Problem Chosen	2023	Team Control Number
C	MCM/ICM Summary Sheet	2308512

Attribute Analysis and Model Building of *Wordel* Game

Summary

The *Wordel* game launched by The New York Times is a word-guessing game that requires players to guess a five-letter word within six steps. The game has gained a lot of popularity worldwide due to its challenging, simple, and social interactive nature. The difficulty of the puzzle is displayed as a percentage of the number of steps required to solve the problem, and is closely related to the words used to create the puzzles. Therefore, the percentage of steps required to solve the puzzle also reflects the difficulty of the words.

In this article, we focus on identifying the word properties that affect word difficulty and developing a word difficulty assessment model based on the percentage of steps required to solve the problem. We also search for methods to predict the percentage of steps required to solve the problem using word properties. To achieve our goals, we have developed three models:

Model 1: Game participation prediction model

Model 2: Game score percentage distribution prediction model

Model 3: Word difficulty classification model

As a predictive model, Model 1 is used to predict the participation range of the game at a specific time in the future. We first consider using the ARIMA model to predict the fluctuations in game participation over time. We found that the model fits well with the data on game participation over the past year. Combined with the gray prediction model, we predict the range of player participation on March 1st to be [7067, 14515].

For Model 2, we decided to use the GBDT regression model to predict the percentage distribution of word scores. We set three word properties: frequency, letter frequency, and letter repetition. Based on the model's output, we obtained the importance of the three word properties to be 50.4%, 32%, and 17.5%, respectively. After parameter adjustments and repeated training, we achieved a good prediction state for the model and finally predicted the percentage of steps required to solve the word "EERIE". The RMSE and MAPE of the model were reduced to a reasonable range, and the model fit well.

For Model 3, we performed cluster analysis on the percentage distribution of game scores, achieving a CH value of 309.688 and a good clustering effect. Finally, we classified word difficulty into three categories: easy, medium, and difficult.

We used these three models to predict game participation, analyze the correlation of word properties, predict the percentage distribution of game scores using word property values, and classify the difficulty of given words based on the predicted percentage distribution of game scores.

Finally, we wrote a letter to The New York Times, outlining our analysis and modeling process and research results related to the WORDLE game data, and providing some suggestions and thoughts.

Key words: ARIMA-LSTM Model，Gray Forecast Model，GBDT Model，Cluster analysis

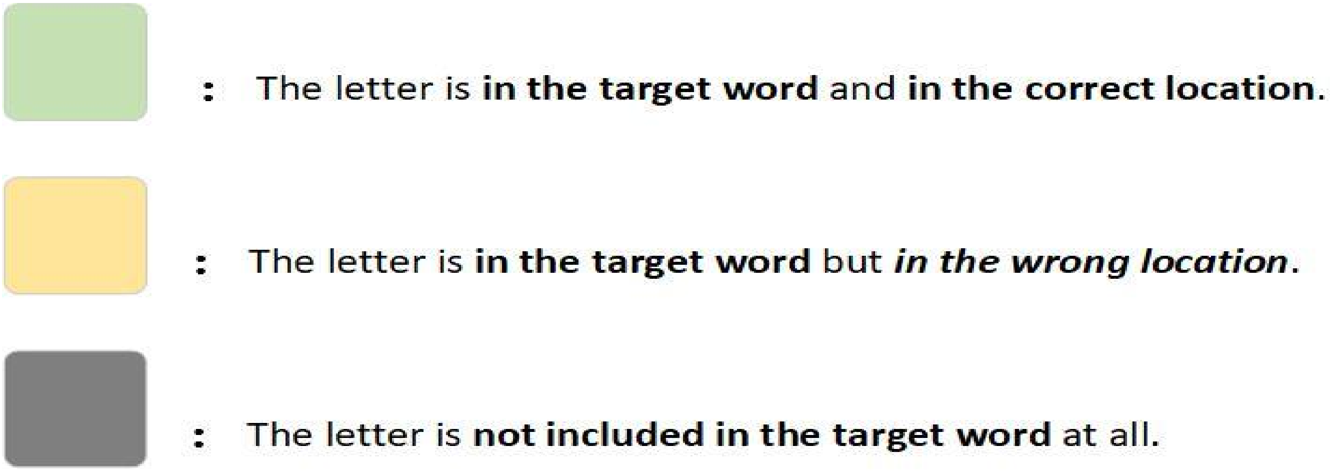
Contents

I. Introduction

1.1 Background

Crossword puzzles have a long history and wide popularity, and they vary widely. The game Wordle is also a type of crossword game, which is much simpler in difficulty and form than other crossword puzzles. Wordle updates only one question per day, and the player's final goal is to guess a five-letter English word that is meaningful in no more than six attempts times. Attempts that are not recognized as words by the game system are not allowed and do not count as one successful attempt.

Wordle's interface is quite simple and intuitive. It consists of a 5x6 array of blocks, and players are expected to complete a single attempt by inputting a letter on the keyboard below. After each try, the player will receive a response related to color. After one guess, the base of the five letters will be represented in corresponding colors to give feedback on the result, and give hint for the next attempt. Different colors of squares indicate different meanings of feedback (Figure 1). Then, the player continues to guess based on the feedback, with a maximum of six chances.



🟩 : The letter is **in the target word** and **in the correct location**.

🟨 : The letter is **in the target word** but ***in the wrong location***.

⬛ : The letter is **not included in the target word** at all.

Figure1: Explanations of the Feedback

The rules of Wordle are simple, however, the scarcity of the game that only updates one question a day contributes to the sense of challenge and anticipation. What's more, players can choose to play in either regular mode or " Hard Mode". In the regular mode, in six tries players can fill in any actual word they think fits the direction of the guess. In the "Hard Mode", players are required to reproduce the correct letters (the tile is green or yellow) in the feedback and then use them in the next attempt. At the end of the game, Wordle will give the player statistics showing the correct rate of each step, the number of consecutive days logged into Wordle, and the countdown to the next game.

1.2 Restatement of the Problem

Considering the background information and restricted conditions identified in the problem statement, we need to solve the following problems:

- Develop a model to explain the variation and predict the interval for the number of reported results on March 1,2023. And Judge whether and how the attributes of the word affect the percentage of scores reported in Hard Mode.
- Establish a model to make a prediction for associated percentages on a future date. Give uncertainties associated with the model, and predict one example word on certain day.
- Identify the attributes of one given word based on the model which classify solution words by difficulty.Show other interesting features of the given data set.
- Write a letter to summarize the results to the Puzzle Editor of the New York Times.

1.3 Our Works

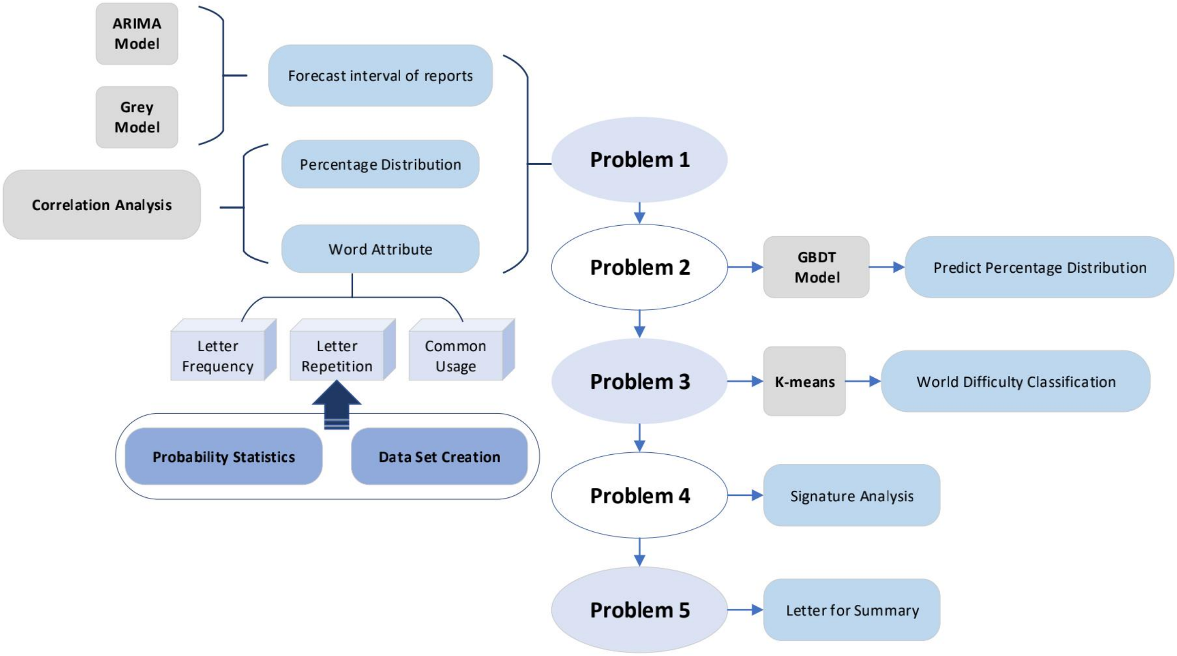


Figure 2 : Our Work

II. Basic Assumptions

Assumption 1: It is assumed that the players who participated in the word-guessing game for the given data records played the game solely with the aim of solving the problem.
Assumption 2: It is assumed that the word bank for the selected words in the word-guessing game is sufficiently extensive and diverse, and that the selection of words for adjacent word-guessing games is not related and is randomly independent.

Assumption 3: Since no repeated words appeared in the given data, it is assumed that each word can only be used once in the Wordle game. This will avoid the repetition of words and increase the uncertainty of the game.
Assumption 4: It is assumed that the words used in the game must be standard English words, excluding abbreviations, proper nouns, and words with irregular pronunciation.
Assumption 5: It is assumed that the words used for the game are not extremely obscure.

III. Notations

The key mathematical notations used in this paper are listed in Table 1.

Table 1: Notations Used in This Paper

Symbol	Description
γ_t	Time series of the number of reported results per day
μ	The constant term
φ	The influence coefficient of series data to predicted points
θ	The influence coefficient of predicted points to series data
ε_t	The prediction error term
f_L	The letter frequency
r_L	The letter competition
f_u	Degree of common use
S	Score of word difficulty

IV. Data Pre-Processing

Whether the data is clean or not is directly related to the deviation and accuracy of the model, and will also affect the final conclusions to some extent. Therefore, we analyze, process ,and clean the raw data to enhance the credibility and trustfulness of our results.

4.1 Data Filtering

Since game Wordle is to guess a five-letter word in six or less tries, words which are not five-letter will be eliminated from the data set when extracting and computing the probability of occurrence of each letter in the five-letter word and the probability of occurrence of each position letter of the word from the data set. We use the relevant **python** code to import the **EXCEL** data file, then select and analyze the data in the “Word” column.

4.2 Data Sequence Adjustment

On the other hand, we will make an interval forecast for the number of reported results on one future date, therefore, our reference data should be represented in a time sequence from the past to the present. However, the given data are just on the contrary, so we arrange the data in chronological order.

4.3 Outlier Processing

The ranked data of **“Number of reported results”** column is fluctuating over time and with the increase of the contest number. Finding some values in which fluctuations are too large, these abnormal value are considered as outliers. We take related measure to adjust and replace these outliers. Firstly we select the data in the column **“Number of reported results”** and use the rolling function to calculate the moving standard deviation. According to the calculated data, we regard the data with standard deviation greater than a certain threshold which is set to 2 as outliers. Next, iterate through the index of outliers and use shift function to obtain the number of reported results in the previous row. Then replace the outliers with the value we just have got, which completes the processing of outliers.

4.4 Percentage Alignment

Given that the percentages of players guessing the words of different times may not always sum to 100% due to rounding. Therefore, we firstly calculate the sum of the percentages of the given data. Since the error due to rounding is generally within three percentage points, data with the error beyond this value is recognized to be unreliable data. Then we will choose to use **method of near replacement** for modification.To ensure the sum is 100%, we move the percentage of errors to the percentage of players who could not solve the puzzle in six or fewer tries as the expected step. Under the principle that the percentage can never be negative, part of data errors will be moved to the percentage of 6 guesses, and so on. After processing, the sum of the percentages of players guessing the puzzle in corresponding tries is 100%.

V. Problem I: Forecasting Intervals and Percentage Distrubutions

5.1 Analysis of Problem I

For a word, we can define and classify it from different perspectives.Different definitions lead to different types of words.What determines the definition of a word is its attribute, which is human's abstract description of an object.For example, in English words, there are noun, verb and other word properties, which can be classified as the word attribute of the part of speech.Take part of speech for example, a word can be either a noun or a verb.By extension, a word may have many word attributes.The word apple, for example, is a noun in part of speech attributes and a five-letter word in letter number attributes.However, in this game scenario setting, the guessing range is limited to the actual English word with five letters.

In Problem 1, we are going to develop a model to explain the variation and predict the interval for the number of reported results on March 1,2023. Next , we will judge whether and how the attributes of the word affect the percentage of scores reported in Hard Mode. We establish two models to predict the interval for the number of reported results on the certain future date.

5.2 Model I : ARIMA Time Series

The problem of predicting the number of reported results in the future is a typical time series problem. The purpose of time series is to predict future trends.ARIMA is the most commonly used time series forecasting model. However, ARIMA requires the timing data to be stationary, or stable after processing differences. And ARIMA can only capture linear relationships in essence, but not nonlinear relationships. Therefore, to predict the time series data using the ARIMA model, the data should be steady. If the given data is unstable, it cannot capture the pattern. For example, the reason why Bitcoin (BTC) data cannot be predicted with the ARIMA model is that it is non-stationary and generally fluctuates due to policies and news. We conduct modeling prediction of data of the number of reported results in a given time.

After pre-processing the data, we build the ARIMA(p,d,q) model on it. The Box-Jenkins method is a highly accurate algorithm for the analysis and prediction of time series data, among which the ARIMA model is one of the outstanding models to predict future data. In the formula below L is the hysteresis operator. By selecting the corresponding data as the variable, the date column as the time item, and the number of participants as the time series data, the backward prediction unit in the corresponding parameter is 61, and the automatic optimization mode of the parameter is confirmed.

5.2.1 Model Establishment

The basic idea of ARIMA is to predict the variable in future , the related formulas of which are as follows. The data set of our model is : $Y_{(t)} = (y_1, y_2, ..., y_T)$.

Since the time series data itself has a lag number, and the lag number of the prediction error is also uncertain whether it has stationarity, we assume that the model is

$$\hat{Y}(t) = u + \varphi_i \sum_{i=1}^{P} yt - i + \theta i \sum_{i=1}^{9} \varepsilon_{t-i}$$

At the same time, we need the weights of the time series data and the weights of the prediction error, so we use the least squares method to solve this.

$$\hat{Y}_t \approx \varphi^\top v_t \quad (t = p + 1, \ldots, T)$$

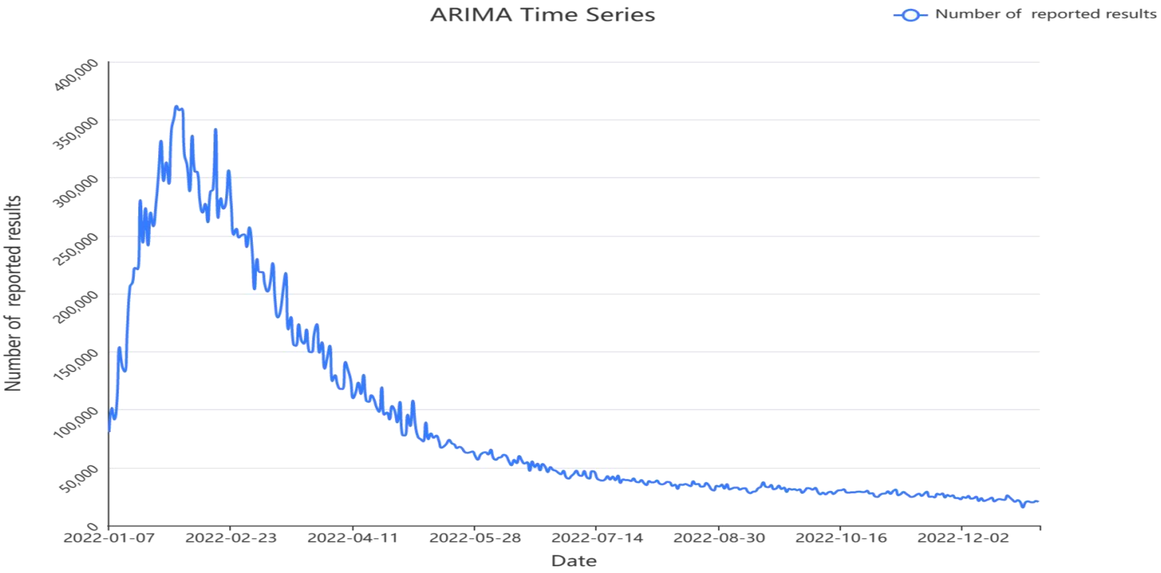
$$v_t = (y_{t-1}, y_{t-2}, y_{t-3}, \cdots, y_{t-d}) \in R^d$$

$$z = (y_{p+1}, y_{p+2}, \cdots y_T) \in R^{T-d}$$

$$Q = \begin{bmatrix} V_{d+1}^{\top} \\ \vdots \\ {V_T}^{\top} \end{bmatrix} \in R^{(T-d)\times d}$$

$$\varphi = \left(Q^{\top} Q\right)^{-1} Q^{\top} z$$

Similarly, the weight coefficient of the prediction error Et can be obtained.



5.2.2 Analysis Procedure

Firstly, the ARIMA model requires that the sequence should be stationary. We test whether the time series is stationary by analyzing the ADF test list(see table: ADF test table). Since the sequence in this model must be stationary time series data, by analyzing the t value, we primarily analyze whether the original hypothesis of sequence instability can be significantly rejected. If it is significant, which means the P value is less than 0.05, the null hypothesis will be rejected. Therefore, the series is a stationary time series. Otherwise, it means that the series is an unstable time series. If the ADF Test result is less than three critical values which are 1%, 5%, and 10% at the same time, we can reject the hypothesis with good confidence. Through the predicted result of the time series and based on the relevant variable "Number of reports", we find that when the difference order is 0, AIC value is 7150.086, and the data shows horizontal significance. Therefore, we reject the null hypothesis, the time series is the most stable among the three-time series.

Table 2: ADF Testing Table

ADF Testing Table

Variable	Difference Order	t	P	AIC	Critical Value		
					1%	5%	10%
Unnamed: 5	0	-3.984	0.001***	7150.086	-3.45	-2.87	-2.571
	1	-4.169	0.001***	7116.656	-3.45	-2.87	-2.571
	2	-9.318	0.000***	7116.629	-3.45	-2.87	-2.571

Tips：*** , **,   *   represent 1%、5%、10% significance level respectively

Secondly,Check the data comparison graph before and after the difference to determine whether it is stable, which means whether the fluctuation range is large or not. At the same time, we conduct autocorrelation analysis (ACF) and partial autocorrelation analysis(PACF) for the time series and estimate its p and q values according to the truncated situation. For ACF, it consists of a coefficient, upper confidence limit, and lower confidence limit, where the horizontal axis represents the number of delays and the vertical axis represents the autocorrelation coefficient. The ACF graph is truncated at order q, and the PACF graph is trailed. The ARMA model can be simplified to MA(q) model. For the partial autocorrelation graph, which concludes coefficients, upper confidence limit, and lower confidence limit. The p-order is truncated, the ACF graph is trailed, and the ARMA model can be simplified to an AR(p) model.

ARIMA model requires that there is no autocorrelation in the residual error where the residual error of the model is white noise. By analyzing the autocorrelation diagram of model residuals, the partial autocorrelation diagram of model residuals, and viewing the model test table, we conclude that the model white noise is tested according to the P value of Q statistic.If the P value is greater than 0.1, it is white noise. And according to the information criteria AIC and BIC values for multiple analysis model comparison,the lower the result is, the better the effect is. R^2 represents the fitting degree of time series, and the closer it is to 1, the better the effect is.

Table 3: Model Parameter Table

	Model Parameter Table					
	Coefficient	Standard Variance	t	P>|t|	0.025	0.975
Constant	-225.559	466.119	-0.484	0.628	-1139.135	688.017
ar.L1.D.Unnamed: 5	-0.359	0.049	-7.269	0	-0.456	-0.262

Tips：*** , **, * represent 1%、5%、10% significance level respectively

5.2.3 Model Fitting Results

Based on the given data of "Number of reported results", the system automatically finds the optimal parameter based on AIC information criterion. The model result is the ARIMA model (1,1,0) test table based on 0 difference data. The model formula is as follows:

$$Y (t) = 225.559-0.359 * y (t - 1)$$

And the relevant time series graph (attached figure) represents the original data graph, model fitting value, and model predicted value of the time series model.

5.2.4 Model Testing

By the analysis procedure above, we test that the established model is appropriately fitting for the related requirements. The sequence is stationary and no autocorrelation of the residual error by testing the ADF, AIC, ACF, and PACF. ARIMA is based on historical period data to predict future period data. Finally, we get the model's fitting degree squared is 0.982. We find that the model performs well in the final result.

5.3 Grey Prediction Model

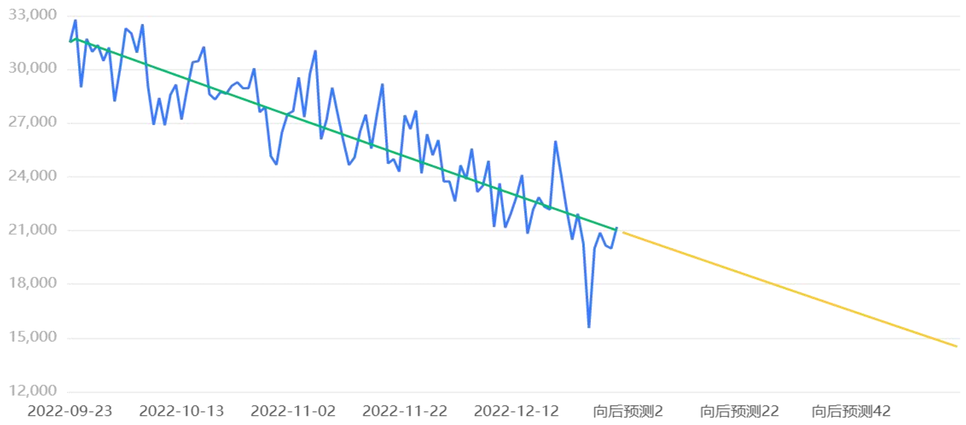
5.3.1 Model Establishment

To improve the accuracy of the prediction interval, two models, time series and gray prediction, were developed to predict the interval of the number of reports at a future date. Gray theory considers that all random variables are gray quantities and gray processes that vary within a certain range and over a certain period. Based on the data of the latter one hundred reported quantities and dates, we build the GM(1, 1) model. Before building the gray prediction model GM(1,1), we perform a cascade test on the time series. If the series passes the rank-ratio test, then the series is suitable for building the gray model, and if it does not pass the rank-ratio test, then the series is "transformed" so that the new series meets the rank-ratio test. The gray prediction model can be tested to determine whether it is reasonable or not, and only the model that passes the test can be used for prediction, and the system mainly uses the posterior difference ratio C value to test the gray prediction model.

5.3.2 Model Prediction Results

We found that all the level ratios of the translation-transformed series lie within the interval (0.98, 1.02), which indicates that the translation-transformed series are suitable for the construction of the gray prediction model. And through the gray model construction (e.g., Table: Gray Model Construction), we analyzed that all the level ratios of the translation-transformed series are within the interval (0.98, 1.02), which indicates that the translation-transformed series are suitable for the construction of the gray prediction model.

By analyzing the relevant model fitting results table, the average relative error of the model is 4.756%, which is much less than 20%, implying a good model fit. (As shown in the figure: model fit prediction graph) By analyzing the prediction result table of the gray prediction model, we found that the number of reported results for the 61st day forward, i.e., March 1, 2023, was 14,515.274. Combined with the time series model analysis, based on rounding, we finally arrive at a prediction interval of (7067,14515).



5.4 Effects of Word Attributes on Percentage Distributions

5.4.1 Analysis of Process

We wanted to examine the effect of word attributes on the percentage of points reported played in hard mode.First, through data processing and research we will determine the attributes of the words that participate in the analysis.According to the algorithm, two word attributes, letter frequency and letter repetition, are considered to have a greater impact on the percentage of the score played in difficult mode.

Next, based on the degree to which the selected word attributes can affect the difficulty of the word, we will quantify the two word attributes to find the correlation degree and establish the relationship between the word attributes and the difficulty of the word.

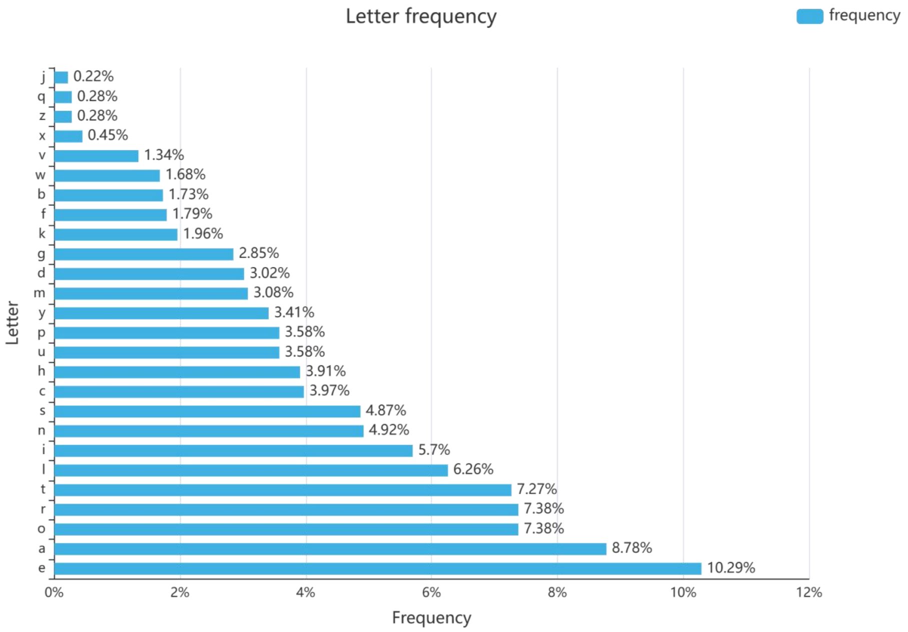
Then, we establish the final score index of the corresponding word under different word attributes through the algorithm, so as to judge the difficulty degree of the word according to the score.Therefore, the higher the score, the more difficult it is for the player to guess the word, and the corresponding percentage of different plays will vary.

Finally, we neutralized different score situations, combined with the table and corresponding graph, and obtained the influence of different word attributes on the percentage of points played in the report of hard mode.

For the final percentage distribution, we convert it into the corresponding score for better quantitative calculation.We will give a score of 1 in the case of successfully solving the problem on a single attempt;2 points for two attempts, and so on up to 6 points.And the corresponding X column, that is, more than six attempts or give up halfway within six attempts, is recorded as seven points.

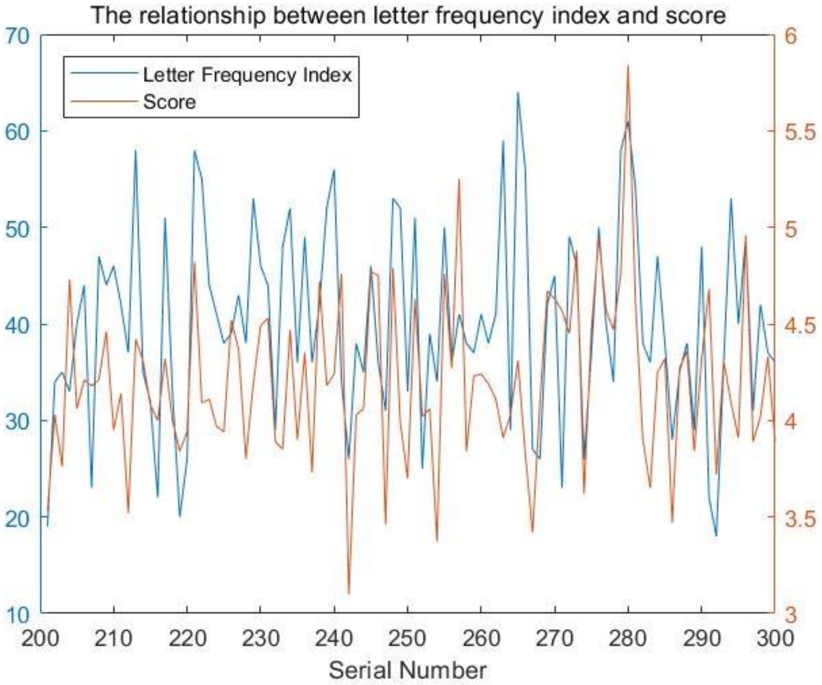
5.4.2 Attribute 1: Letter Frequency

Letter frequency is the frequency distribution of letters appearing in the word bank.Generally, letters that occur more frequently in A word (e.g., E, A, I) are easier to guess than letters that occur less frequently (e.g., J, Z, X), so the frequency distribution of letters in a word may affect difficulty.Relevant research shows that in all English single words, the top five letters with the highest frequency among the 26 letters are E,T,A,O and I, with corresponding frequencies of 12.25%,9.41%,8,19%,7.26% and 7.1%, respectively.



In this game scenario setting, only English words composed of five letters with practical meaning are eligible, and the corresponding word bank is the latest period of data given on Twitter.If the dictionary is used as the object of the word bank, the result is very inaccurate, and the error is very large, so we lock the data given by the title, the frequency of letters in the word for statistics and processing, and finally get the corresponding letters and letter frequency as shown in the following table , the first five are E,A,O,R,T.,

In general, the more frequently a letter appears in a word, the easier it is to guess, while the less frequently it appears, the harder it is.Thus, the word attribute of letter frequency can affect difficulty.We assign points to different letters, and the letter with the highest frequency is considered the easiest to guess, so we assign 1 point, followed by 2 points, and so on.We assign 1 to 26 points to the letters, and the final score is the number of points added up among the target words.The lower the score a word receives, the more difficult the letter frequency indicator is considered. Under the same target word, the relationship between the letter rating index and the final score and the word respectively, we found that the coincidence showed almost synchronous changes.Therefore, we draw a final conclusion: letter frequency as a word attribute has an effect on the percentage of reported scores played in hard mode.



5.4.3 Attribute 2 : Letter Repetition

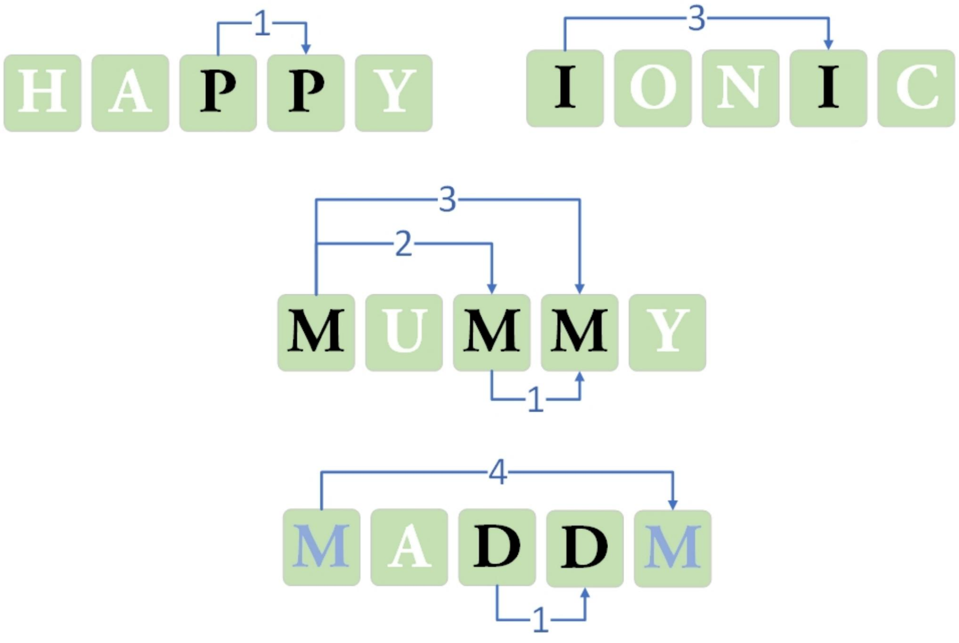
Letter repetition is the number of repeated letters in a word and their corresponding times.The number of repeating letters in a word may affect the difficulty of guessing.If more than one repeating letter appears in a word, you can more quickly eliminate words that do not contain repeating letters, and lock in the word bank of words with quasi-repeating letters in five-letter words, thus improving the probability of guessing the word correctly within a specified number of times.Thus, to some extent, the degree of letter repetition affects the percentage of scores reported playing in hard mode.

We have established defining indicators for how to measure the number of repeated letters in a word and their corresponding times.By analyzing the repetition of the words in the corresponding word bank, we find that there are four categories in a five-word word.According to these four categories, we use "repeated letters -- corresponding number of repetitions" for measurement, which is divided into 1-2, 1-3, 2-2, 1-1.The four categories are illustrated in a table: (in the process of measuring the degree of repetition, including but not limited to whether the letters are linked together)

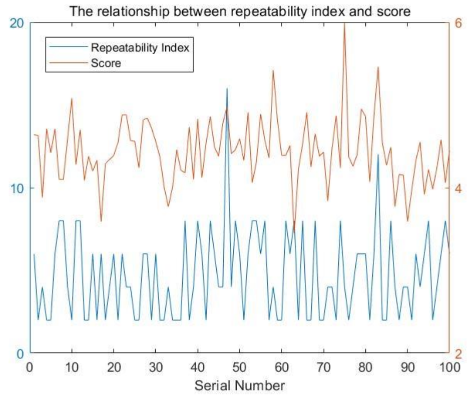
1-2: A word in which a letter appears twice, e.g., IONIC;
1-3: three occurrences of a letter in the word, e.g. MUMMY;
2-2: There are two types of repeated letters in the word, each appearing twice, e.g. MADAM;
1-1: there are no repeated letters in a word, so for easier description and calculation, we define them as 1-1 rather than 0-0, such as UPSET;

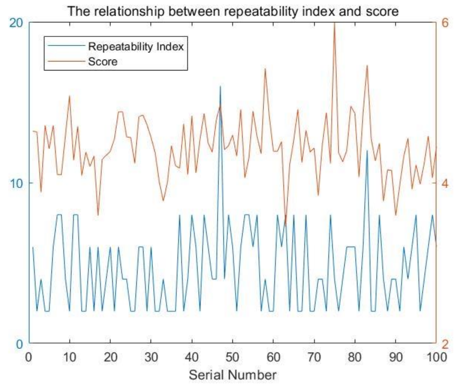
To some extent, the above four categories reflect the frequency of letter repetition in the corresponding word library.Statistically, we found that words with more repeated letters were less frequent, and therefore less common, and more difficult to guess, and required more attempts, thus affecting the overall reporting percentage distribution.In addition, for words with repeated letters, the distance between the repeated letters also had an effect on the difficulty of the word.When it comes to apple and ionic, it's clear that we're more likely to guess the apple, because people tend to associate the word more easily with duplicates.

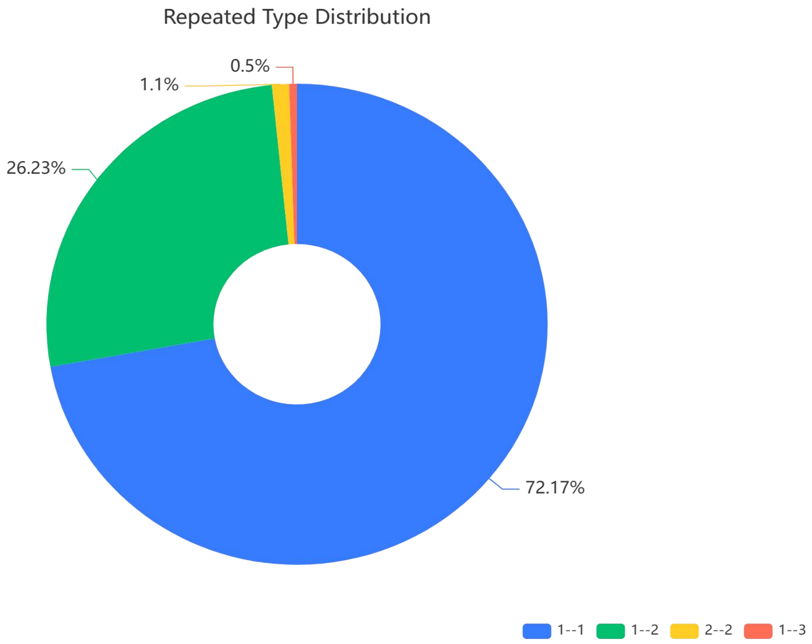
In this regard, based on the above measures and the corresponding distance, we use the algorithm to establish a system to assign points for the word attribute -- letter repetition.Within a word, one letter to another letter, if the distance between one position is assigned 1 point, two positions are assigned 2 points, and so on.The scoring of the distance between our two letters is directional, so it requires repeating the calculation of the distance.(See the following flow chart: Distance 1-2 (to write the overlapped word, apple, ionic);1-3 (mummy);2-2 (madam) example)



Therefore, a word with a higher repetition index is more difficult to correspond to, which translates into a lower score, and to some extent affects the percentage distribution of the report.Through the visual analysis of the graph (two figures), we found that under the same target word, the relationship between the letter repetition and the final score and the word respectively, we found that the coincidence showed almost synchronous changes, and the range between 200 and 300 was more consistent with our expected results.Therefore, we draw the final conclusion that the word attribute of letter repetition has an effect on the percentage of scores reported in hard mode.

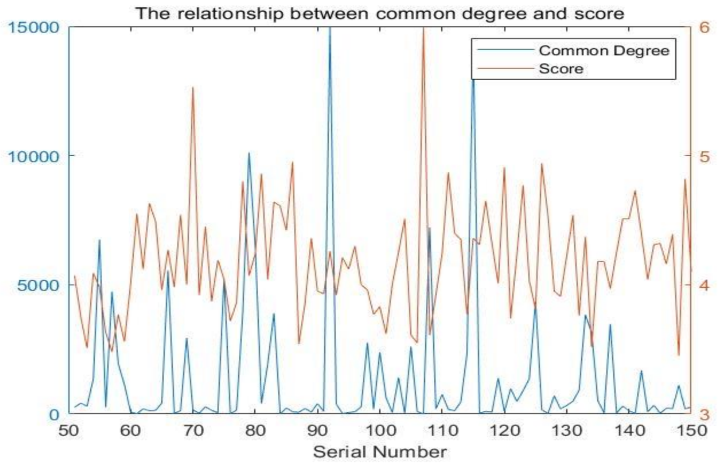






5.4.4 Attribute 3 : Degree of Common Us

Like above analysis, We multiply the frequency of the word in the last three years by 10 to the power of 6, and finally get the corresponding quantitative index. Finally, we obtain the relationship among them.



VI. Problem II : Prediction of Degree of Common Use

6.1Analysis of Problem 2

In problem 2, we are going to develop a model to predict the distribution of the reported results for a given target world. To forecast the relevant 7 types of percentages (1,2,3,4,5,6,X) for a future date, we establish a model called Gradient Boosting Decision Tree (GBDT) to make prediction of the percentage distribution.

6.2Essential Ideas

GBDT is an integrated algorithm based on decision tree called CART regression tree, and is combined with gradient descent, which has a remarkable effect in data analysis and prediction.Three basic ideas are included in GBDT.

6.3 CART Regression Model

6.3.1 CART

CART, also known as Categorical Regression Tree, is a decision tree optimized on the basis of ID3. The relevant formula is as follow:

$$f(x) = \sum_{m=1}^{M} cmI \qquad (x \in R_m)$$

The key variable M is to divide the data set into M units. The cm means the output value is corresponding output of unit. And the I is the indicator function.

6.3.2 Boosting

Boosting, which belongs to a class of ensemble learning, is a way to generate learners. Generally, there is strong relationship among individual learners. We add multiple weak learner to produce a new reinforced learner.

$$\mathrm{F}(x) = \sum \left(f(x_i)\right)$$

6.3.3Gradient Descent

According to the specified step size along the opposite direction of the gradient, we carry out the iterative search, where a represents the step size and is a constant. We keep updating the value of x while minimizing the objective function until it converges.

$$\mathrm{X}(t) = X_i + a\, f'(x_i)$$

In short conclusion, GBDT model combines three ideas which are listed above.Its core idea is that Using the negative gradient, that is, the reverse direction of the gradient,approximates the residual error.

6.4 Theory of GBDT

6.4.1 Initializing the Weaker Learners

Suppose the Loss function L (y,f(x)) :

$$\mathrm{f}_0(x) = arg\ min + \sum_{i=1}^{N} L(\,y, f(x))$$

We assume that the loss function is a square loss, then the function is a convex function. Next, we take the derivative of c , and make the full derivative equal to 0, where the value of c is the mean value of all the training sample label values.The initial learner is :

$$\mathrm{f}_0(x) = \mathrm{c}$$

6.4.2 Iteration Training

The purpose is to select the m=1,2,...,M trees. There are four steps:

(1) For each sample i = 1,2,...,N , we calculate the negative gradient, which also is called the residual: $r_{mi} = -\left[\frac{\partial L}{\partial f(x_i)}\right]$

(2) We take the residual r_{mi} obtained in setp(1) above as the new true value of the sample, and convert the data (x_i, r_{mi}) where i = 1,2,...,N as the training data of the next tree, a new regression tree is obtained, whose corresponding leaf node region is R_{mj} .

(3) We make full use of each leaf node, and calculate the best fitting value.

(4) Finally, we update the strong learner.

6.4.3 Getting the Final Learner GBDT

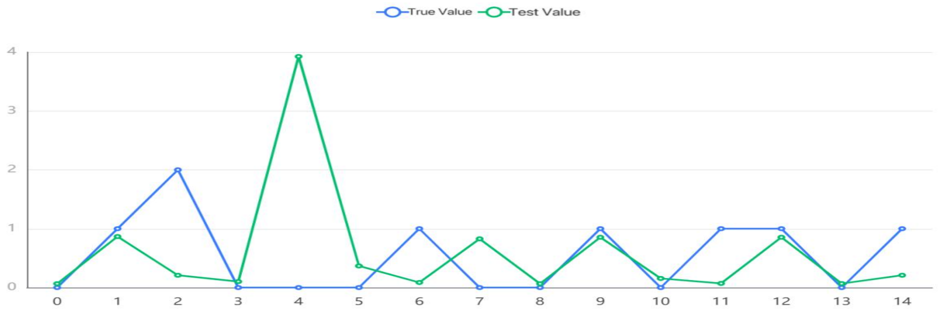
$$f(x) = f_M(x) = f_0(x) + \sum_{m=1}^{M}\sum_{j=1}^{J} c_{mj} I_{x\in R_{mj}}$$

6.4.4.Testing

Next, we use SPSSPRO to train the GBDT model based on the three word attributes including Letter Frequency, Degree of Common Use, and Letter Repetition Index. Since there are 7 situations (1,2,3,4,5,6,X), we train the GBDT model 7 times respectively. And the corresponding results including tables and figures are as follows.

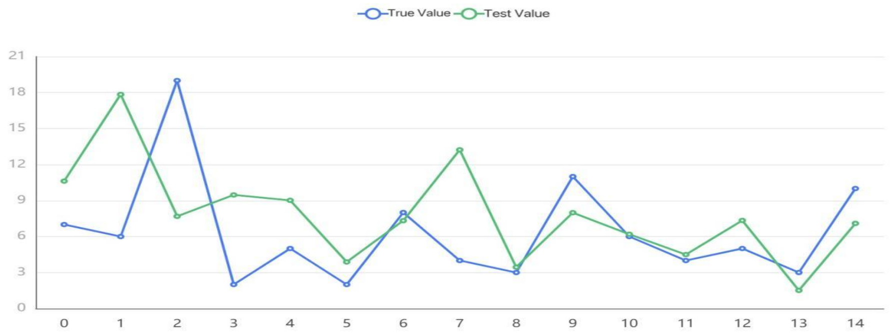
1 Try

	MSE	RMSE	MAE	MAPE	R²
Training Set	0.031	0.176	0.142	70.638	0.949



2 Tries

	MSE	RMSE	MAE	MAPE	R²
Training Set	0.84	0.917	0.723	15.389	0.949



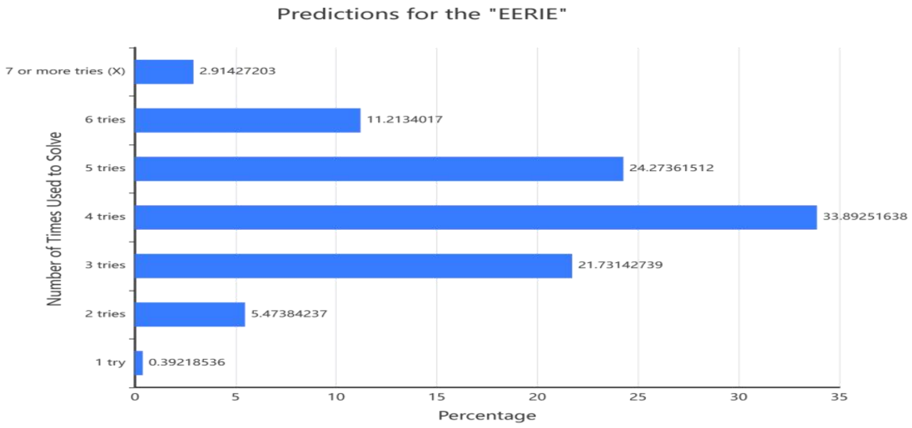
We were surprised to find that the seven models after training all had a high degree of fitting (large than 0.85) , so we think the models we have built are successful. And, we brought the test set into the model for prediction and obtained the corresponding model evaluation results. We find that the fit degree was 0.726, indicating that the fit degree was objective. Then, GBDT classification model was used to predict the uploaded data, and the predicted results were obtained, as shown in the following table.

	MSE	RMSE	MAE	MAPE	R^2
Test Set	0.167	0.409	0.348	88.841	0.726

6.5 Prediction for the word EERIE

We use GBDT model to predict the word EERIE on the certain future date -- March 1,2023. First, we quantified and evaluated 3 word attributes respectively. Then we train the model with the corresponding index of difficulty. Our final percentage score is as follows. And we get the distribution of the word EERIE based on different tried times:

1 try	0.39218536
2 tries	5.47384237
3 tries	21.73142739
4 tries	33.89251638
5 tries	24.27361512
6 tries	11.2134017
7 or less tries (X)	2.91427203



6.6 Uncertainties of Model and Prediction

There are many influence factors related to the distribution of the result percentage, like the number of vowels or consonants in a word, the part speech and so on. And the culture degree of the player may affect the final result of the report percentage distribution.For a word, we can define and classify it from different perspectives. Different definitions, the corresponding classification of words will be different.

What determines the definition of a word is the attribute of a word, which is human's abstract description of an object. Take part of speech as an example. A word can be either a noun or a verb. By extension, a word may have many word attributes. The word APPLE , for example, is a noun in part of speech attributes and a five-letter word in letter number attributes. In this game setting, the scope of guessing words is limited to five letters of actual English words. Therefore, there are many uncertainties which are associated with the GBDT model and prediction, though we have high confidence that the model's prediction is quite objective.

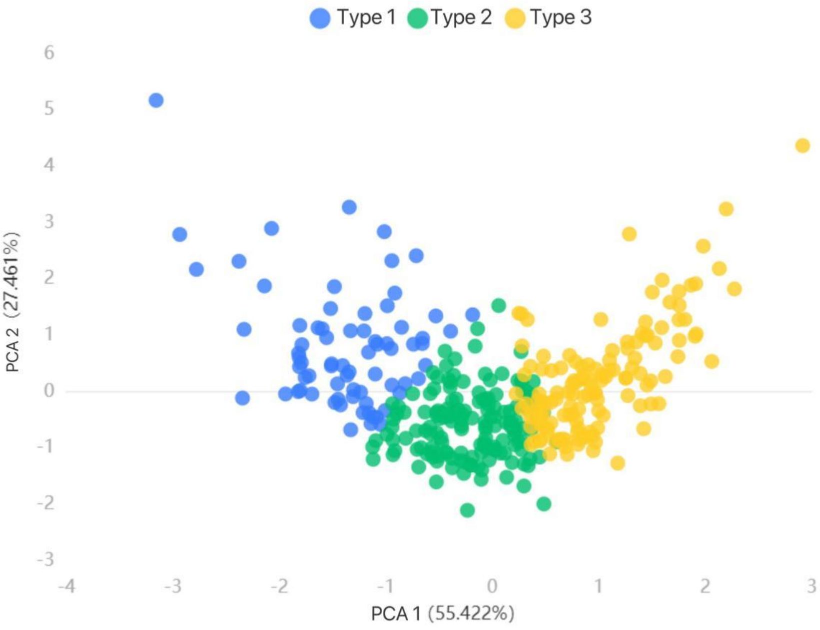
VII. Problem III : World Difficulty Classification

7.1 Analysis of the Problem

We use the method of clustering analysis to classify the words according to the difficulty. Clustering is a technology that finds the internal structure between data. The cluster organizes the instance of the data set into some similar groups. The data instances in the same cluster are the same, and the instances in different clusters are different.

7.2 Cluster Analysis Model

We bring the percentage distribution of the daily number of people into the difficulty of the difficulty of the daily designed cluster analysis model, as shown in the figure.



7.3 Model Evaluation

We can see from the figure very intuitively that we divide the data into 3 categories and use 1, 2, 3. The distribution situation can be seen that the clustering effect is objective, and we also evaluate the accuracy of the model:

silhouette coefficient	DBI	CH
0.367	0.919	309.688

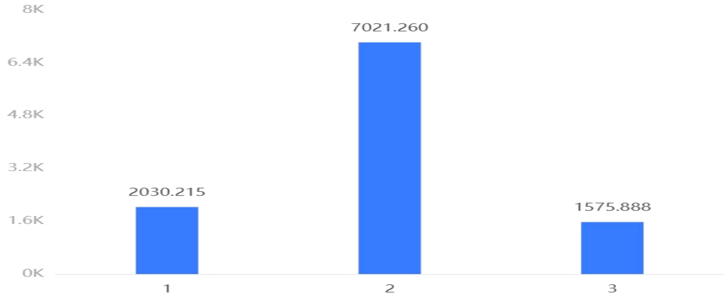
Outline coefficient: For a sample set, its contour coefficient is the average value of all sample contour coefficients. The range of the contour coefficient is [-1,1]. The closer the sample distance of the same type, the farther the different category sample distance, the higher the score, and the better the clustering effect.

DBI (Davies-BOLDIN): This indicator is used to measure the distance between the cluster distance from the cluster of any two clusters. The smaller the indicator, the better the clustering effect.

CH (CALINSKI-Harbasz Score): By calculating the tightness (division) of the distance between the distance between the points and the centers in the class and the category of the category, through the calculation of the center point of the category and the data set point distance and the caller data set, The separation (molecule), the CH indicator is obtained by the ratio of the separation and the tightness. The larger the CH, the better the clustering effect.

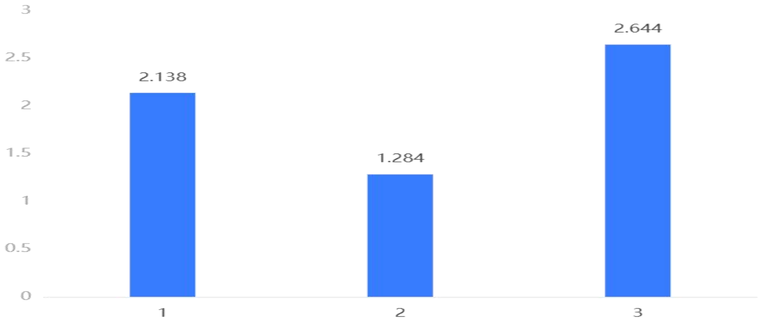
From this objective, the effect of clustering is very considerable, and the effect is good.

For identifying the attributes of the derived word associated with each classification, we use a simple classification summary average analysis to analyze the impact of each word attribute on the category.



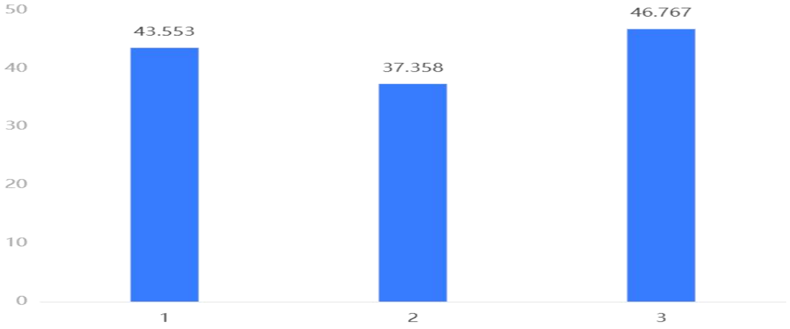
Letter Frequency Index Classification Mean Value

From the perspective of the perspective of the figure, the higher the average value of the common degree, the more common the words represent, the easier it is to guess the words.



Degree of Common Use Index Classification Mean Value

From the figure, from the analysis of the alphabet frequency indicator, the higher the frequency indicator of the letters, the easier it is to guess the words.



Letter Repetition Index Classification Mean Value

From the figure, the analysis of the alphabet repetition indicator, the higher the repetitive indicator of the letters, the easier it is to guess the words.

Overall, from the three aspects of analysis, there is no doubt that the difficulty of the three categories is from low to high:

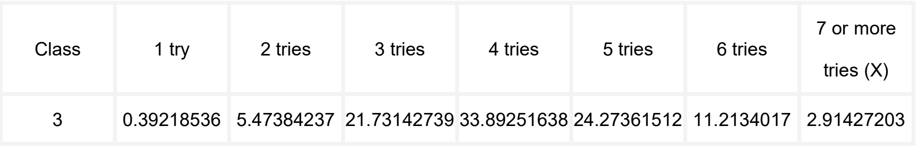
2, 1, 3, simple, medium, difficulty, three types.

The following is an example of some of our prediction results:

Type of Cluster	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (X)
2	0	2	17	37	29	12	2
2	0	4	21	38	26	9	1
2	0	2	16	38	30	12	2
2	0	3	21	40	25	9	1
2	0	2	17	35	29	14	3
1	0	2	8	16	26	33	14
2	1	5	20	35	28	10	1
3	2	11	34	32	15	6	1
3	0	7	26	35	20	10	3
2	0	1	13	34	34	15	2

7.4 Case Prediction:

By introducing the clustering analysis model, we can see that the corresponding clustering types and difficulties of "EERIE" are as follows:



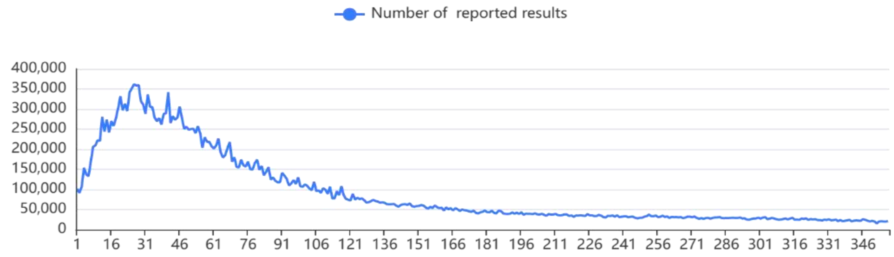
Class	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (X)
3	0.39218536	5.47384237	21.73142739	33.89251638	24.27361512	11.2134017	2.91427203

From the predicted results，“EERIE” corresponds “difficulty“.

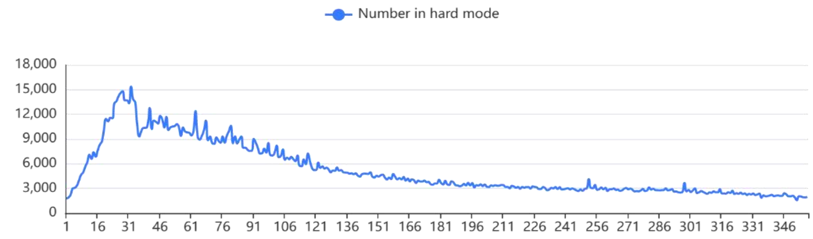
VIII: Problem 4 : Signature Analysis

8.1 Interesting features of this data set

Changes in the number of game participants over time; Change in the number of people playing hard mode over time; Change in the percentage of people playing Hard Mode as a percentage of the total number of people playing that day's game.

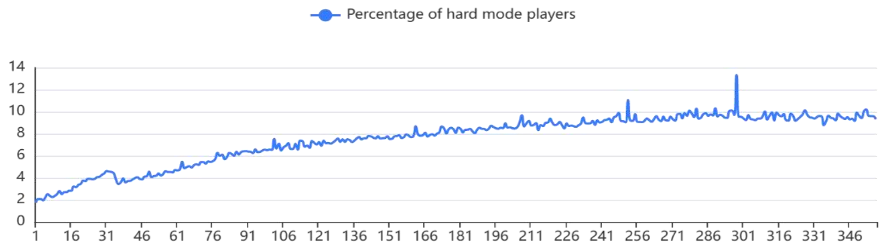


Change in the number of players over time; In both cases, the number of people playing hard mode increased sharply over time, then reached a peak and then declined rapidly before leveling off and leveling off. The information reflected by these two data is in line with the basic rules of a game. People tend to be interested in and fascinated by this game due to its freshness and popularity, which brings great traffic and great growth of players to this game. But over time, the novelty of the game wore off, and a summary of the best game strategies for the game modes made the game less fun and harder to play. The number of players began to drop to a stable range.

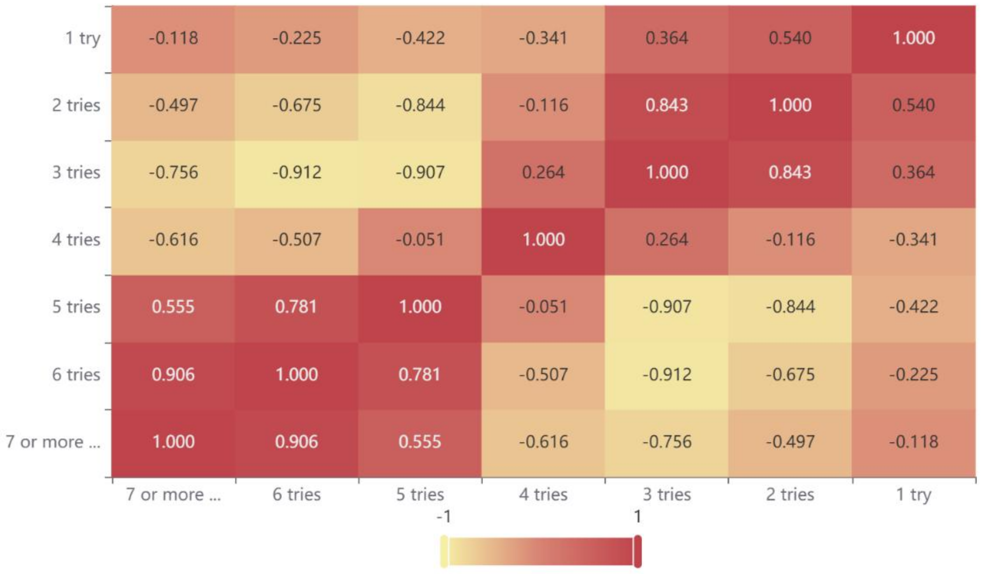


1. As can be seen from the parameters of the function fitting this proportion fluctuation, the proportion of players playing the hard mode to the total number of players playing the game on that day increases gradually over time, but the rate of increase decreases gradually. This is not hard to explain, as the strategy of the game is summed up, and people tend to choose more difficult modes. As for why this rate of increase is gradually slowing down, a reasonable explanation is that as the number of sessions increases, players lose the fun and freshness of the game in normal mode

In the case of the game, more players chose not to play and gave up on the game than to challenge the hard mode. So the ratio is going to curve like this.



2. The correlation coefficient between the percentage of The Times of completing the guessing game.



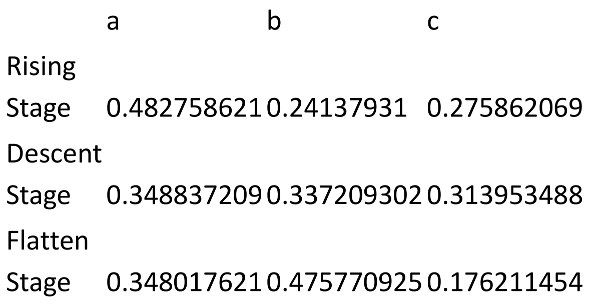
	7 or more ...	6 tries	5 tries	4 tries	3 tries	2 tries	1 try
1 try	-0.118	-0.225	-0.422	-0.341	0.364	0.540	1.000
2 tries	-0.497	-0.675	-0.844	-0.116	0.843	1.000	0.540
3 tries	-0.756	-0.912	-0.907	0.264	1.000	0.843	0.364
4 tries	-0.616	-0.507	-0.051	1.000	0.264	-0.116	-0.341
5 tries	0.555	0.781	1.000	-0.051	-0.907	-0.844	-0.422
6 tries	0.906	1.000	0.781	-0.507	-0.912	-0.675	-0.225
7 or more ...	1.000	0.906	0.555	-0.616	-0.756	-0.497	-0.118

It is convenient to see from the heat map of correlation coefficient that the percentage of 1~3 guesses is positively correlated, and the percentage of 5~X guesses is also positively correlated. The correlation between 1 to 3 and 4 to X is basically negative.

For this feature is also relatively easy to understand, if the problem is easy to guess, then the number of times must be less, more people guessed twice, three times will be more people. The same is true if the problem is hard to guess. We can call 1 to 3 the easy interval, 4 to X the hard interval, and the sum of these two intervals is fixed, so obviously they are negatively correlated.

The percentage of successful guessing attempts for each of the three difficulty types

We classified the words by difficulty and wanted to find out how many attempts it took for players to guess the word with a high probability at different word difficulties. To do this, we averaged the seven attempts on each of the three difficulties and found something interesting. As can be seen from the table, the majority of players were able to guess the word correctly on the fifth try when the word difficulty was the hardest of the day. Most players were able to guess the word correctly on the fourth try when the word difficulty was medium. On the easy word of the day, most players would also need four tries to guess the word correctly. This phenomenon shows that although we use accurate classification models to classify the difficulty of words, due to the probability of force majeure, most players need three trials and errors to narrow down the possible range of correct words. Ensure the playability of the game, accurately control the word difficulty limit.



	a	b	c
Rising Stage	0.482758621	0.24137931	0.275862069
Descent Stage	0.348837209	0.337209302	0.313953488
Flatten Stage	0.348017621	0.475770925	0.176211454

The proportion of three types of difficulty levels in the game varies during the period of rising, falling, and stable participation. Specifically:

1.The proportion of easy questions decreases from the rising period to the falling period, and remains relatively constant in the falling and stable periods.
2.The proportion of medium questions increases in all three periods.
3.The proportion of difficult questions first increases and then decreases in all three periods, with a low proportion of difficult questions in the stable period.
4.The proportion of the three types of questions is relatively balanced in the falling period.

Initially, when the game was first launched, the proportion of easy questions was the highest, which may have been a strategy employed by the New York Times to quickly attract readers and let them experience the fun of the game. However, as players became more familiar with the game strategy, their interest gradually declined. Therefore, in the falling period, the proportion of easy questions decreased, and the proportion of medium and difficult questions increased to retain players. However, in the stable period, increasing the proportion of difficult questions may further lead to player attrition. Thus, the proportion of medium questions was further increased, while the proportion of difficult questions decreased, and the proportion of easy questions remained constant.

Category	Mean 1	Mean 2	Mean 3	Mean 4	Mean 5	Mean 6	Mean 7
Simple	0.81	9.49	30.78	33.62	17.77	6.43	1.08
Normal	0.26	4.06	20.39	35.65	26.30	11.39	1.93
Hard	0.29	2.88	12.81	25.99	28.86	21.33	7.78

IX. Reference

[1] Scientific Platform Serving for Statistics Professional 2021.SPSSPRO. (Version 1.0.11)[Online Application Software].Retrieved from https://www.spsspro.com.

[2] Wang Yan.Applied time series analysis [M].Beijing: China Renmin University Press 2005.

[3] Deng Julong. Grey Prediction and Grey Decision [M]. Wuhan: Huazhong University of Science and Technology Press,2002.

[4] Zhou Zhihua. Machine Learning [M]. Tsinghua University Press, 2016.

[5] Liu Yue, Hao Shuxin, Song Jie, et al. Study on Rapid cleaning and Automatic Classification and summary of outpatient and emergency data in Time series analysis of the relationship between air pollution and disease [J]. Health Research

X. Letter to New York Times

Dear The New York Times,

I am a reader of The New York Times and have recently become interested in the WORDLE game that the company has launched. I have conducted mathematical modeling and data analysis on the related data and obtained some results.

I have developed three models to analyze this game: the "game participation prediction model," the "game score percentage distribution prediction model," and the "word difficulty classification model." I have identified word attributes that affect the difficulty of the questions, including word frequency, letter frequency, and degree of common use. I have also developed a model that evaluates word difficulty based on the player's score percentage and found a method to predict player scores based on word attributes. I have categorized words into three types: easy, normal, and difficult.

According to my analysis, the reason why game players have seen a sharp increase followed by a sharp decrease, and then a plateau is that as the number of games increases, people gradually come up with game strategies, which greatly reduces the fun and challenge of the game.I believe that The New York Times can consider the following suggestions to better integrate the WORDLE game:

Firstly, The New York Times can consider increasing the diversity of the WORDLE game. Perhaps by adding more game modes to increase the fun and difficulty of the game, such as a "first word only" mode or a "word exclusion" mode where players can exclude words to narrow down the choices based on real-world hot topics, daily news reports, or holiday-specific articles. This can increase the difficulty of the game while also giving it variety, bringing new vitality to the game. Alternatively, it can be given emotional color, and the words used for questions can be related to current events, cultural-related columns, etc. in The New York Times to attract more user groups. These changes can bring more challenges and fun while also making the WORDLE game more closely integrated with The New York Times, allowing users to feel the innovation and change of The New York Times.

Secondly, I suggest that The New York Times refer to my "game score percentage distribution prediction model" and word difficulty classification model to predict and evaluate the score and difficulty of the questions when selecting words. This will ensure that player scores are distributed more evenly and the game difficulty is moderate. Quantitative analysis and adjustment of the questions can be used to respond to fluctuations in the number of players and retain more traffic.

Additionally, I suggest that The New York Times consider integrating the WORDLE game with other content and services, such as connecting the game with English language learning, current events, blogs, social media, and other elements. This approach can bring more user interaction and participation while also increasing the brand value and user loyalty of The New York Times.

Finally, thank you for your time and attention. I hope these suggestions can bring some inspiration and value to The New York Times' WORDLE game. If you have any questions or suggestions about my recommendations or model analysis, please feel free to contact me.

Best regards,

A reader of The New York Times