Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2321423

Prediction And Classification Models For Word Puzzles: Take Wordle As Example

As the computer games prevail and sink rapidly, enduring appeal becomes a vital issue. **Wordle**—an once-popular online game offered by the New York Times, serves as an instance to explore the key. Based on the game mode and popularity trend of Wordle, the aim of this report is to analyze the necessity and provide measures for maintaining numerous players. We are expected to provide a universal forecast model of players and estimate their performance to adjust game difficulty as incentives. Therefore, three models are established: **Model I: Player Number Prediction Model; Model II: The Quantification of Words' Attributes; Model III: RF-PSO Regression for score estimation.**

For Model I, the daily number of participants from 2022/1/7 to 2022/12/31 is firstly collected. Then, based on the Z-Score Truncation of the data, the validation of using **ARIMA** model to predict the player's number is proved. Next, historical data is used to attain the parameters of ARIMA, with introduction of ACF and PACF functions to identify the final prediction model as ARIMA(0,1,1). Finally, according to ARIMA(0,1,1), an iterative program is used to simulate the prediction interval. Robustness test and the Goodness of Fit is used to evaluate the model's performance.

For Model II, through process of data collection, we first filter out the unquantifiable attributes. Then, **the Pearson Correlation Analysis** is used to calculate the Pearson Correlation Coefficient to quantify the relationships between the Hard Mode Percentage and other quantifiable attributes(in Figure 13), which proves that the Hard Mode Percentage rarely relates to these attributes. Next, using **K-Means cluster analysis**, we classify the words into 4 categories by difficulty and rated "EERIE" as "normal" word(the second easiest). The result accords with the theory well.

For Model III, multiple latent factors determine participants' performance, and the Random Forest(RF) regression is introduced for feature selection(shown in figure 27). Then, to tackle with the parameter choosing problem, the Partial Swarm Optimization(PSO) is availed. Thus, **the Random Forest regression based on Particle Swarm Optimization(RF-PSO) model** is established (procedure shown in figure 17). Via stimulation, it can be estimated that the ratio of attempts to guess "EERIE" are 3.73%, 18.66%, 37.76%, 27.91%, 11.13%, 3.13%, 0.65% for try times between 1 and 7. Afterwards, we fit the predicted results with real ones in figure 18-24, and find excellent fitting degree.

In addition, this report discusses the increase trend of Wordle players and the changes of Hard-Mode chooser ratio combined with Survival Theory. With nonlinear fitting strategy and contagion spread theory, we visualize and generalize the change rules of players. The result is compliant with exponential change on initial stage and diminishes till reaching a relatively steady level.

Eventually, **robustness and sensitivity analysis** of the models are tested. For one thing, the estimated participant number invariably falls within the confidence interval with some multiple noise deviating from original data as large as 40%. For another, as unconsidered factors affect the model, the RF-PSO is capable of feature selection, which also proves the rationality and robustness of the estimation of player performance.

Keywords: Wordle, ARIMA, Word Attributes, Cluster, Random Forest

Contents

1 Introduction

1.1 Problem Background

Wordle is a daily word game[1], which attains great popularity all over the world. Every day, the netizens are greeted with a fresh word puzzle, made up of 5 letters, that can only be solved using a series of process-of-elimination clues. The color of each letter's box changes as you input different words. The specific rules are as follows:

- ◇ **Green** means the guessed letter is in the daily word and you've placed it in the right spot.
- ◇ **Yellow** means the guessed letter is in the word but you have it in the wrong position.
- ◇ **Gray** means the guessed letter isn't in the word at all.

In order to assist the developer of the game in better estimating the difficulty of each word as well as predicting the number of participants and the accuracy of guessing of a given date, we conduct the following research.

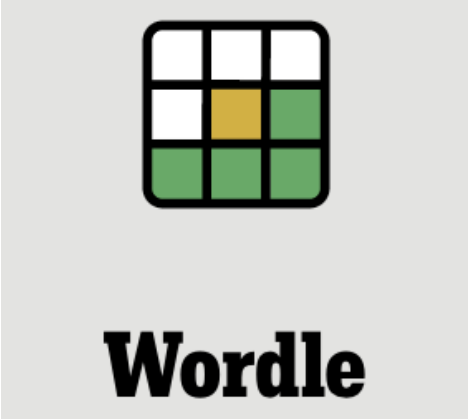

Figure 1: Wordle Game

Figure 2: Word Cloud from the Database

1.2 Restatement of the Problem

Considering the background information and given conditions identified in the problem statement, we complete the following tasks using the models that we have established.

- ◇ Develop a model to explain the variation of the number of daily reported results, then create a prediction interval for the number of reported results on a certain date.
- ◇ Figure out the relationships between the attributes of words and the percentage of scores reported that were played in Hard Mode, which is the same as the percentage of people choosing Hard Mode.
- ◇ Develop a model to predict the distribution of the reported results and explain the uncertainties associated with your model and predictions. Give an example about it and discuss its rationality.

◇ Develop a model to classify solution words by difficulty, give a example of it and discuss its accuracy.

◇ List other interesting features of the database.

1.3 Our work

In summary, the whole modeling process can be shown as follows:

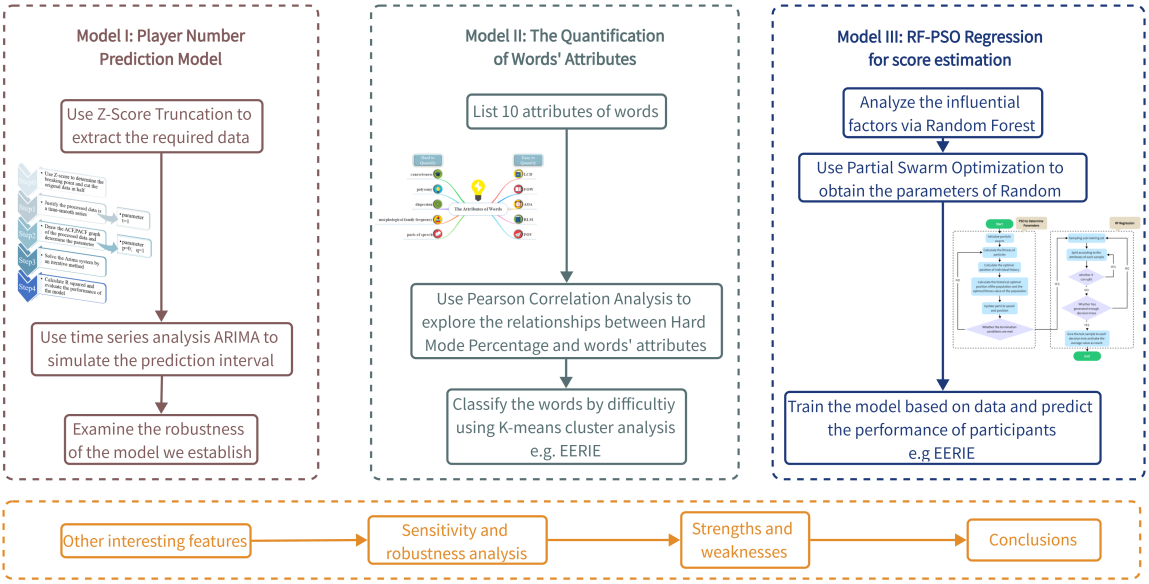

Figure 3: The Flow Chart of the Article

2 Assumptions and Justifications

To simplify the problem and in line with the reality, we make the following basic assumptions, each of which is properly justified:

◇ **Assumption 1:** The ratio of Hard Mode chooser is stable at around 0.9 after Sep.10^{th}.

⇒ **Justification:** The attraction of game elements are often short-lived and the game attract a steady players after a period of fluctuation[2]. Through observation of the Hard-Mode choosing trend, we find it basically stable at 0.9 after September 10, 2022. Thus, we can predict the performance of participants afterwards.

◇ **Assumption 2:** The number of player each day have some kind of inertia, which means we can assume that there is no sudden change in the data.

⇒ **Justification:** A person's decision to play Wordle on a given date is purely a random event. However, when large number of random events are added up, according to the theory of statistics, it can become a predictable model. Therefore, we assume that the daily number of participants has a certain inertia and will only change slowly with time in accord with a certain trend.

3 Definitions and Notations

3.1 Definitions

◇ **LCD:** The letter complexity of a certain word.

Firstly, we calculate the frequency of occurrence of each letter in each position, ranging from Position 1 to Position 5, as shown in this heatmap in Figure 4.

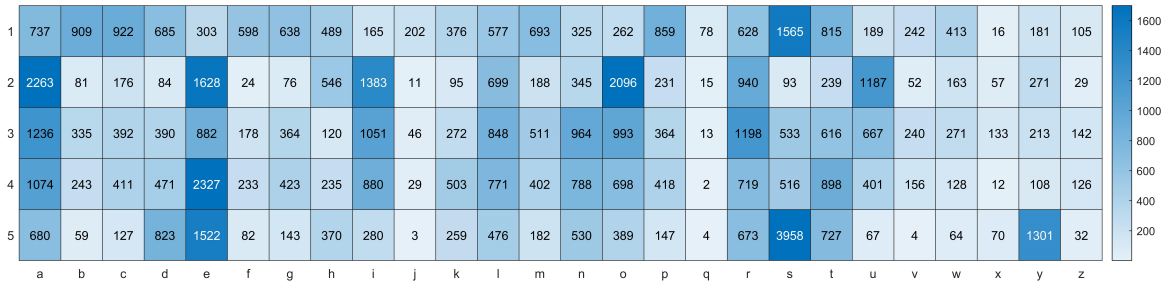

Figure 4: The Occurrence Times of Letter in Specific Location

Then, we multiply the probability of each digit to get the probability of the letter combination, which can help us measure the complexity of the word. Since the difference between the maximum value and the minimum value is several orders of magnitude, we use the logarithmic method to deal with the measurement. The formula is as follows:

$$M_{LCD} = -\log\left(\prod_{i=1}^{5} p_i\right) \tag{1}$$

where p_i represents the frequency of occurrence of each letter in $i\text{-}th$ position, M_{LCD} represents the intermediate processing value.

Then we carry out the Min-Max normalization to make the score distributed in the [0,100] interval.

$$T_{LCD} = 100 \times \frac{M_{LCD} - M_{LCD\_MIN}}{M_{LCD\_MAX} - M_{LCD\_MIN}} \tag{2}$$

where M_{LCD_MIN} and M_{LCD_MAX} represents the minimum and maximum of M_{LCD} in our lexicon of Wordle, which would be further illuminated in the following Data Section, and the symbol T_{LCD} represents the degree of letter complexity of a certain word.

◇ **FOW:** The frequency of a certain word.

Firstly, we attain the frequency f of each word in our lexicon of Wordle, whose data resources would be further illuminated in the following Data Section. Because the data is too small, we divide it by a uniform number, such as the maximum of f to expand the data. Since the difference between the maximum value and the minimum value is still several orders of magnitude, we choose to use the logarithmic method and the Min-Max normalization like above.

$$T_{\text{FOW}} = 100 \times \frac{\left[-\log\left(\frac{f}{f_{\max}}\right)\right]}{\left[-\log\left(\frac{f_{\min}}{f_{\max}}\right)\right]} \tag{3}$$

where the symbol T_{FOW} measures the frequency of a certain word.

◇ **AOA:** The age of acquisition of a certain word[3].

Since the data has the same order of magnitude, we only need to use the Min-Max normalization to standardize it.

$$T_{AOA} = 100 \times \frac{O_{AOA} - O_{AOA\_MIN}}{O_{AOA\_MAX} - O_{AOA\_MIN}} \tag{4}$$

where O_{AOA} represents the original value of AOA, the symbol T_{AOA} represents the degree of the age of acquisition of the word.

◇ **RLM:** The maximum number of repetitive letters in a certain word.

Literally, we define the symbol T_{RLM} to measure the letter attribute of the certain word.

$$T_{RLM} = 100 \times \frac{N_{RLM}}{N_A} \tag{5}$$

where N_{RLM}, N_A respectively represents the maximum number of repetitive letters and the total number of letters in a certain word.

◇ **POV:** The proportion of vowels of a certain word.

we define the symbol T_{POV} to measure the vowel attribute of the certain word.

$$T_{\text{POV}} = 100 \times \frac{N_V}{N_A} \tag{6}$$

where N_V, N_A respectively represents the number of vowels and the total number of letters in a certain word.

3.2 Notations

Notations are shown in Table1.

Table 1: Notations used in this thesis

Symbol	Definition
T_{LCD}	the degree of letter complexity of a certain word
T_{FOW}	the frequency of the word
T_{AOA}	the age of acquisition of the word
T_{RLM}	the maximum number of repetitive letters in a word
T_{POV}	the proportion of vowels of the word

4 Data

4.1 Data Resources

Data	Resource
The lexicon of Wordle	https://github.com/benton- anderson/wordle-opt
The AOA of a given word	http://crr.ugent.be/archives/806
FOW	https://reference.wolfram.com/language/ref/WordFrequencyData.html

4.2 Data Cleaning

For the purpose of better analyzing the data in later studies, data cleaning is done. After searching through all the data, we find 6 bad data points. Among all the six points, five points have spelling mistakes which cause the word to have more or fewer than five letters. For these five points, we simply correct the spelling mistakes(e.g. rprobe->probe) and retain the other data. The remaining bad point of the six is the word "study" (contest number: 529), this point has the number of reported results far fewer (roughly 1/10) than that of the day before or the day after, yet the number in hard mode remains approximately the same. All these indicates that there is a statistical error on 2022/11/30. Our solution is to delete this row of information and supplement it using the Mean difference method (illustrated in the following picture).

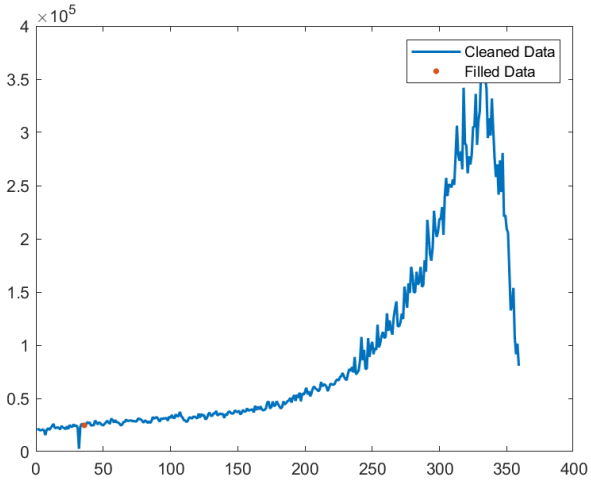

Figure 5: Data Cleaning

where the red point indicates the filled data.

5 Model I: Player Number Prediction Model

5.1 General Ideas and Results Display

In order to better predict the variation of the number of Wordle participants from time to time, we naturally introduced the ***ARIMA model***. But before we do so,we need to check whether the data meet the conditions of applicability of the ARIMA model, which mainly requires the data to be in a wide smooth state.

Applicability Verification.In order to check whether the data meets the applicability conditions of the ARIMA model, we need to do the first-order difference of the data. Then, we find that the first half of the data fluctuates a lot, while the second half of the data fluctuates very little.

Reason Analysis.Based on our observations, the data suggests that the number of visitors per day in the first half of the data increased and varied significantly from day to day, presumably due to the fact that Wordle was known through various channels after its launch. Eventually the increase in visitors brought by this information diffusion diminishes over time (due to a limited total number of people). At this point, another effect starts to gradually replace the former and takes over, leading to a decrease in the number of daily participants in the second half of the data. We believe that this may be the forgetting effect of the game, which leads to a gradual decline in the number of daily participants from some point near 2022/6/1, eventually reaching a stable value (those who remain are die-hard fans).

To predict the number of participants in the next 60 days, we only need to consider the forgetting effect and ignore the propagation effect in the first half of the period. So removing the first half of the data does not have a significant impact on the prediction.

Z-Score standardization. We use the Z-Score standardization method[2] to remove the first half of the data that do not conform to the wide smoothing, and the specific formula is as follows:

$$Z\text{-}\boldsymbol{Score} = \frac{x_i - \bar{x}}{s} \tag{7}$$

where $\bar{x}$ is the average of all x and s is the standard deviation of all x

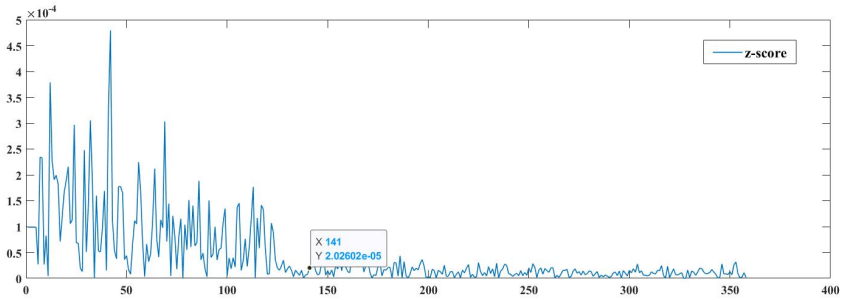

Figure 6: Z-Score

With the Z-Score graph, we can clearly and intuitively derive the inflection point of the image (we set the point to be 2022/6/6), which represents the threshold where the two effects (information diffusion and forgetting effects) are at odds with each other. In turn, the data can be split relatively easily.

After we deferentially split the processed data, hypothesis testing can be used to determine whether the series is a time-smooth series. The original hypothesis is that the series is an unstable time series(taking a significance level of 0.001).

At the difference of order 1, the significance p_value is 0.000, thus the original hypothesis is rejected. So we have justify that the series is indeed a smooth time series, satisfying the conditions for the applicability of the ARIMA model. Thus the formula below is established:[3][4]

$$y_t = \mu + \sum_{i=1}^{p} r_i y_{t-1} + \varepsilon_t + \sum_{i=1}^{Q_1} \theta_i \varepsilon_{t-1} \tag{8}$$

where p suggest the relation between y_t and the historic data y_i while q represents the relation between y_t and the error term ε.

All we need to now before applying the Arima model is to determine the parameters p, q according to the their ACF and PACF graph.[7]

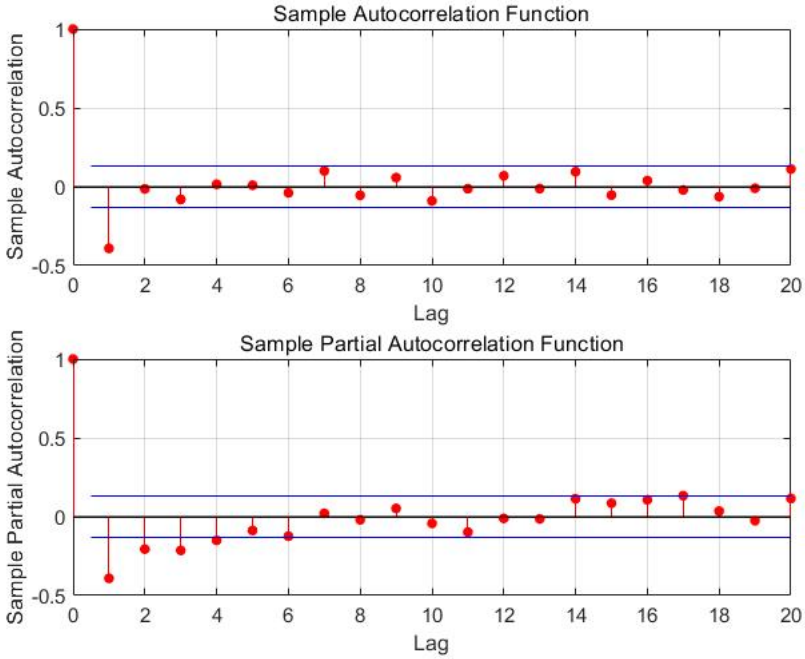

Figure 7: ACF&PACF

Based on the image we can find that the auto correlation and partial auto correlation plots drawn from the data are trailing. The most significant order in the PACF, ACF plot is 0&1. So we can know that p=1 & q=1.[7]

Then we use an iterative method to solve for the prediction interval of the participants' number on $Mar.1^{st}$(the forecasting error of each day is set at 5%). Detailed implementation are shown in Algorithm 1.The result is shown below:

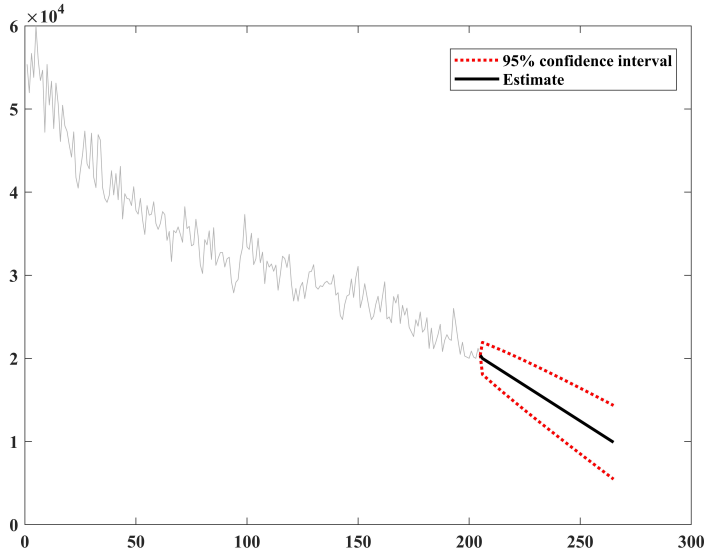

Figure 8: ARIMA_Result

The prediction interval is estimated to be [855,18956].

The final process is to evaluate the model's performance. Based on the parameters we can find that the Goodness of Fit R^2 of the model is 0.923, which indicates that the simulation is excellent and the model meets the requirements. Further ways of evaluation are shown in the graph below, all showing a satisfying result:

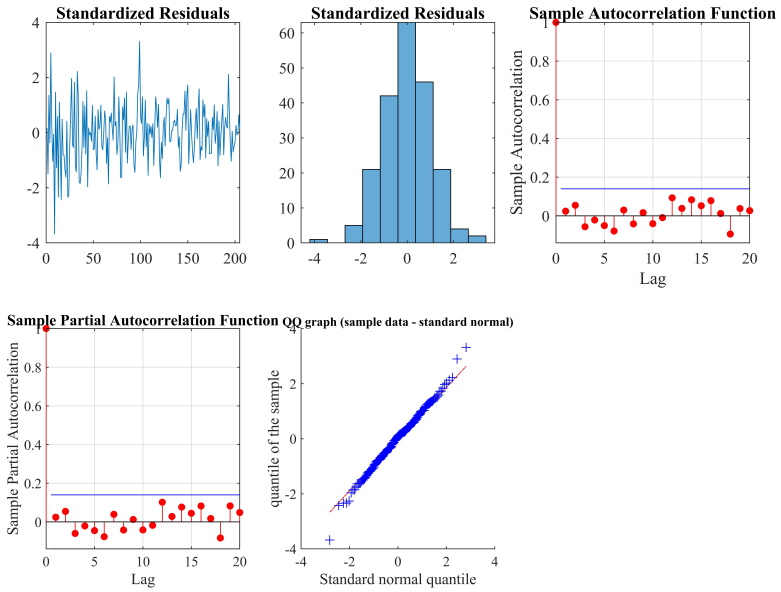

Figure 9: Model I Evaluation

We use a flow chart to summarize the operation process of model I:

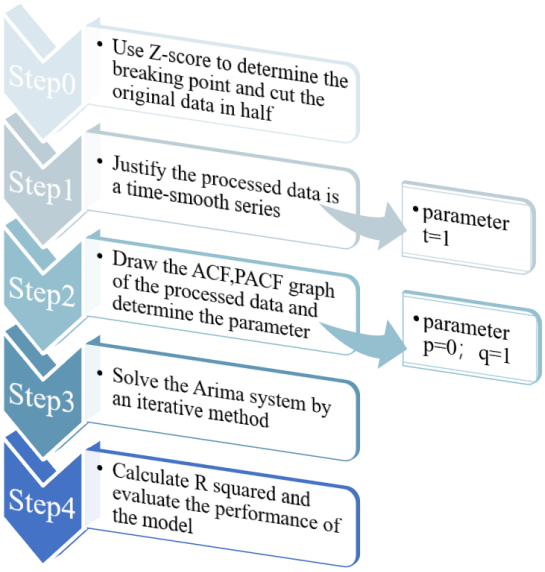

Figure 10: Flow Chart of Model I

5.2 Detailed Implementation

The main part of the interactive function is posted below to help readers better understand the iterative method used in the model(in Pseudo code).

Algorithm 1 algorithm of Model I

Input: a list of the number of the player each day num_player
the required prediction date $length$
parameter p
parameter d
parameter q
bool index max

Output: $list\_num\_pred$

```
while length! = 0 do
    length=length-1
    temp=ARIMA(list,p,d,q)*(1+0.05*max)
    #call the ARIMA function to estimate the next data and set the forecast error at 5%
    list.append(temp)
    list_num_pred.append(temp)
end while
print(list_num_pred) #draw the graph of the prediction
return list_num_pred
```

6 Model II:The Quantification of Words' Attributes

The attributes of words can have various aspects. It is widely recognized that length is an essential attribute in words, but since the length of each word is fixed at 5 in wordle, this factor can be temporarily discarded in this thesis. According to Nagy and Hiebert's study in 2011[7] and our rational analysis, eventually we intend to take these following factors into consideration: LCD, FOW, parts of speech, morphological family frequency, dispersion, RLM, POV, poly-semy, AOA and concreteness.

6.1 Two categories of the attributes of words

According to whether each factor can be easily quantified, we divide these attributes into two categories: Attributes that are easy to quantify and that are challenging to quantify.

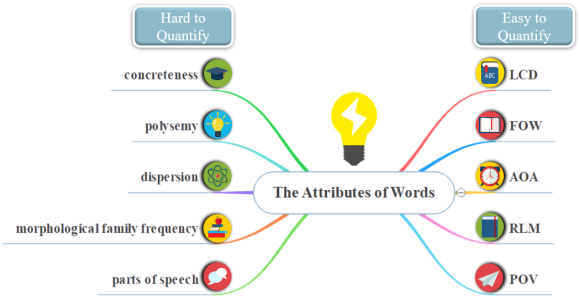

Figure 11: The Attributes of Words

Easy to quantify

As the attributes mentioned in the Definitions and Notations above, namely the LCD, FOW, AOA, RLM and POV, we can quantify them with ease, and explore whether they will affect the percentage of scores reported that were played in Hard Mode. We will perform this operation later.

Hard to quantify

On the one hand, it is difficult to obtain accurate data about these factors; on the other hand, it is hard to quantify these factors. To simplify the problem, we do not take these fac-tors into consideration in this problem.

6.2 Relationships between Hard Mode Percentage and Quantifiable Factors ——Based on Pearson Correlation Analysis

Based on the Z-Score standardization method that is mentioned in Model I, we simply consider the data after June 6, 2022 to figure out the relationships between Hard Mode Percentage and the quantifiable factors.

Data Pre-Filtering

According to box diagram below, we can conclude that the attributes——LCD, POW, AOA approximately satisfy the normal distribution while the other two attributes——POV, RLM don't satisfy.

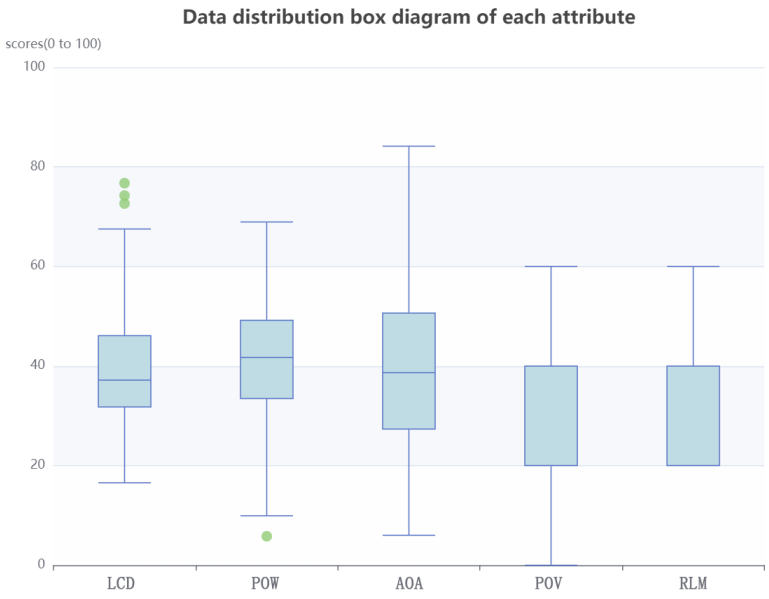

Figure 12: Data Distribution Diagram of Each Attributes

Pearson Correlation Analysis

First of all, we only consider 3 attributes of words—— LCD, POW, AOA because Pearson correlation analysis can only analyze the relationship between two variables that follow the normal distribution. Here is the formula for the Pearson correlation coefficient.

$$r = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} \sqrt{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}} \tag{9}$$

where (X_i, Y_i) is a set of sample points, $\bar{X}$ and $\bar{Y}$ are the sample means.

Using this powerful formula, we make the following heat map of correlation coefficients.

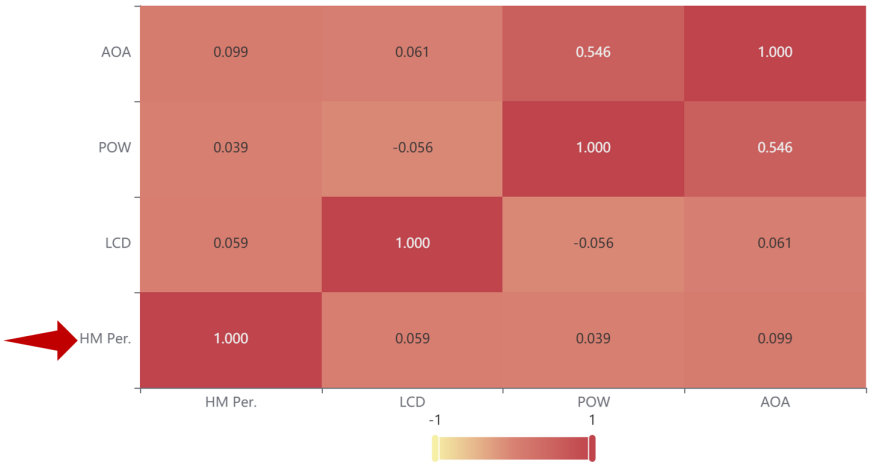

Figure 13: The Pearson Correlation Coefficient Among 4 Factors

We mainly focus on the last line of data, which reflects the correlation coefficient be-tween the three factors and the percentage. We find that the correlation coefficients between Hard Mode Percentage and LCD, POW, AOA are all less than 0.01, which implies that they have almost no linear relationship.

Result Analysis

In fact, it is very reasonable to conclude that Hard Mode Percentage has nothing to do with LCD, POW and AOA. When game players open Wordle, they do not know the attrib-utes associated with this word but habitually turn difficult mode on or off. Or someone among their friends is playing hard mode, so they intend to try it too.

6.3 Classifications Based on Difficulty

In this section, we introduce the method of cluster. We first rate the performance of the player daily using the formula $S_p = \log(400i_1 + 100i_2 + 7i_3 + 6i_4 + i_5)$ where S_p represents the score of the players performance while i_n represent the percentage of people who gets their answer in their nth try.

The coefficient of each i_n is set exponentially and we take the log of the sum.This formula is mainly based on the idea of information entropy. We observed that the first few guesses give us much more information than the others, which help us rule out many possibilities. So success in the first few tries is far harder, thus should be given far more scores than the other(success after more than fifth try is negligible regarding the rating of performance).

Then we use the cluster method to process the players data ,the results are shown below on the left.

The score of EERIE is 3.5995, referring to the estimated performance processed by the formula above. Through cluster method, it is classified to category 3, meaning “simple” (the second easiest level).

To test the viability of the attribute of words mentioned above. We also process the word into cluster according to their FOW and LCD. The results are shown below on the right.

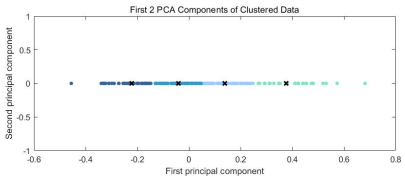

Figure 14: Classification by Difficulty

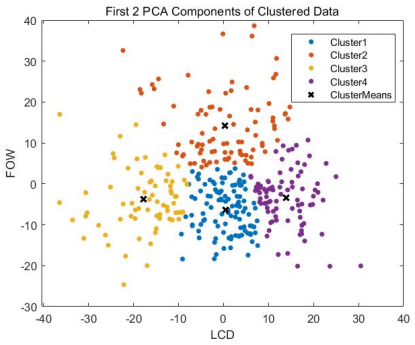

Figure 15: Classification by FOW and LCD

By analyzing two figures, we observe that the yellow cluster and the blue cluster in figure2 is mostly correlated with the most difficult and the second most difficult cluster in figure1, which indicate that the word with low frequency and extreme LCD(either extremely low or extremely high) are difficult to guess. This is in line with our intuition thus proving the rationality of the model.

7 Model III: RF-PSO Regression for Score Estimation

7.1 Model Choice and Optimization Based on Data Features

The number of attempts before success is relevant to numerous factors in reality. For one thing, word attributes such as LCD, FOW, AOA(which mentioned in 4.3) affects whether the solution easily occur to people. For another, the surge of players, the play-in-hard-mode ratio, and proficiency improved day by day are taken in account as well. Nonetheless, emergency elements such as the wide use of AI to solve the problem is not considered, though it may make waves. The relevance map is showed below:

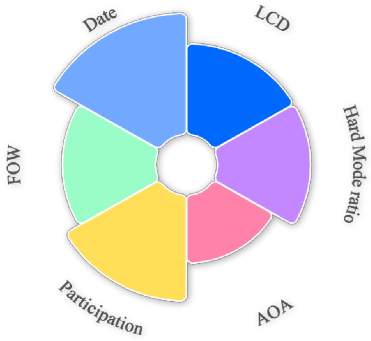

Figure 16: Factors Included

Multi-factor, small amount of data, and certain reality randomness are core characters of the problem as seen above, distinguishing among masses of existing non-sequential prediction models(shown below). Judging between the merits and shortcomings of those algorithms, the combination of the Random Forest(RF) regression and the Partial Swarm Optimization(PSO) are availed to adapt to these features for optimal prediction.

The Random Forest(RF) regression is an ensemble technique that combines multiple disconnected decision trees and the ultimate decision is made jointly. The key advantage of it is the capacity to handle high-dimensional datasets with many features without feature selection(we call it Auto Feature Choice), which is the second-to-none bet to cope with multi-factor, as it did in the case, considering 5 elements above. Additionally, since build multiple decision trees is demanded, it requires massive memory and computation resources for large data. Thus, since the case with appropriate number of data, the RF model can do the prediction excellently.

Otherwise, to amend the weak robustness of RF and solve the discrete data without gradient information, we introduce PSO as a compliment for RF. It simulating the predation behavior of birds, cooperating to get the optimal solution. This message-exchange procedure improves the efficiency of data utilization and the stability of dealing with noise interference, since the overall results of Wordle may differs in virtue of some random events.

The RF-PSO model is established for attempt times prediction, based on the analysis above. In our proposed RF-PSO based algorithm, the best parameters in the random forest model is gained by the PSO procedure.

7.2 Structure of RF-PSO

As is known, the first step for RF algorithm is to determine the parameters. Inspired by an article in IEEE conference [8], we utilize the PSO algorithm to find the best parameter set in the random forest model among the possible parameter sets. The update process of the particle i about its own speed and position is based on the following formula:

$$V_i = w^* V_i + {c_1}^* \operatorname{rand}()^* \left(p\,\mathrm{Best}[i] - X_i\right) + {c_2}^* \operatorname{Rand}()^* \left(p\,\mathrm{Best}[g] - X_i\right) \tag{10}$$

$$X_i = X_i + V_i \tag{11}$$

Below are explanations about these significant parameters for PSO model and gives the assignments of each one in the actual case of Wordle results prediction:

- $\diamond$ w is the inertia weight, which depict how the velocity of the latter generation affects that of this generation. To balance local and global optimal in the realistic occasion, w is taken as 0.9.
- $\diamond$ $c1$ and $c2$ are constant, which are called learning factor. We take both C1 and C2 as 2 to avoid missing the target or being stuck in local optimal value.
- $\diamond$ N is the initial number of particles. The smaller N is, the easier it is to converge, but it is easy to fall into the local optimal solution. So we let $N = 50$ to ensure the efficiency and accuracy.
- $\diamond$ V_{max} is the highest velocity to determine the maximal times of iterations, and the current generation speed is expressed in V_i. We assign V_{max} as 150 to ensure convergence.
- $\diamond$ X_i is just represented as a RF parameter vector and $pBest[i]$ is the best parameter calculated in Wordle prediction.
- $\diamond$ $Rand()$ and $rand()$ are the cable number of the position of "the best" particle in the group which randomly changes from 0 to 1.

Afterwards, our program generate a corresponding number of decision trees according to the parameters obtained in the PSO process, and determine the node splitting rules and the size of the tree. (As it is widely used, we will not repeat the concrete principle and train procedure of RF [8]) It shuffles the given data and use the 70% of it to train the model, while the other 30% to test the accuracy of the established model.

The whole process of RF-PSO is shown in the figure below:

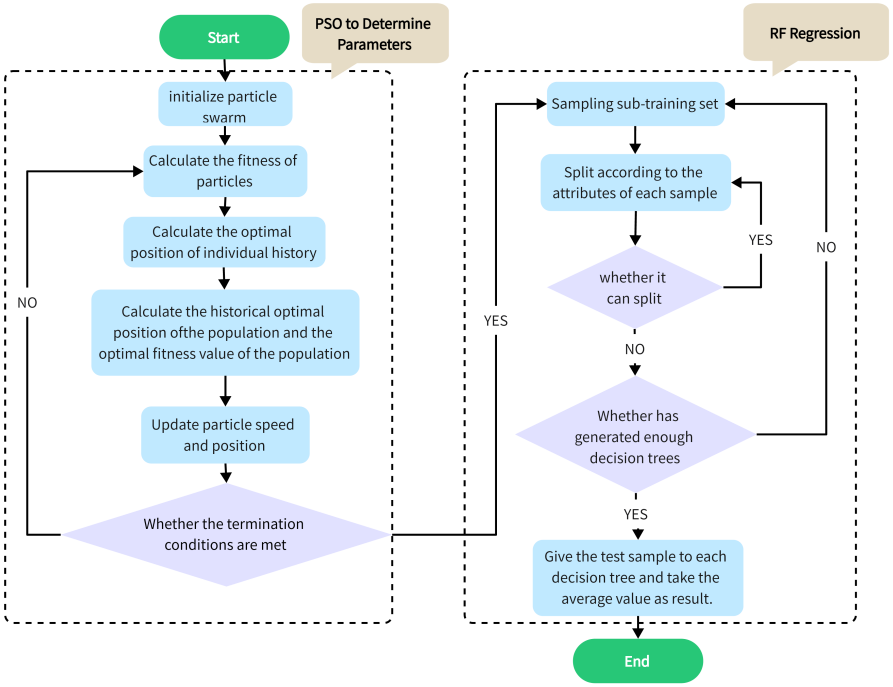

Figure 17: RF-PSO Process

7.3 Estimate Results of RF-PSO

In this application of this model, we use the data cleaned and processed in former parts. Before specific modeling, we classify these data as dependent variable and independent variable. The prediction target is the percentage of each attempts times, which means the dependent variable of RF-PSO. And to avail the Auto Feature Choice of RT, other influential factors are all considered, including Date, the ratio of Hard Mode, the number of reported results, LCD, FOW, AOA. Especially, the percentage prediction of latter times with take the former result in account. In other words, when predict the percentage of success with 3 tries, the RT-program include the predicted percentage of 1 tries and 2 tries as an independent variable. The factors considered is shown in the figure below:

After input these shuffled data to train the RF-PSO model, we get the trained Model. Then, we input the data of EERIE (In particular, since the proportion of people who choose Hard Mode is basically stable at around 0.09 after 2022/12/01, the value of 2022/03/01 is also assumed to be 0.09), and get the predicted results. From the trained RF-PSO model, the predicted results are listed in table 2.

word	1 tries	2 tries	3 tries	4 tries	5 tries	6 tries	7(X) tries
eerie	3.73	18.66	37.76	27.91	11.13	13.13	0.65

Table 2: result prediction of EERIE

7.4 Accuracy Test

By extension, to know the accuracy of our model, we use multiple measures to test the fitting degree of our trained RF-PSO model. The below table shows the prediction and evaluation results of

the training set and test set, and measures the prediction effect of the trained RF-PSO model through a series of quantitative indicators.

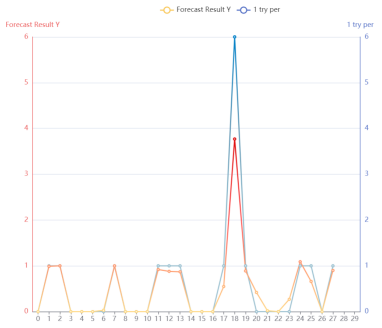

Figure 18: 1 try

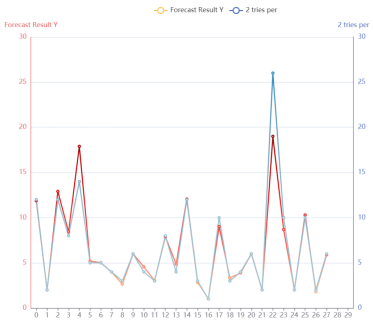

Figure 19: 2 tries

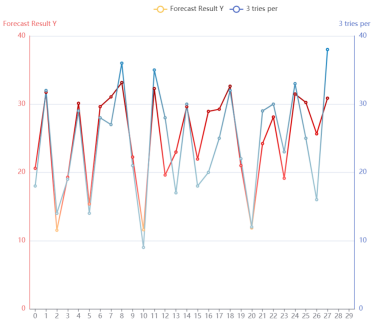

Figure 20: 3 tries

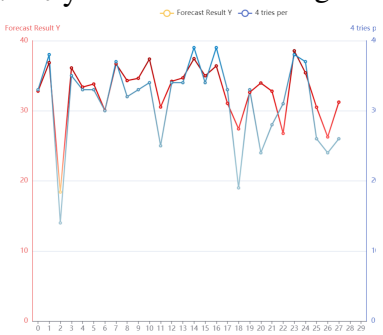

Figure 21: 4 tries

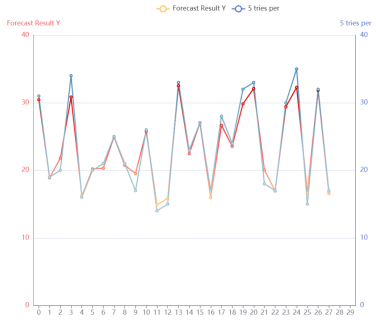

Figure 22: 5 tries

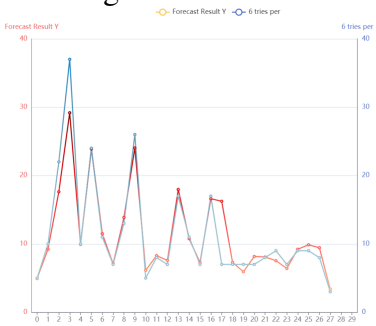

Figure 23: 6 tries

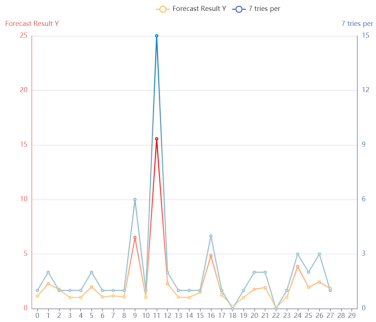

Figure 24: 7 tries or more

From the evaluation results for the prediction of each times, we find that, under the training, the RF-PSO model has a good fitting degree for the probability prediction of each number of times used to guess the right answer, and has a good prediction effect.

In order to more intuitively feel the fitting degree of the model, we will make the following curve for the true value and predicted value of the test set after the disruption, and we can see that the prediction result is accurate.

8 Other Interesting Features

8.1 The Propagation Effect of Wordle

In the first model we have studied the forgetting effect of the model. It is also interesting to investigate the first half of the data to have a better picture of how this game is spread worldwide. After going through the literature ,we find that modeling information propagation by the survival model[9]can best suit our need.

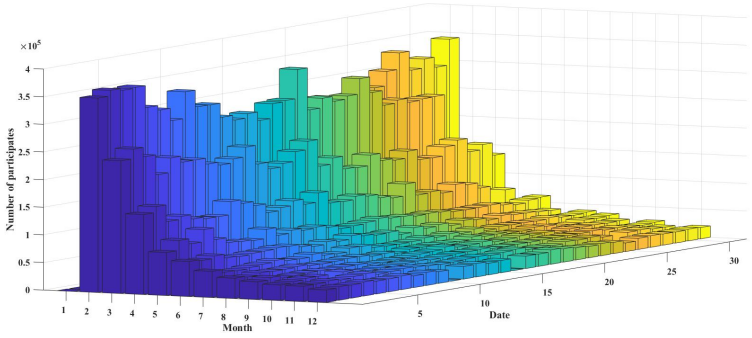

Figure 25: participant vs time

Inspired by previous literature,our approach is to develop a nonlinear model to explain the first half of the data,namely the information propagation model. In the model we consider information of the game wordle as contagions and potential player of wordle as a node. Thus,the whole picture of information propagation becomes contagions spreading across a fixed population of nodes. The contagion spreads by nodes forcing other nodes to switch from being uninfected to being infected, but nodes cannot switch in the opposite direction. Therefore, we can represent whether a node is infected at any given time as a nondecreasing (binary) counting process. We then model the instantaneous risk of infection,simply put in our case,is just how likely you will go to play wordle once you hear a person talking about this game. By inferring which nodes influence the hazard rate of a given node, we discover the edges the underlying network over which propagation takes place. In particular, if the hazard rate of node i depends on the infection time of node j, then there is a directed edge (j, i) in the underlying network.

By training the model with the existing data, we get to know the approximate value of i&j, which draws a blueprint for future studies.

8.2 Trend of HM Choose Rate Over Time

The trend of Hard Mode choosing rate is another case worth studying.To begin with, we first visualize the rate on a time scale. There is a graph below showing that change

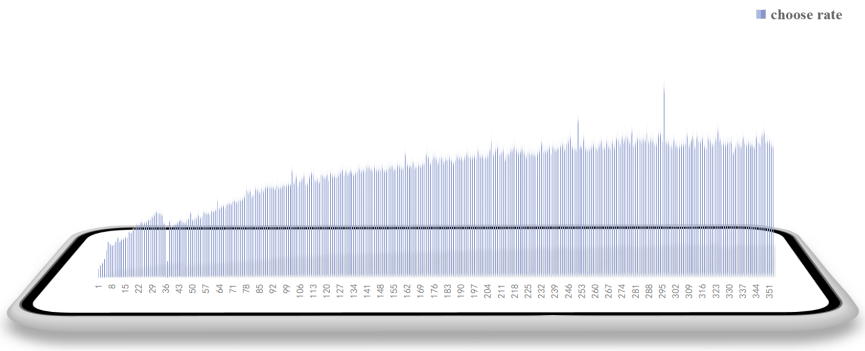

Figure 26: Hard Mode Rate vs Time

Two possible reasons are that who remained are all experienced (indicated by the drop in numbers of the participants) or people prefer the Hard Mode since the result looks good on their TikTok or Instagram page.

9 Sensitivity and Robustness Analysis

9.1 Sensitivity: RF-PSO's Comprehensive Feature Selection

RF algorithm is considered as a insensitive algorithm to noise and outliers in the input data, and so does RF-PSO. This is because it is based on a collection of swarms and decision trees, and each of them is trained on a random subset of the features and data, making it more immune to noise and overfitting. In other words, the RF-PSO model can handle a certain degree of noise and outliers in the input data without significantly affecting its performance.

However, the sensitivity of the RF-PSO can still vary depending on the type and degree of noise in the input data, since the significant ratio of factors may vary with time. The ratio of attempts before correct answer can be dominated by the number of people at first, but greatly influenced by proficiency afterwards when the participants is enough. In these cases, RF-PSO will always judge the importance ratio of each feature and keep pace with the current situation. The uncounted feature selection is listed in figure [***].

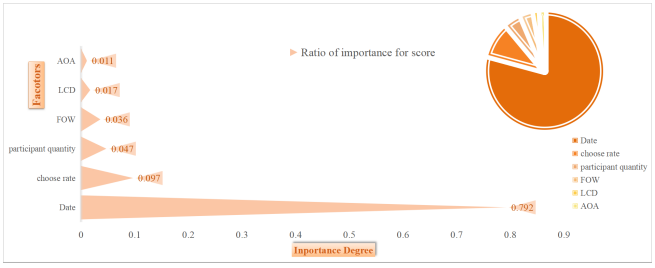

Figure 27: Sensitivity Test & Feature Selection

Overall, while the random forest algorithm is relatively insensitive to noise and outliers, it is still aware of the type and degree of noise in the input data and selects the relevant features accordingly to

get the best results.

9.2 Robustness

We have tested the robustness of the ARIMA model presented in section 1.Some noise were added based on the officially provided data. We preset the variance of noise to be 0.1 and 0.4 times the original value.The noise added here follows a Gaussian distribution. We have visualized the noise in the following graph.

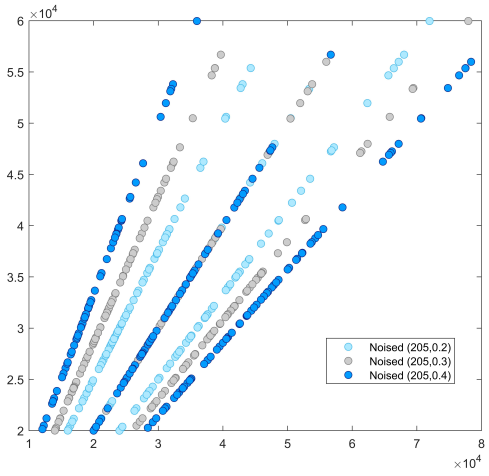

Figure 28: Noise Visualization

where the y axis is the original data and the x axis is the result when noise is added. We then put the processed data back to our program.The following Figure shows the simulation results.

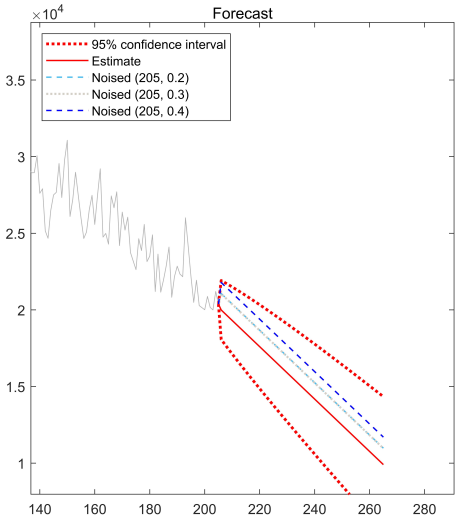

Figure 29: Robustness Test

We can know clearly from the figure that even if we add a maximum of 40% noise to all of the original data,the probability that the number of participants on 2023/3/1 will still most likely fall in our prediction interval(95% likely).Thus we ensure the excellent robustness of our study.

10 Model Evaluation and Further Discussion

10.1 Strengths

◇ ***Multi-dimensional evaluation.*** We have considered a lot of attributes of words to establish a multi-dimensional evaluation system. he split procedure of RT trees satisfies a significant demand in this problem with a lot of variables.

◇ ***Great robustness.*** The sensitivity analysis of the model demonstrates the effectiveness of the model under different parameter combinations and prove the robustness of the model.

◇ ***High-level Accuracy.*** For one thing, the RF model is accurate with part of lost features, which is tailored for the prediction since some factors can not be quantified. For another, PSO solves the parameters determination problem of RF. The combination goes a further step of accuracy.

10.2 Weaknesses

◇ ***Unable to formulate prediction function.*** Machine learning plays like a black box, obtaining the result without formula or knowable relationship. That makes it hard for users to adjust the variables.

◇ ***Poor data set for training.*** The data from Wordle contains only around 500 words with a few errors. The lack of data may affect the accuracy of the trained model.

11 Conclusions

In the article, we have established three models to solve the problems: Model I: Player Number Prediction Model; Model II: The Quantification of Words' Attributes; Model III: RF-PSO Regression for score estimation.

◇ **Model I: Interval Prediction through ARIMA.** Through the model of Z-score and ARIMA combined, we calculate the prediction interval at Mar 1, 2023 to be $[855, 18956]$.

◇ **Model II: Words' Attributes and Classifications by Difficulty.** We find that the Hard Mode Percentage rarely relates to the quantifiable attributes of words, namely the LCD, POW and AOA within the scope of consideration in this article. Using K-Means cluster analysis, we classify the words into 4 categories. And the word "EERIE" is classified to category 3, meaning "simple" (the second easiest level).

◇ **Model III: Estimated Result of "EERIE".** Through PF-PSO model, the ratio of attempts used to guess "EERIE" are respectively 3.73%, 18.66%, 37.76%, 27.91%, 11.13%, 3.13%, 0.65%

for 1,2,...,6 and 7 or more tries. And the model has an excellent fitting degree.(shown in figure 18-24)

- **Other Features: Survival Theory.** Through Survival Model, we provide a method to study the nonlinear part of the data, pointing the direction for future studies.

References

[1] "Wordle-The New York Times." The New York Times, 2022. Accessed December 13, 2022 at https://www.nytimes.com/games/wordle/index.html.

[2] Suh, A., Cheung, C., Ahuja, M., & Wagner, C. (2017). Gamification in the workplace: The central role of the aesthetic experience. Journal of Management Information Systems, 34(1), 268–305.

[3] Kuperman, V.; Stadthagen-Gonzalez, H.; Brysbaert, M. Age-of-acquisition ratings for 30,000 English words. Behav. Res. Methods 2012, 44, 978–990.

[4] The Mystery of the Z-Score Alexander E. Curtis , Tanya A. Smith , Bulat A. Ziganshin , John A. Elefteriades Thieme Medical Publishers 333 Seventh Avenue, New York, NY 10001, USA.

[5] ARIMA model building and the time series analysis approach to forecasting Paul Newbold January/March 1983

[6] M. S. Peiris, B. J. C. ON PREDICTION WITH FRACTIONALLY DIFFERENCED ARIMA MODELS Perera Journal of Time Series AnalysisVolume 9, Issue 3 p. 215-220 May 1988

[7] Nagy, W.E.; Hiebert, E.H. Toward a theory of word selection. In Handbook of Reading Research; Kamil, M.L., Pearson, P.D., Moje, E.B., Afflerbach, P.P., Eds.; Longman: New York, NY, USA, 2011; Volume 4, pp. 388–404.

[8] H. Li, W. Guo, G. Wu and Y. Li, "A RF-PSO Based Hybrid Feature Selection Model in Intrusion Detection System," 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), Guangzhou, China, 2018, pp. 795-802, doi: 10.1109/DSC.2018.00128.

[9] Modeling Information Propagation with Survival Theory Manuel Gomez-Rodriguez, Jure Leskovec, Bernhard Schölkopf Proceedings of the 30th International Conference on Machine Learning, PMLR 28(3):666-674, 2013.

To：Puzzle Editor of the New York Times

From: MCM Team 2321423
Subject: Some research about Wordle and further proposals
Date: February 21,2023

Dear Wordle Editors:

As loyal fans of the game Wordle, we appreciate the game very much, spellbound by its interesting design and challenging puzzles. We are so addicted to Wordle that we have done some research about the game. Next, we will explain our research results and give some relevant reasonable suggestions as follows:

- Firstly, we notice that the player number has dropped about 75% from 2022/5/4 till now, mainly due to the forgetting effect of the players. It is of great urgence to perform some action to optimize the game before we lose more die-hard fans, hoping to attract more new blood. Strategies like setting different challenge levels more than just the Normal Mode and the Hard Mode, may help keep players engaged over time. While Wordle is a single-player game at present, adding multiplayer options and launching online P.K. Mode could make it even more attractive.
- Secondly, we've observed that the percentage of player choosing to play in the Hard Mode shows a general upward trend. A reasonable guess is that those who continue playing are all experienced or prefer the Hard Mode, since the result looks ostentatious on their TikTok or Instagram Page. Whichever is the case, we strongly recommend adding more social functions to fascinate more people.
- Last but not least, the word chosen for the puzzle provides much fruit for thoughts. Our findings reveal that we should take comprehensive consideration of attributes of words (the frequency, the letter complexity, the age of acquisition of the word and so on) when choosing the riddle answer. The excellent choice of words should be harmonious among these factors. Therefore, the game can retain its challenging characteristic, and at the same time, the player won't be frustrated with the problem of excessive difficulty.

All in all, we are in favor of the game's design, which is concise and delightful, making it easy to focus on the task itself. The interface is user-friendly and the color scheme is friendly on users' eyes. The purpose of proposing these suggestions is not to blame the game team, but to hope that Wordle can become better and better in the near future.

We want to express our sincere appreciation for the time and effort that you and your team have put into reading this letter and creating Wordle.

Best wishes!

Yours Sincerely,
Team 2321423