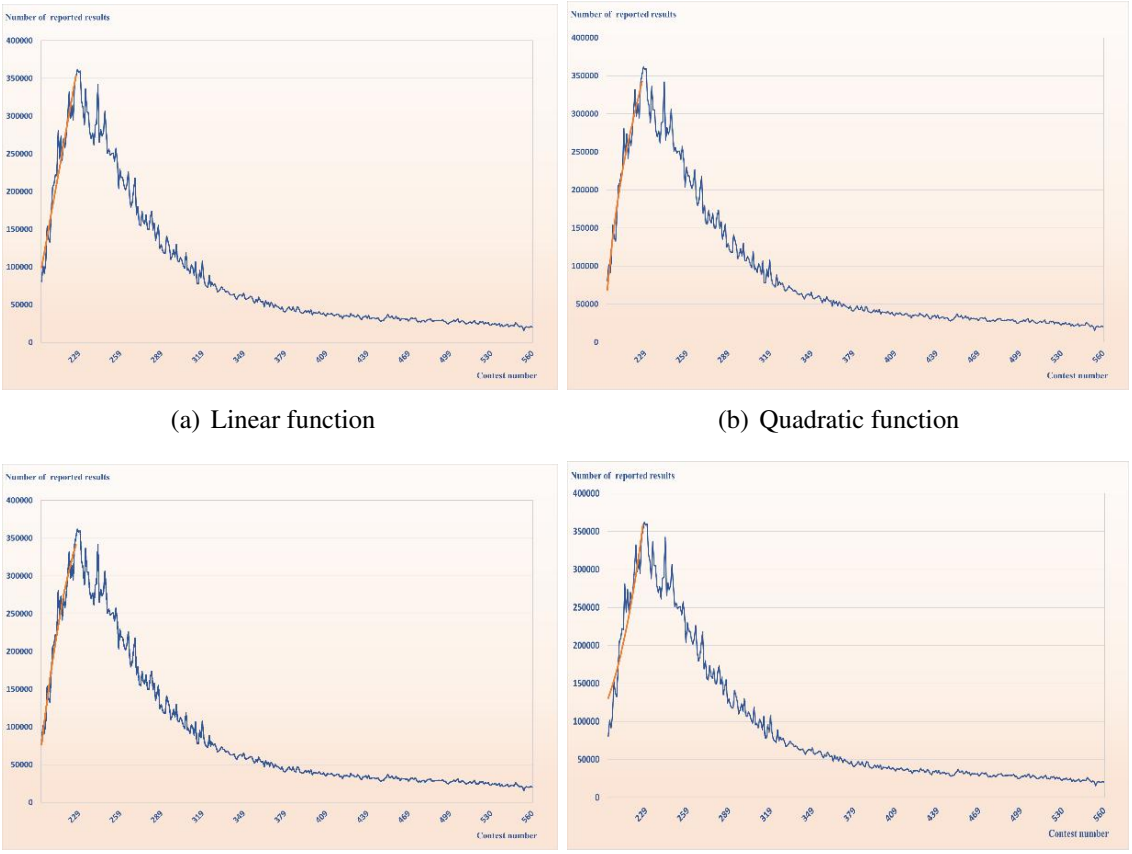
4.2.2 Upward Process

Like the descending process, since the ascending process has a relatively short time, we consider fitting the data using the 1∼3 degree polynomial function and the 1 order exponential function respectively, and the fitting effect is shown in the **Figure 6**.

The performance of the fitted function is shown in the following **Table 3**.

Table 3: Comparison of fitting effects

Fitting method	Functional form	R^2
Linear function	$y = ax + b$	0.9381
Quadratic function	$y = a_1x^2 + a_2x + a_3$	0.9555
Cubic function	$y = a_1x^3 + a_2x^2 + a_3x + a_4$	0.9575
First-order exponential function	$y = a\exp(bx)$	0,8828



(a) Linear function (b) Quadratic function

(c) Cubic function (d) First-order exponential function

Figure 6: Function fitting

According to the conclusion in the table above, the correlation coefficients of quadratic function and cubic function fitting have reached more than 95%, so both can be used as good fitting methods. In order to reduce the subsequent calculation difficulty, we choose the quadratic function as the change function of the number of reported results in the rising stage, whose expression is

$$f(x) = -192.2x^2 + 1.576 \times 10^4 x + 6.122 \times 10^4 \tag{2}$$

Based on the analysis and solution of the above two processes, we have built a preliminary model based on more than 300 existing data to fit the change trend of the number of reported results in the past. The detailed results are as follows.

$$y = \begin{cases} -192.2c^2 + 1.576 \times 10^4 c + 6.122 \times 10^4, & x < 228, \\ 5.418 \times 10^5 \exp(-0.01749c) + 1.946 \times 10^4 \exp(-0.0006212c), & x \geq 228. \end{cases} \tag{3}$$

Of course, we hope that the above model can not only fit the past data well, but also make a reasonable prediction for the possible future situation. Therefore, we will try to use the constructed model to predict the number of reported results on a certain day in the future, and give a reasonable and reliable prediction interval.

4.3 Application of Model I

Reviewing the construction process of the above model, we have fitted the change function of the number of reported results through the existing data. But generally speaking, there must

be some deviation between the actual changes and the theoretical values. Therefore, in practical applications, we should not only get a specific real value result based on the model, but also consider the impact of the error. Therefore, we will model the error based on Model I to get the improved model, and create a prediction interval for the number of reported results on March 1, 2023.

According to the assumption that the errors affecting the number of reported results are random, we have good reason to assume from the knowledge of error theory[2] that these errors (ε_i) obey a Gaussian distribution (where μ, σ are unknown), so we consider using the available data to determine the exact distribution of the errors. The histogram of the relative error distribution is shown in **Figure 7**.

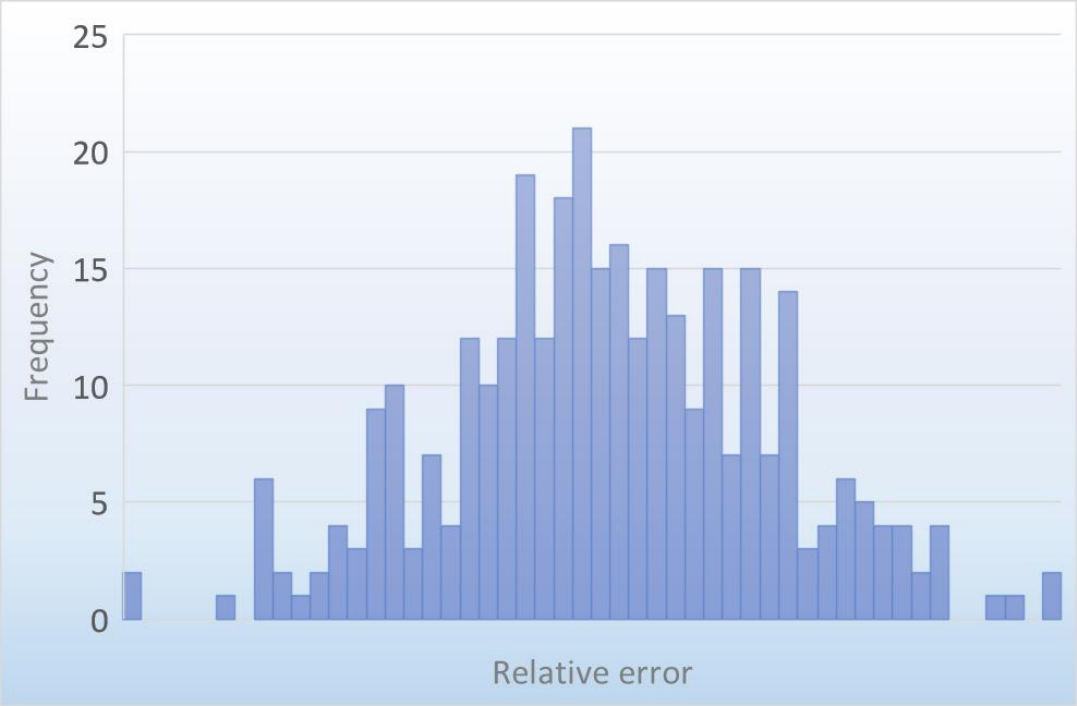


Figure 7: Error distribution histogram

The probability function of relative error obtained by fitting the above data with MATLAB is

$$f(\varepsilon) = \frac{1}{\sqrt{2\pi} \times 0.0755} \exp(-\frac{(\varepsilon - 0.01)^2}{2 \times 0.0755^2}) \tag{4}$$

Namely ε has Gaussian distribution $N(0.01, 0.0755^2)$, and its specific distribution curve is shown in the **Figure 8** below.

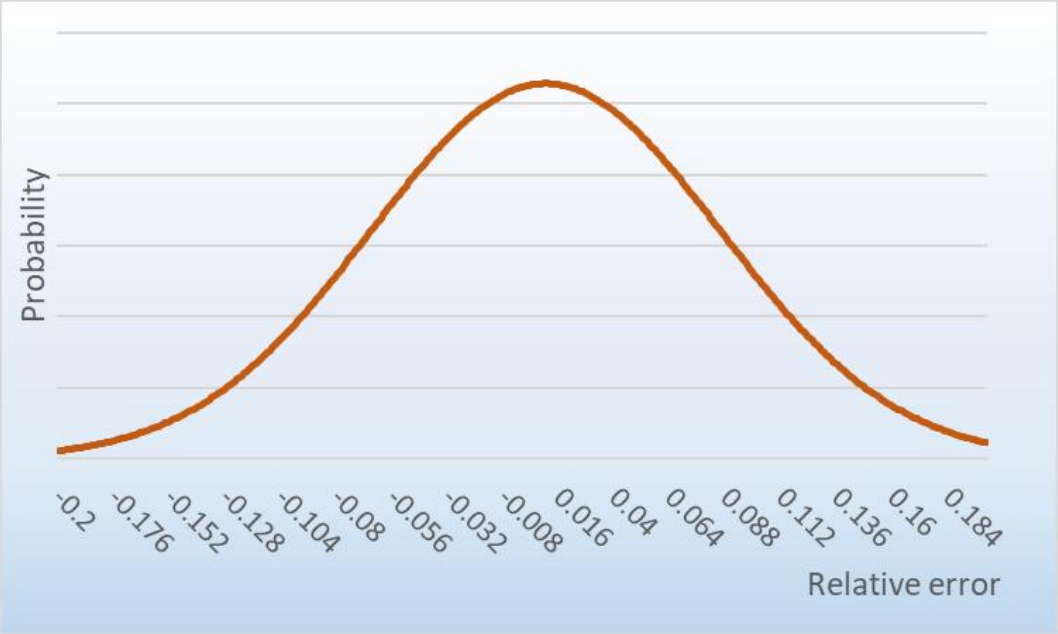


Figure 8: Error fitting

4.4 Result of Model I

Based on the above analysis, we have obtained the accurate distribution of the error. Using the idea of hypothesis test, we can obtain the confidence interval of the error under the condition of given confidence level. Here we choose the confidence level $\alpha = 95\%$, then according to the pivot method, we can know

$$u = \frac{\varepsilon - \mu}{\sigma} \sim N(0, 1) \tag{5}$$

If $P(|u| \geq c) \leq 1 - \alpha$,then $c = u_{0.975}$, where $u_{0.975}$ is the 0.975 quantile of the standard normal distribution, so the 95% confidence interval of error ε is$(\mu - u_{0.975}\sigma, \mu + u_{0.975}\sigma) = (-0.1380, 0.1580)$.According to the conclusion of model I, the predicted value of the number of reported results on March 1, 2023 is 18623, so after considering the error, we give the predicted result of the number of reported results on March 1, 2023 as follows.

Table 4: Model prediction results

Model I Predicted Value	Y=18623
Relative error range	$(\varepsilon_1, \varepsilon_2) = (-13.80\%, 15.80\%)$
Prediction interval expressions	$(Y + Y * \varepsilon_1, Y + Y * \varepsilon_2)$
Forecast interval	$(16053, 21565)$

5 Model II Partial Correlation Analysis Model

5.1 Preparation of Model II

5.1.1 Analysis of Independent Variables

To explore the effect of word attributs on the percentage of scores reported that were played in Hard Mode, we first analyzed the attributes of the words themselves. The frequency of word

use affects the difficulty in guessing words, with words that are used less frequently being more difficult to guess. The composition of the letters in a word also has an impact, as words made up of more frequently used letters are used more frequently, and words containing more vowels are used more frequently, so people tend to guess those words that are used more frequently. In addition, the structure of the word itself also has an effect, with words containing two or more repeated letters being more difficult to guess. It is easy to overlook that in addition to the attributes of the words themselves that may have an effect on the percentage of scores reported that were played in Hard Mode, time may also have an effect on this, and we should fully consider all possible independent variables and analyze each one individually.

We have preliminarily summarized five factors that determine the difficulty of words: frequency of word use, frequency of letter use, number of vowels, number of repeated letters and time. Before analysis and calculation, we first deal with these five independent variables. We rank the use frequency of all words, and divide them into five grades from 1 to 5 according to the frequency from small to large, to express the use frequency of words. For the use frequency of letters, we add the use frequency of five letters to represent this factor. The number of vowels and repeated letters in each word represents the number of vowels and repeated letters. Time is considered from the first day of statistics.

Then analyze these five factors separately to determine whether they affect the percentage of scores reported that were played in Hard Mode, and analyze the degree of influence.

5.1.2 Model Selection

When considering correlation analysis, a more classical method is to use Pearson correlation coefficient to make a judgment, and the strength of correlation is obtained according to the correlation coefficient r. However, the Pearson correlation coefficient needs to control the other independent variables to be consistent, and in the treatment of this problem, because each word is specific, it is impossible to control the other independent variables unchanged while considering the influence of one independent variable, and it is not possible to exclude the interference of other independent variables. On this basis, we consider the use of partial correlation analysis for the solution.

Partial correlation analysis is used to solve for the correlation between one of the independent variables and the dependent variable when there are multiple independent variables and correlations between them[3]. The partial correlation analysis can exclude the influence of other independent variables to obtain the partial correlation coefficient between the required independent variables and the dependent variable, and then determine whether the independent variables have an influence and the magnitude of the influence by the positive or negative and magnitude of the partial correlation coefficient.

5.2 Establishment of Model II

According to the Pearson correlation coefficient, when there are only two variables x_1,x_2, the correlation coefficient between them is

$$r_{x_1x_2} = \frac{Cov(x_1, x_2)}{S_{x_1}S_{x_2}} \tag{6}$$

where $Cov(x_1, x_2)$ is the covariance of x_1 and x_2, and s_{x_1}, s_{x_2} are the standard deviations of x_1,x_2, respectively.

When three variables x_1,x_2,y exist, the correlation coefficient matrix between them is listed

as

$$\begin{pmatrix} 1 & r_{x_1x_2} & r_{x_1y} \\ r_{x_2x_1} & 1 & r_{x_2y} \\ r_{x_3x_1} & r_{x_3x_2} & 1 \end{pmatrix}$$

From the correlation coefficient matrix of the above equation, the partial correlation coefficients of x_1 and y (x_2 in the same way as y) can be obtained as

$$r_{x_1y\cdot x_2} = \frac{r_{x_1y} - r_{x_1x_2}r_{yx_2}}{\sqrt{(1 - {r_{x_1x_2}}^2)(1 - {r_{yx_2}}^2)}} \tag{7}$$

Similarly, pushing the problem to the higher dimention case, for x_1,x_2,... ,x_n,y, the formula for the partial correlation coefficient between x_i and y becomes

$$r_{x_iy\cdot l_1l_2\cdots l_n-1} = \frac{r_{x_iy\cdot l_1l_2\cdots l_{n-2}} - r_{x_il_{n-1}\cdot l_1l_2\cdots l_{n-2}}r_{l_{n-1}y\cdot l_1l_2\cdots l_{n-2}}}{\sqrt{(1 - {r_{x_il_{n-1}\cdot l_1l_2\cdots l_{n-2}}}^2)(1 - {r_{l_{n-1}y\cdot l_1l_2\cdots l_{n-2}}}^2)}} \tag{8}$$

where l_1,l_2,... , l_n are the excluded irrelevant variables This problem has four independent variables and taking n=5, the partial correlation coefficient between x_i (i=1,2,3,4,5) and y is

$$r_{x_iy\cdot l_1l_2l_3l_4} = \frac{r_{x_iy\cdot l_1l_2l_3} - r_{x_il_4\cdot l_1l_2l_3}r_{l_4y\cdot l_1l_2l_3}}{\sqrt{(1 - {r_{x_1l_4\cdot l_1l_2l_3}}^2)(1 - {r_{l_4y\cdot l_1l_2l_3}}^2)}} \tag{9}$$

5.3 Evaluation of Model II

Substituting the data according to the above equation using MATLAB, the partial correlation coefficients of the five independent variables with the dependent variable were obtained as shown in **Table 5**.

Table 5: Partial correlation coefficient

Attributes	w	l	r	v	t
r	-0.1959	-0.0105	0.1058	0.0240	0.9243

From the table, it can be seen that time has the greatest influence on the percentage of people with difficulty mode, with a partial correlation coefficient of 0.92, indicating that time shows a significant positive correlation with the dependent variable. As for the four attributes of the word itself, the frequency of the word with the largest partial correlation coefficient is just less than 0.11, and the partial correlation coefficients of the other three factors are less than 0.1.

Since the partial correlation coefficient between time and the dependent variable is large and the partial correlation coefficient between the attributes of the words themselves and the dependent variable is very small and significantly different, it is reasonable to assume that the percentage of the percentage of scores reported that were played in Hard Mode is determined by time only and is not related to the attributes of the words themselves.

In reality, this result is also logical. With the passage of time, the level of users' word guessing gradually improves. The classic mode may become less challenging and interesting for users. At this time, more and more users will choose to try the Hard Mode, which is consistent with our results. On the other hand, the user does not know the difficulty of the word when selecting the Hard Mode. The event of participating in the Hard Mode will not be affected by the difficulty of the word. That is to say, the attribute of the word itself has no effect on the percentage of scores reported that were played in Hard Mode.

6 Model III XGBoost Predictive Model

6.1 Preparation of Model III

Like Model II, we hope to explore the extent to which the summarized factors will affect the distribution of the reported results. It is worth noting that considering that players can improve their game level by summarizing their skills and consulting online strategies after the game has been online for a period of time, we intuitively know that the average times of players guessing words will decrease over time, which will also directly affect the distribution of the reported results. So we also take the contest number into consideration as the possible influencing factor of the distribution of the reported results.

6.2 Establishment of Model III

In order to ensure that the fitted model has good generalization ability and accurate prediction ability, we build a multi-output prediction model based on the XGBoost model for the distribution of guessing times.

XGBoost model is a new deep learning model proposed by Chen et al. in 2016. It is a large-scale parallel boostedtree model.[4] It uses multiple CART trees for prediction, and adds the predicted values of each tree as the final predicted value. The algorithm flow is shown in the **Figure 9**.

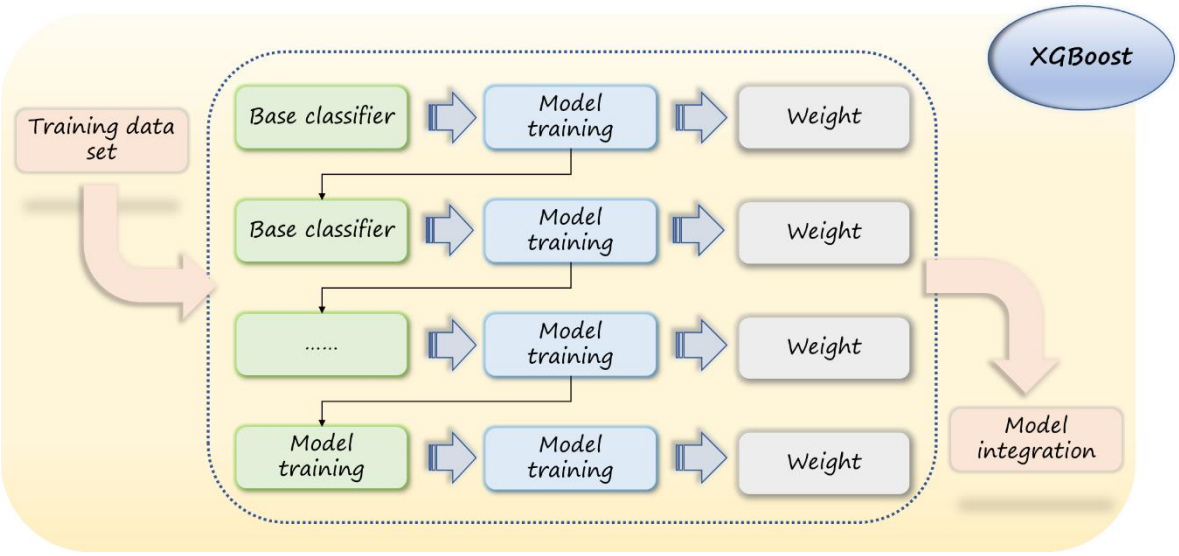


Figure 9: Flowchart of XGBoost algorithm

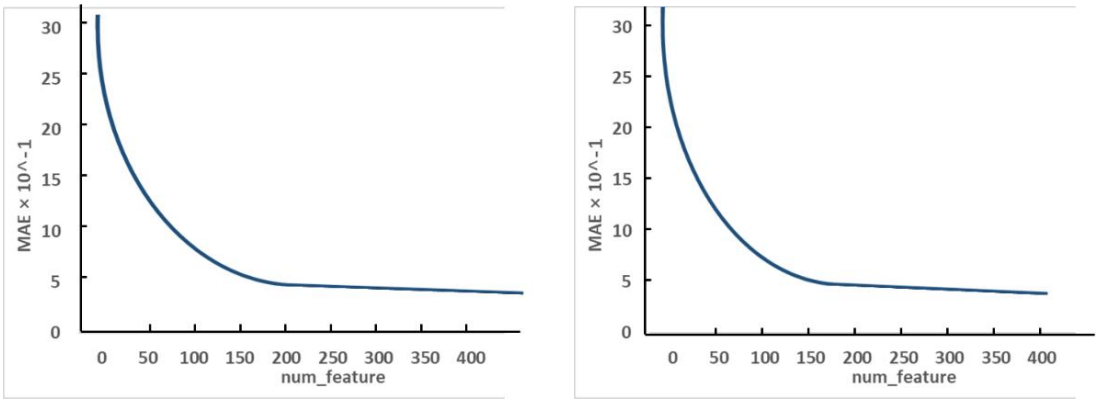
There are three types of parameters in the XGBoost model: General Parameters, Booster Parameters and Task Parameters. General Parameters determine which ascending model is selected in the ascending process, usually linear and tree models; Booster parameters depends on the type of ascending model selected; Learning objectives parameters determine learning strategies, define learning tasks and corresponding learning objectives. Use the data of Attachment1 to randomly select 20% data to form a test set and 80% data to form a training set. Taking the above "word usage frequency, letter usage frequency, number of vowel letters, number of repeated letters and the contest number" as input characteristics, the model parameters are set as shown in the following **Table 6**.

For the selection of the parameter num_feature, we determined the following experiment: by continuously adjusting the num_feature we recorded the mean absolute error (MAE) on the training set and plotted the trend of MAE with num_feature, the results of which are shown below (taking the distribution fit with the number of guess pairs of 3 and 4 as an example).

Table 6: Model parameter setting

	General Parameters	Booster Parameters			Task Parameters
Parameters	silent	eta	max_depth	$n\_estimatores$	$eval\_metric$
Value	0	0.05	4	undertermined	mae

Note: The parameter values not shown in the table are set as default values



(a) MAE change curve (the number of guesses is 3) (b) MAE change curve (the number of guesses is 4)

Figure 10: MAE change curve

According to the above figure we find that when num_feature≥200, the MAE of the model has dropped to a low level and remains basically unchanged afterwards, so we choose the parameter value of 200 for num_feature.

Based on the above parameter settings, we use Python to realize the construction of XGBoost model, and train the dynamic multi-output prediction model with the number of guess times as the output.

6.3 Application of Model III

Through the above process, we successfully built the XGBoost-based learning model, and its accuracy was 68.6205% on the test set and 63.1282% on the test set. We also input the word "EERIE" on March 1, 2023 into the model and got the following prediction results.

Table 7: The prediction of frequency distribution

	Percent in (%)						
Word	1	2	3	4	5	6	X
EERIE	0.0001	4.2321	23.2835	32.0824	24.9822	13.3721	2.0474

Combined with the prediction of model I, we can infer that the histogram of the number of times to guess the right words in the game on March 1, 2023 is shown in the **Figure 11** below.

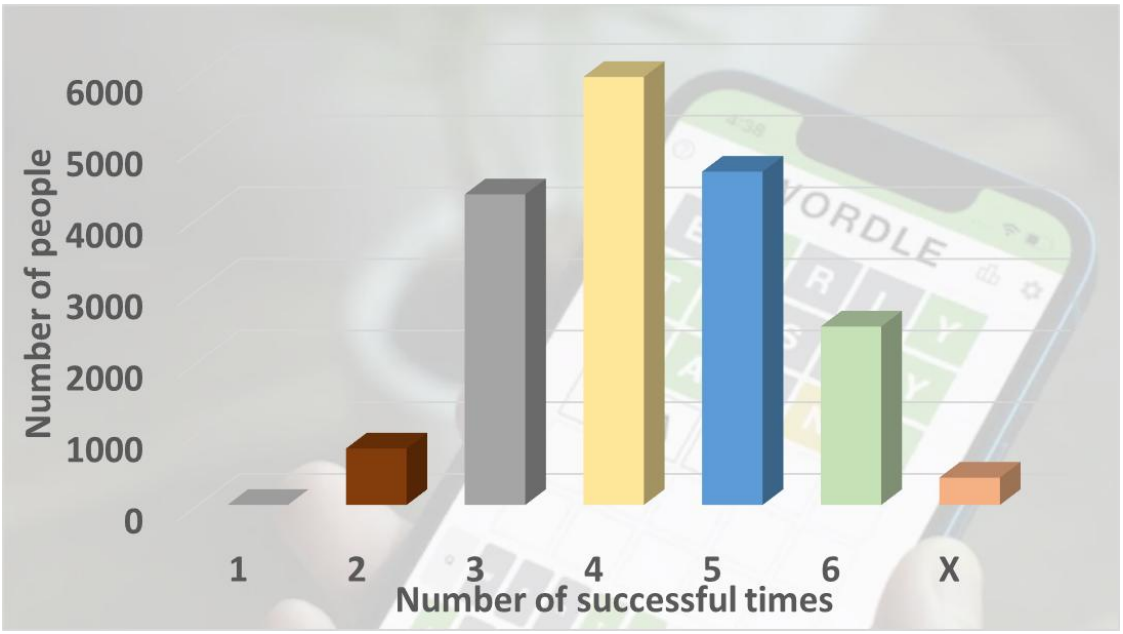


Figure 11: Predicted frequency distribution histogram

According to the above figure, we found that the prediction results were basically in line with our expectations, and the frequency distribution was in the form of "high in the middle and low at both ends". Since the prediction model has more than 63% accuracy in the test set, we have great confidence in the accuracy of the prediction result.

6.4 Evaluation of Model III

For model III, although it has good performance on the training and test sets, we still have the following uncertainties that may affect the accuracy of the model when evaluating it.

1.Feature selection. We only considered five factors as model features in the construction process, but there may be other influencing factors that are not considered or there are features that have no impact on the results among the factors considered, which bring uncertainty to the model and prediction.

2.Since the dataset used to build the model is derived from website information, there is no guarantee that the statistics of the data are completely accurate and reliable, and the model trained from this dataset will therefore have uncertainties.

3.Since the XGBoost learning model has many parameters to be selected, the setting of parameters may introduce uncertainty to the model and prediction.

4.In real life there are some uncertainties that affect the prediction uncontrollably, such as the mutual communication between players, the release of online tips, and the company's reform of the game mechanics.

7 Model IV Factor Analysis and Clustering Model

7.1 Establishment of Model IV

The model IV problem solving process is shown in **Figure 12**.

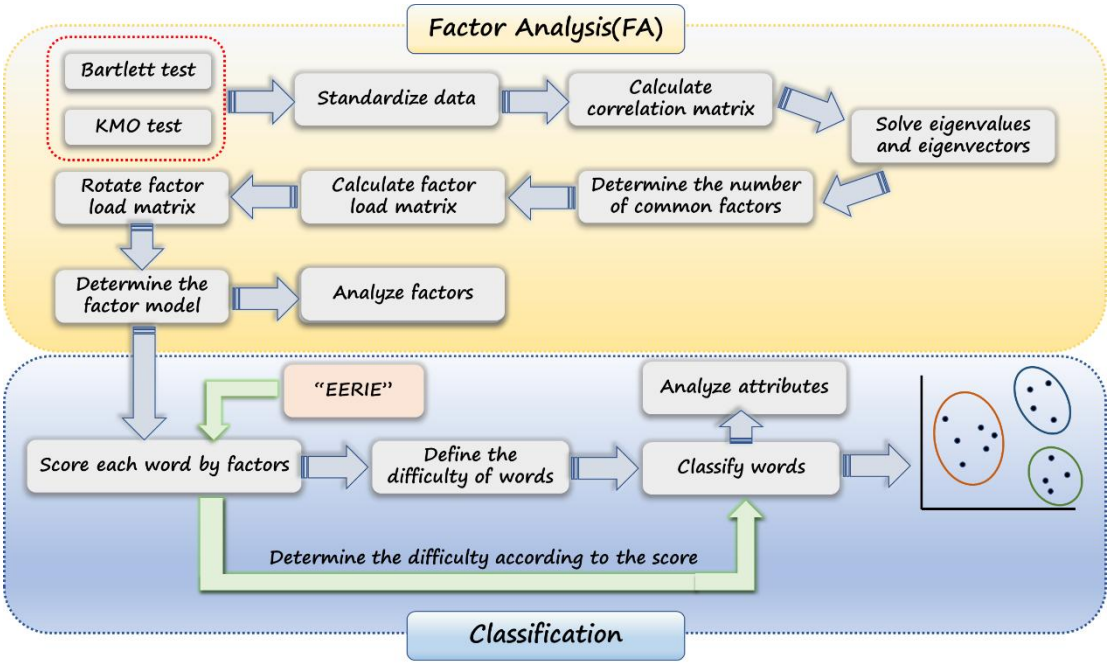


Figure 12: Flowchart of model IV

7.1.1 Factor analysis model

1.Model preparation (KMO inspection, Bartlett spherical inspection)

Before factor analysis, we need to confirm that the independent variables used have a good correlation, otherwise the independent variables do not meet the conditions for using factor analysis. In this problem, we use KMO test and Bartlett spherical test to determine whether the variable satisfies this condition.[5]

When the square sum of the simple correlation coefficients between all variables is far greater than the square sum of the partial correlation coefficients, the closer the KMO value is to 1, the stronger the correlation between variables. When the KMO value is greater than 0.5, the independent variable meets the conditions for using factor analysis. Bartlett's spherical test judged that if the correlation matrix is a unit matrix, then the variables are independent, and the factor analysis method is invalid. The smaller the value of Sig, the better. When it is less than 0.05, it means that the variable meets the conditions of factor analysis.

2.Model Building

Common factors in factor analysis are common influencing factors that cannot be directly observed but exist objectively. Each variable can be expressed as the sum of linear functions of common factors and special factors, namely:

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \varepsilon_i, (i = 1, 2, ..., n) \tag{10}$$

Where F_1, F_2,..., F_m are called common factors, $_i$ is called the special factor of x_i. The model can be expressed as:

$$X = AF + \varepsilon \tag{11}$$

The matrix A in the model is called the factor load matrix, and a_ij is called the factor "load", which is the load of the ith variable on the jth factor.

(1) Solution of factor load matrix & Principal component analysis.

First of all, we require that the cumulative contribution rate of factors should not be less than 90%, so we can determine m factors.

Calculate covariance matrix $\sum$. The characteristic root of, expressed in numerical value $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. The corresponding unit eigenvector is expressed as $e_1, e_2, \ldots, e_p$. The eigenvector matrix is expressed as U. At this time, the covariance matrix $\sum$ As follows express:

$$\sum = \sum_{i=1}^{m} \lambda_i e_i e_i^T + \sum_{j=m+1}^{n} \lambda_j e_j e_j^T \tag{12}$$

Based on formula above and the model hypothesis, we can also get the covariance matrix $\sum$. The expression method of formula is as follows:

$$\sum = AA^T + D_\sigma \tag{13}$$

$$D_\sigma = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2) \tag{14}$$

Combine formulas and we can obtain the estimation of factor load matrix:

$$A = \left(\sqrt{\lambda_1}e_1, \sqrt{\lambda_2}e_2, \cdots, \sqrt{\lambda_m}e_m\right) \begin{pmatrix} \sqrt{\lambda_1}e_1^T \\ \sqrt{\lambda_1}e_1^T \\ \cdots \\ \sqrt{\lambda_m}e_m^T \end{pmatrix} \tag{15}$$

where

$$a_{ij} = \sqrt{\lambda_i}e_{ij}(i, j = 1, 2, ..., n) \tag{16}$$

where λ_i represents the ith eigenvalue, e_{ij} means the jth component of the ith eigenvector of λ_i.

After obtaining the load matrix, we can express the factor model as:

$$\begin{cases} X_1 = \sqrt{\lambda_1}e_{11}F_1 + \sqrt{\lambda_2}e_{12}F_2 + \cdots + \sqrt{\lambda_m}e_{1m}F_m + \varepsilon_1, \\ X_2 = \sqrt{\lambda_1}e_{21}F_1 + \sqrt{\lambda_2}e_{22}F_2 + \cdots + \sqrt{\lambda_m}e_{2m}F_m + \varepsilon_2, \\ \cdots \\ X_n = \sqrt{\lambda_1}e_{n1}F_1 + \sqrt{\lambda_2}e_{n2}F_2 + \cdots + \sqrt{\lambda_m}e_{nm}F_m + \varepsilon_n, \end{cases} \quad m \leq n \tag{17}$$

(2) Rotation of the factor loading matrix

Let Q be an orthogonal matrix of order m, and B=AQ, then:

$$BB^T = AA^T \tag{18}$$

The factor load matrix A obtained above is not unique. In fact, the new matrix obtained by orthogonal transformation of matrix A can be regarded as the factor load matrix.

When we get the estimation of a factor load matrix, it may occur that multiple variables

have larger factor loads on the same factor, or a variable has larger loads on multiple factors, which is difficult to explain or name the factor. At this time, we can obtain a new simplified factor load matrix by rotating the factor load matrix to avoid the above situation, and finally get a better factor load matrix, This makes the factor load more distinguishable and convenient for factor analysis and naming.[6]

(3) Get factor score

The score of each factor can be obtained by regression method, and the final score function of the model can be obtained by this mean. The influence of independent variables on the difficulty of words can be analyzed by the composition of each factor, and the difficulty of each word can be quantitatively described by the score function.

7.1.2 Cluster analysis model : K-means++ algorithm[7]

After obtaining the factor scoring model, we can conduct a more scientific quantitative evaluation of the difficulty of words. Next, we will establish a classification model to classify words according to the difficulty.

(1)In the classification model, we use the K-means++ algorithm to optimize the selection of the initial center of mass

①Randomly select a sample point c_1 from sample set X as the $1st$ clustering center.

②Calculate the distance $d(x)$ from the other sample points x to the nearest cluster center.

③Select a new sample point c_i to be added to the set of cluster centroids with probability $\frac{d(x)^2}{\sum d(x)^2}$, where the larger the distance value $d(x)$, the higher the probability of being selected.

④Repeat steps ② and ③ to select k clustering centers.

The K-means operation is performed based on these k clustering centers.

(2)K-means algorithm

①Taking each initial center of mass as a class, for each remaining sample point, calculate their Euclidean distances to each center of mass and assign them to the cluster in which the center of mass with the smallest distance from each other is located. Calculate the center of mass of each new cluster.

②After all sample points are divided, the location of the center of mass of each cluster is recalculated according to the division, and then the distance from each sample point to the center of mass of each cluster is iteratively calculated, and all sample points are re-divided.

③Repeat ① and ② until the center of mass no longer changes or until the maximum number of iterations is reached.

7.2 Application and Evaluation of Model IV

7.2.1 Solution of Factor Analysis Model

First, KMO test and Bartlett spherical test were carried out for variables. KMO value was 0.632, greater than 0.5, and sig value was 0.047, less than 0.05. All meet the conditions of factor analysis. We determine two main factors F_1 and F_2 with a contribution rate of not less than 90%,and obtained the formula:

$$\begin{cases} F_1 = -0.0477w + 0.4125l + 0.5041r + 0.2450v, \\ F_2 = -0.6660w + 0.2647l + -0.0814r + 0.7378v. \end{cases} \tag{19}$$

It can be seen that F_1 mainly describes the impact of word frequency and the number of repeated letters, and F_2 mainly describes the impact of letter frequency and the number of vowel letters. In other words, F_1 and F_2 qualitatively describe the difficulty of words on the macro and micro levels, and quantitatively describe the difficulty of words on the whole. The scoring formula is as follows:

$$F = \frac{0.3646F_1 + 0.3050F_2}{0.6696} \tag{20}$$

Score all words according to the scoring formula. According to the final result, the higher the score, the simpler the word is. For example, “voice” score 1.8042, “other” score 1.5527, and the lower the score, the more difficult the word is. For example, “tryst” score 0.0099, “fluff” score 0.1662. This is also consistent with people’s habit of using words.

7.2.2 Solution of Clustering Model

The data are classified according to the obtained scoring formula. First of all, we need to determine the number of categories. We have tested three categories, namely, 3, 4 and 5, and found that three categories are the most effective and persuasive. So set the three difficulty divisions of ”difficult”, ”moderate” and ”simple”, and use the clustering algorithm to get the words with three difficulty levels. It can be seen that the words in category 1 are more difficult, such as nymph, coyly; The second category of words is moderately difficult, such as Album, Train; The third category of words are relatively simple, such as black and light. The data results are in good agreement with the actual situation.

The analysis is based on the main factors and three categories, as shown in the following **Figure 13**.

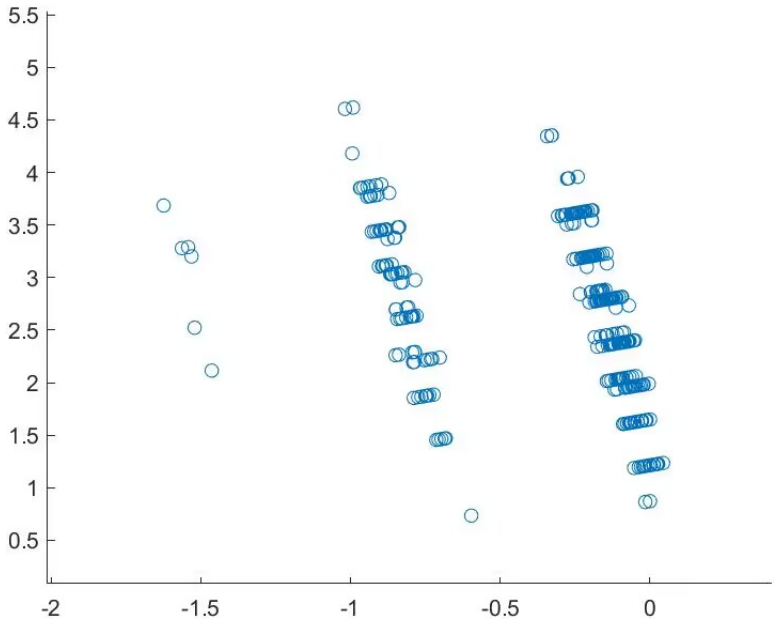


Figure 13: Sample distribution map

The abscissa is the main factor F_1, and the ordinate is the main factor F_2. It can be seen that F_1 and F_2 present an obvious linear relationship, and each category constitutes a linear family, that is, the elements in each family can be significantly determined by F_1 and F_2, and then it can be analyzed from the indicators of F_1 and F_2 that word frequency and letter frequency are the decisive factors of word difficulty.

By substituting the indicators of "EERIR", F_1 is -1.6236 and F_2 is 4.3685. It can be seen from the **Figure 13** that EERIE falls significantly in the first category on the left, which is a difficult word.

Considering the accuracy of the model, the expected formula for scoring and answering times is as follows:

$$F = -0.8212y + 3.4426 \tag{21}$$

It can be seen that there is a significant negative correlation between the two, that is, the greater the expectation of the number of answers, the smaller the score, and the greater the difficulty of the words. From the point set in the figure above, we can see that the fitting effect of the model is very good, and there are significant differences in each category. We selected some sample points for verification, and the results are in good agreement with our expectations. We have reason to think that the accuracy of the model is quite high.

8 Other Interesting Features

In fact, in addition to the data features used in the above model, we also summarized many noteworthy and interesting features in this data set:

1. After reviewing the relevant strategies, we found that there is a saying on the Internet that "slate" is the best start. Therefore, on December 9, 2022, many people have successfully guessed the right word at one time with "slate" as the starting point, so the proportion of players who guessed the right word at one time is far higher than the average level.

2. Theoretically, the probability of a player guessing the word correctly at one time is very small, with the order of magnitude of $10^{-5} \sim 10^{-4}$, and the proportion of guessing correctly at one time is more than 1% in some days, that is, thousands of people guess the data correctly at one time, so we speculate that these players may have the suspicion of searching for answers.

3. On December 25 (Christmas), the number of people participating in the game decreased significantly compared with the neighboring days (a decrease of about 25%), but on Halloween, the number of people participating in the game increased significantly compared with the neighboring days (an increase of about 30%).

4.The correct rate of the two words Dream (2022/11/5) and train (2022/5/4) to be guessed at once has also reached 5% and 6%. We speculate that it is simply because they are relatively simple.

5. In the statistics of these words, the most difficult word is parer in 2022/9/16. Half of the people guessed it more than six times or even didn't guess it. Next are foyer (2022/4/19, 26%) and catch (2022/10/15, 23%).

6. On Thanksgiving Day (2022/11/24), the word given is "feast". Because it is closely related to the holiday tradition, the number of people who guessed correctly at the first time reached 5%.

9 Sensitivity Analysis and Model Validation

9.1 Stability of Changes

From Model I, the functional relationship between the number of reported results (y) and the guest number (c) is

$$y = 5.418 \times 10^5 \times \exp(-0.01749 \times (c-199)) + 1.946 \times 10^4 \times \exp(0.0006212 \times (c-199))$$

According to the requirements of model I, we hope that the function can continue smoothly in the later stage and the function value will not reach 0. Therefore, we will draw the actual data

of 2022 /5 /6 2022 /12 /31 and our model function on the same figure (as shown in the **Figure 14**). According to the results in the figure, our model fits the actual data well and will be stable and fluctuate slightly in the range of 20000∼30000 after 2022 /12 /31. This has well verified the rationality of our model in practical application, indicating that the number of reported results will be stable in a range after the game goes online for a period of time, and will not change significantly without the interference of external factors.

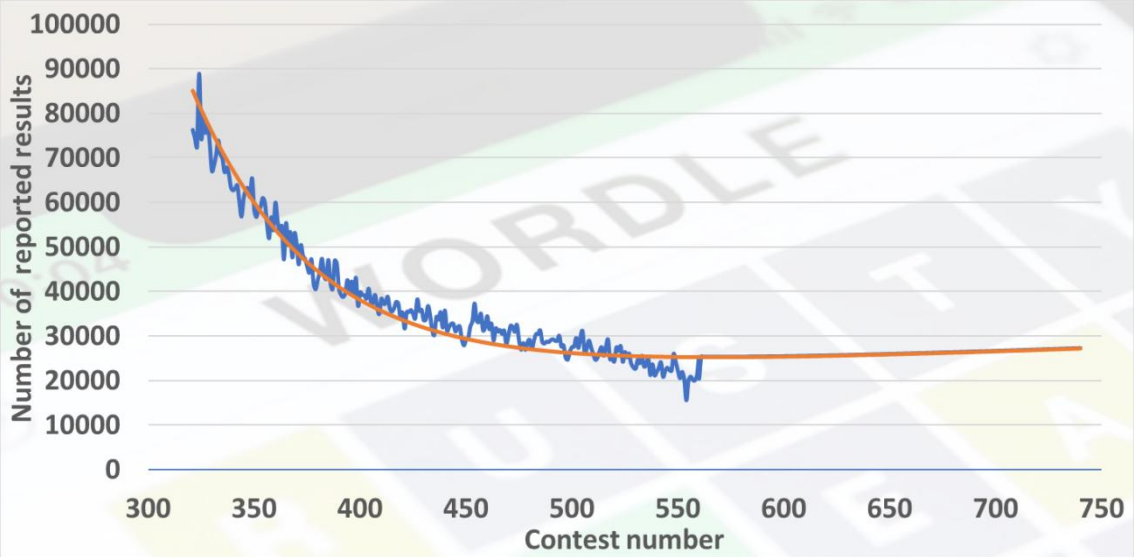


Figure 14: Prediction curve of model I

9.2 Influence of Feature Selection

Effect of the number of features on the accuracy of XGBoost model In model III, we built XGBoost deep learning model, and trained the five factors we selected in advance as characteristics to get a model to predict the distribution of success times. Here we want to explore the relationship between the accuracy of the model and the number of selected features, so we try to use any four, three or two of them as sample features to train a new model in turn, and verify the accuracy of the model on the test set together with the previously obtained model. Finally, we get the following conclusions:

Table 8: Model accuracy under different feature changes

Feature Changes	Training set correct rate	Test set correct rate
Removing the feature ”constant number”	66.9122%	62.1186%
Removing the feature ”letter usage”	50.0928%	50.0024%
Excluding the feature ”letter usage” and ”word usage”	49.9827%	46.5283%
Divide the characteristics ”number of vowels” ”letter usage rate” ”number of consecutive letters”	49.2819%	47.2923%

From the data in the **Table 8**, we can speculate that the influence of ”const number” on the prediction results is small, while the influence of letter usage rate and word usage rate on the prediction results is large, and they can be considered as indispensable factors for predicting the distribution. In general, the features selected in the process of constructing the model have a reliable influence, which also indicates that the final model is more accurate.

9.3 Influence of Factor Quantity Selection

The effect of the number of main factors on the accuracy of the model In Model IV, we identified 2 main factors. We want to explore whether the number of main factors affects the final judgment criteria, so we choose to identify 3 main factors for re-modeling and repeat the scoring process in Model IV, and we can see the following results.

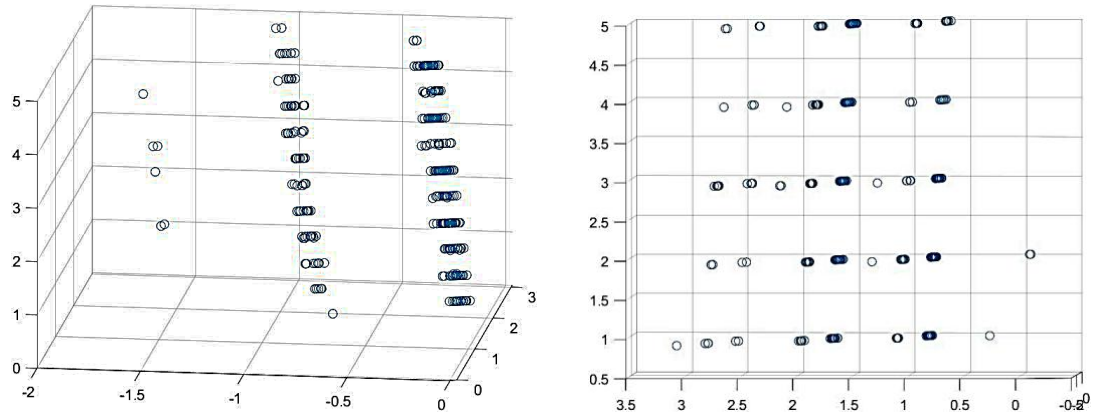


Figure 15: Sample distribution when the number of principal factors is 3

Based on the comparison of the above two figures, we can find that adding main factors does not increase the clustering effect, and from one of the dimensions, the data almost presents random distribution in disorder, so we can conclude that the characteristics of this dimension do not help the clustering effect very well, and we select two main factors in model IV is reasonable and correct.

10 Strength and Weakness

10.1 Strength

1. Through the establishment of models 1, 2, 3 and 4, we have successfully completed the analysis and construction of the number of reported results prediction model, the attributes of the word that affect the percentage of scores reported that were played in Hard Mode, the distribution model of success times and the difficulty evaluation model, which have good performance in fitting the existing data and predicting the future.

2. In the process of building the model, we have used various methods such as linear and nonlinear fitting, partial correlation analysis, XGBoost learning model, FA and so on to process and analyze different data and different objectives in a targeted way, which has good application value for diverse data in practical tasks.

3. In terms of model optimization, we use factor analysis method to determine the judgment criteria of word difficulty in model IV, which has better interpretability and accuracy than the traditional component analysis method; In classification, we use the optimization algorithm of K-Means++, which improves the accuracy of classification to a certain extent compared with traditional classification methods.

10.2 Weakness

1. In models II, III and IV, we consider a limited number of influencing attributes. Some attributes that have a strong impact on the results may not be considered, which may have a certain impact on the accuracy of the model.

2. In model III, we do not further consider the impact of each attribute on the results, but directly study the model by taking all the attributes considered as the characteristics of the sample, so that the unfiltered features may cause over-fitting of the trained model, thus affecting the accuracy of the model.

References

[1] Akima, & Hiroshi. (1970). A new method of interpolation and smooth curve fitting based on local procedures. Journal of the Acm, 17(4), 589-602.

[2] Gauss, Carl Friedrich. "Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections." (1957).

[3] Velicer, W. . (1976). Determining the number of components from the matrix of partial correlations. Psychometrika, 41(3), 321-327.

[4] CHEN T Q, GUESTRIN C. XGBoost: A Scalable Tree Boosting System[C] // ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [s.l.]: ACM, 2016: 785-794.

[5] Kaiser, H. F. . (1974). An index of factorial simplicity. Psychometrika, 39(1), 31-36.

[6] Jolliffe, I. T. . (2002). Principal component analysis. Journal of Marketing Research, 87(4), 513.

[7] Arthur D , Vassilvitskii S . K-Means++: The Advantages of Careful Seeding[C]// Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007. ACM, 2007.

Letter

Dear Puzzle Editor of the New York

I am very happy to write this letter to you. For your company's recently popular Wordle game, our team has done relevant modeling and analysis, and achieved the interpretation of the past data and the prediction of the future. Now we will introduce and show you the relevant conclusions.

1 **Number of users.**We establish a piecewise function composed of quadratic polynomial function and quadratic exponential function for the number of users. The data shows that the number of players continued to grow at a high speed during the initial boom period of Wordle, reaching the peak on February 2, 2022. And then the number of players began to decline slowly, and the decline rate gradually decreased. This trend is also in line with the market reaction after the product explosion: after a period of rapid rise, there will be a period of slow decline. At the same time, we forecast the number of players in the future, giving the number of players expected to be between 16053 and 21565 on March 1, 2023.

2 **Percentage of people in hard mode.**For the percentage of people in hard mode, we establish a partial correlation coefficient model. For time, word frequency, letter frequency, number of repeated letters and number of vowel letters, calculate the partial correlation coefficient with the percentage of people in hard mode. The result shows that time is the decisive factor affecting the dependent variable, and the attribute of the word itself has no significant effect on the dependent variable. That is to say, with the passage of time and the improvement of people's level of word guessing, more people are willing to try to play the hard mode and try greater challenges. However, people do not know the attribute and difficulty of the word before guessing it. Therefore, the characteristics of the word itself will only affect people's performance in guessing, and will not affect the number and percentage of people selecting hard mode. In this respect, the result is consistent with the logic.

3 **Future distribution trend.** In view of the future distribution trend of the number of guess, we established XGBoost model, which can predict the distribution of the number of guesses of any word in any future day. For the word EERIE on March 1, 2023, the prediction results of the model are: 0.00,4.23,23.28,32.08,24.98,13.37,2.05. The model has high accuracy and good generalization ability in both training set and test set, so there is reason to believe that our model is reasonable.

4 **Word difficulty rating.** As for how to define the difficulty of the word itself, we have established the word scoring model, factor analysis model and the word classification model K-means++ clustering model. We extracted two main factors from the four independent variables and obtained the word difficulty scoring formula. The higher the score, the easier the word difficulty is. For example, the higher the score of "voice" and "other", the more difficult the word difficulty is. For example, the score of "tryst" and "fluff" is very low. Then, we classify the results according to K-means++, and get three difficulties: "difficult", "moderate" and "simple". For example, the words in "simple" include "black" and "light", while the words in "difficult" include "nymph" and "coyly". So we have successfully quantified the difficulty of words and can give the difficulty level of any word. For example, the difficulty level of "EERIE" is "difficult". By analyzing the specific characteristics of words, we found that the determinants of word difficulty were word frequency and letter frequency, and were less related to the number of repeated letters and the number of vowels.The more words are used, the easier they are to be guessed, which is also very good in line with the actual situation.

The above is our team's modeling analysis and relevant conclusions for wordgame, and we hope they can provide the company help!

Finally, thank you again for your reading.

Yours sincerely,
Team #2320339