Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2301434

title wordle

Summary

At the start of 2022, images of green and yellow squares began cropping up on Twitter, when the pandemic spreading wide on earth with people quarantined at home. It a flash in the pan and quickly gets off-booming in June which provides a good sample for Dataminer to find the regular pattern for multitude and the sociology behind.

Our paper firstly proposes the regression process model with ARIMA principal to address the Variation of Number of Reported Results through Twitter with Time Issue.

For the sake of analysis, we takes into account the following attributes of words: Using Frequency Globally, Typical

Keywords: ARIMA

1 Introduction

1.1 Problem Background

Wordle is a daily, online brainteaser, a single-player word-guessing game which the goal is to discover a secret word w having the same length 5 that has been chosen from a certain database D. In order to discover w, the player can make at most 6 guesses. Each time a player makes a guess, the game provides feedback in the form of colored squares that indicate whether a letter in the guess is present in the target word and in the correct position (green), present in the target word but in the wrong position (yellow), or not present in the target word (gray). The objective of the game is to guess the word correctly with as few attempts as possible, using logic and deduction to narrow down the possibilities.

Wordle has become a cultural phenomenon, with players all over the world competing to get the highest scores and share their strategies and tips for success. The game has also sparked debate and discussion around issues such as the ethics of using external tools to cheat and the impact of the game on mental health and productivity.

1.2 Literal Review

Morning Consult, a global data intelligence company, found that Wordle was one of the top ten most popular mobile games in the United States in January 2022, with an estimated 9.6 million monthly active users.[1]

However, it's important to note that these surveys were conducted early in the game's rise to popularity and may not reflect the current level of interest or engagement with the game. As Wordle is a relatively new phenomenon, there is likely to be more data and surveys on the topic in the future.

Another survey conducted by OnePoll in January 2022 found that nearly three-quarters of respondents (73%) had played Wordle at least once, and of those who had played the game, 40% reported playing it every day.[3]

Through utilization of the extant survey, we were afforded a panoramic view of its ubiquity, culminating in a refined, precise, and attainable blueprint for our architectural design.

1.3 Our Work

After processing the data to a more accurate version, we do a prediction of number of people sharing on Twitter in the upcoming few days. It is the Variation towards time model.

Furthermore, we discover a connection between Word Attributes and peoples' specific casting preferences in the hard mode.

In the fifth section, we forecast how individuals would do given a specific phrase and predefined time.

We created a Classify Words Model after extensively learning from existing models. It can give a profit of a word from the result side.

What's more intriguing is that, in contrast to the surveys mentioned above, we focused on the

interesting features of this game, from which we created new neural network learning algorithms to abstract "a changing throng" and analog the distribution of the actual number of players when playing this game, regardless of whether they tweeted their results or not. There after, We discovered a correlation between the frequency with which people solve this challenge and their desire to share it on Twitter from the Analog Gamer Mass, which is in line with psychological studies.

2 Preparation of the Models

2.1 Data Processing

After examing the provided excel, we pictured the properties and found noises inside.There are four words "naïve","tash","clen","rprobe" appeared invalid during the String Matching Process. According to the daily use practice and later model, we change them to "naive", "trash", "clean" and "probe" respectively.

To make the supplied Attempt-partition more practicable, we use 3 Sigma Outlier Identification and found the sum of all ratios happened to be far away from the rational value 1.

$$\Sigma_{i=1}^{7} t_i = 1.26$$

Owing to this spot, we manipulated

$$t_i' = 1/1.26 * t_i$$

to acquire

$$\Sigma_{i=1}^{7} t_i' = 1$$

which fits the general knowledge and the topic requirements.

2.2 Notation

Table 1: Notations

Symbol	Definition
A	the first one
b	the second one
α	the last one

2.3 Overall Flowchart

3 Variation towards time model

3.1 Problem Analysis

3.1.1 Problem Restatement

According to the subject, we clarify the requirement to two parts.

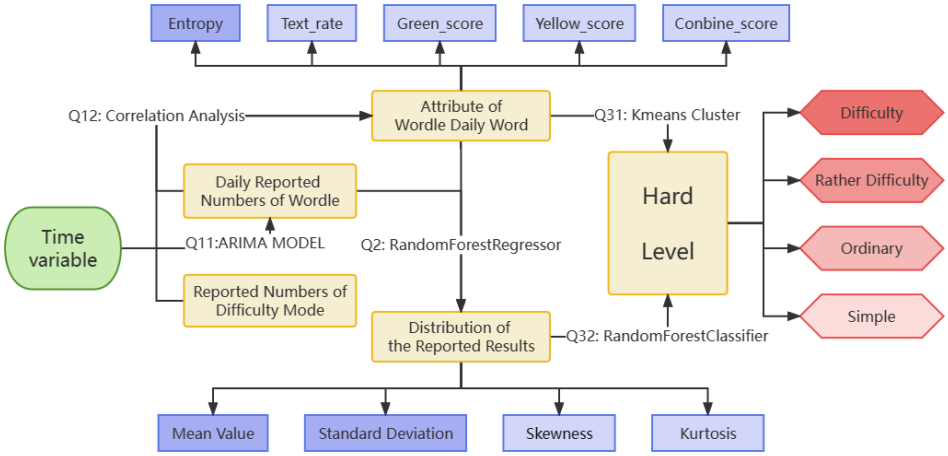
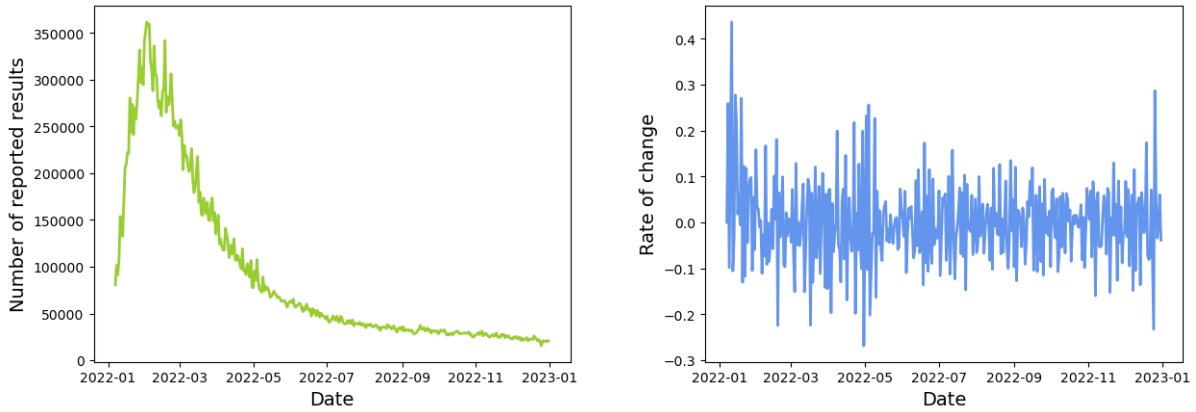


Figure 1: Overall Flowchart

- Develop the model to explain the variation (Regression Fitting)
- Use the model for predictive analysis (Forecasting)

The first problem is mainly a Time Series Analysis issue. Thereafter preprocessing the data, we found the outliers and corrected the data by approximate mean. Here is the Time Series Plot of the Original Data and a Time Series Plot of the First-order Difference Data.



(a) Time-series Diagram of First-order differencing (b) Time-series Diagram of Zero-order differencing

Figure 2: name of the figure

Since this data has no obvious periodic characteristics, there is no periodicity in the amount of data for one year. Observing that the first-order difference plot predicts the first-order difference is a smooth data, we are ready to use the traditional ARIMA model for interpretation and prediction. In order to verify the suitability of the model, we also adopt the Desicion Tree Regressor model for comparison prediction.

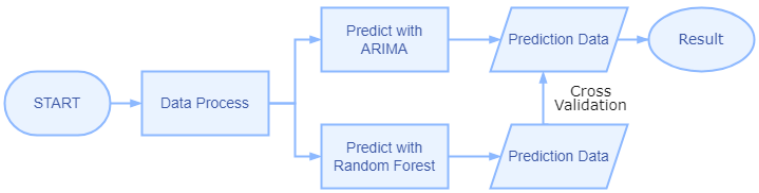


Figure 3: Flow Chart

3.1.2 Flowchart

3.2 Model Construction

3.2.1 Model Principle

ARIMA model is a Time Series Prediction Model which can be used to forecast future time series trends. ARIMA has three terms that correspond to the three main components of the model, respectively stands for self-regression (AR), difference (I) and moving average (MA). Autoregressive (AR) refers to the use of past observations to predict future observations. The model assumes that future values are a linear combination, where the weights are controlled by the autoregressive coefficients (AR coefficients). Moving average (MA) refers to the use of past forecast errors to predict future observations. The model assumes that the future values are a linear combination of past forecast errors, where the weights are controlled by the moving average coefficients (MA coefficients). The weights are controlled by the moving average coefficients (MA coefficients). Differencing (I) refers to the differencing of the time series to make it smooth (i.e. mean and variance do not change over time).

By differencing the time series one or more times, the ARIMA model can remove the trend and seasonality, thus making the time series smooth. A smooth time series can be more easily modeled and predicted.

A regression tree is a decision tree based machine learning algorithm for regression analysis of continuous data. The basic idea of regression tree is to recursively divide a data set into regions, each region is represented by a constant that represents the output values of all data points within that region. The regression tree recursively divides the dataset into regions until only one data point remains in each region or a predefined stopping condition is reached.

The general formula we use for ARIMA(p, d, q) model can be written as:

$$y_t = c + \phi_1 y_t - 1 + ... + \phi_p y_t - p + \theta_1 e_t - 1 + ... + \theta_q e_t - q + e_t$$

where:

- y_t is the value of the time series at time t
- c is a constant (or intercept term)
- $\phi_1, ..., \phi_p$are the autoregressive coefficients of lag 1 through p

- $\theta_1, ..., \theta_q$ are the moving average coefficients of lag 1 through q
- e_t is the error (or residual) term at time t, which is assumed to be normally distributed with mean zero and constant variance
- d is the order of differencing required to make the time series stationary

The autoregressive component of the model (AR) captures the dependence of the current value on the previous p values of the time series. The moving average component (MA) captures the dependence of the current value on the past q error terms. The order of differencing (d) determines the number of differences required to make the time series stationary, which is a necessary condition for an ARIMA model.

The parameters ($\phi_1, ..., \phi_p, \theta_1, ..., \theta_q$) are estimated from the data using various methods, such as maximum likelihood estimation or least squares estimation. Once the parameters are estimated, the model can be used for forecasting future values of the time series.

3.2.2 Model Assumption

1. the number of daily reports is only time-dependent, i.e., the difficulty coefficient of the word of the day, the number of people who chose the difficult mode and the percentage of the number of attempts and other indicators not mentioned are not relevant for the number of daily reports.

2. the first-order difference of the data is divided into a smooth time series

3. Only the data from January 7, 2022 to December 31, 2022 are known to make the prediction, without considering the influence of chance factors such as the increase of game attention due to the American game on the game's enthusiasm.

3.2.3 Main Model

1. Testing the smoothness of first-order difference time series with ADF unit root.

Table 2: Test Outcome

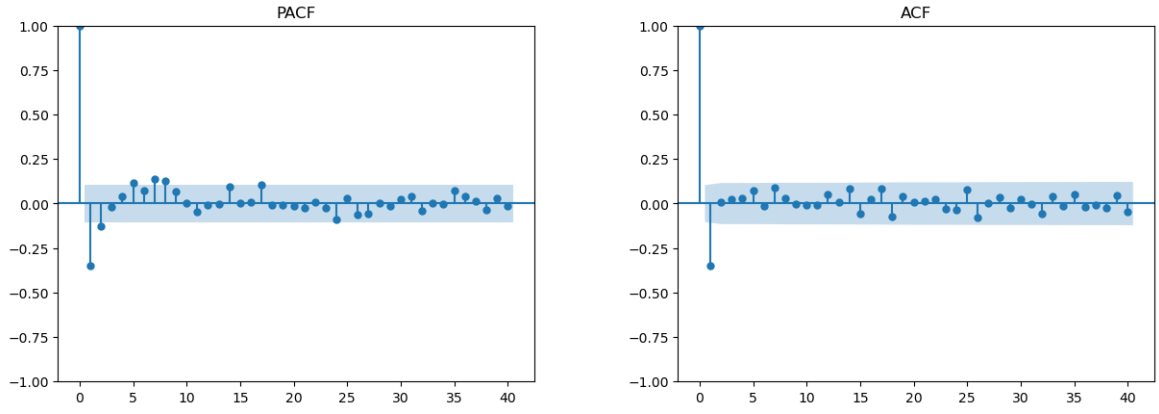
Parameter	Value
ADF	-7.1359413592169885
p	3.4210335686568374e-10
1%Confident Value	-3.4496162602188187
5%Confident Value	-2.870028369720798
10%Confident Value	-2.5712922615505627

There is greater than 99% probability of rejecting the original hypothesis, so the series is a smooth time series.

2. Ljung-Box test is used to verify whether the series is white noise

Each p-value is less than 0.05 or equal to 0, which means that the data is not white noise data and the data is valuable to continue the analysis.

3. Draw pacf plot and acf plot



(a) PACF (b) ACF

Figure 4: ACF and PACF

4. Model tuning

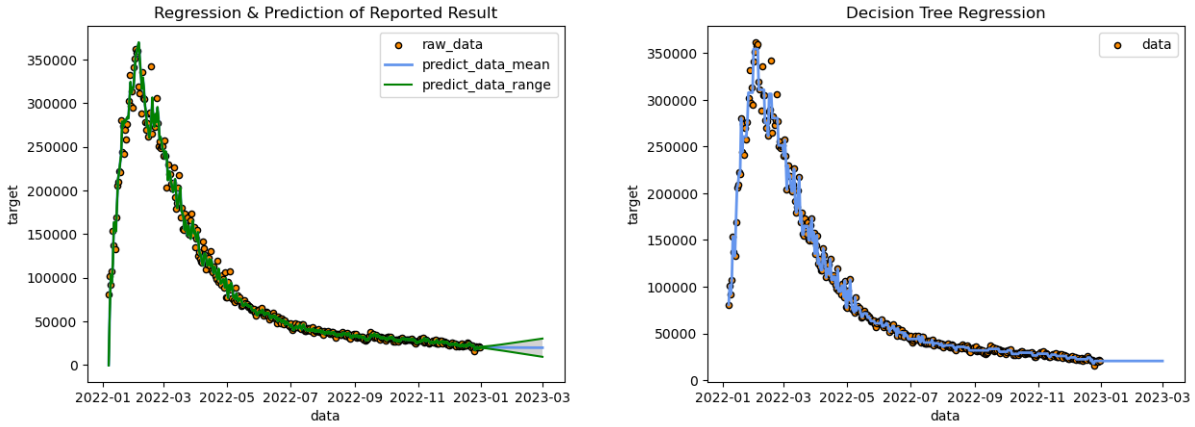
Determination of AR() and MA() model parameters by heat map

Figure 5: Thermodynamic Graph

5. With comprehensive consideration, we made a decision to fit the ARIMA(4,1,2) model to the time series, where we found the ultimate answer for our prediction.

$$\left(1-\sum_{i=1}^{4}\alpha_i L^i\right)(1-L)^1 y_t = \alpha_0 + \left(1+\sum_{i=1}^{2}\beta_i L^i\right)\varepsilon_t$$

In order to preclude the deviation caused by the conventional model, we put up the Random Forest Method for further validation.



(a) ARIMA(4,1,2) (b) Regression Tree

Figure 6: Predicts

3.3 Result

3.3.1 Specifically on March 1, 2023

From the ARIMA model, we figure out the prediction interval for the number of reported results on March 1, 2023

- The mean of the prediction interval is: 19697.53 = 19698
- Upper limit of the prediction interval: 30148.69 = 30149
- Lower limit of the prediction interval: 9246.36 = 9246

3.3.2 Random forest method

This is for further validation with a training accuracy of 0.98259.

Predicted value: 20439.16 = 20439.

3.4 Sensitivity Analysis

From the ARIMA model, we can predicte the number of reports on a "future time" such as Feburary 1th.

- The mean of the prediction interval is: 19847.08 = 19847
- Upper limit of the prediction interval: 25321.29 = 25321
- Lower limit of the prediction interval: 14372.87 = 14373

The realistic result basically lies in the center of the prediction interval, which proves that the model fits well. The sensitive stability can be strengthened by the two different prediction models.

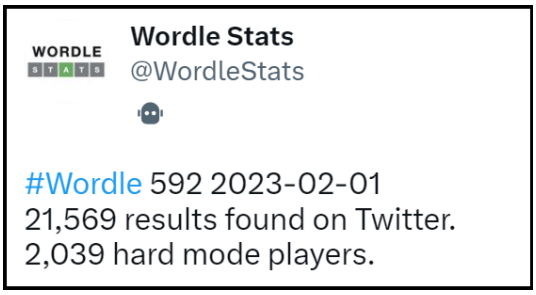


Figure 7: Data From Twitter

4 Influence of Word Attributes on Hard Choices

4.1 Problem Analysis

In this section, we aim to figure out whether there are attributes of the word that affect the percentage of scores reported that were played in Hard Mode.

Attributes means the characteristics or properties that can be used to describe or define a word. Some common word attributes include part of speech, synonyms, antonyms, definition, pronunciation, etymology, and usage. [4] After condensing this classification, we focus on three basic attributes: Appearance Rate (AR), Sentiment Value (SV) and the part of speech (POS).

Here is the chart for better comprehension of what we did on this problem.

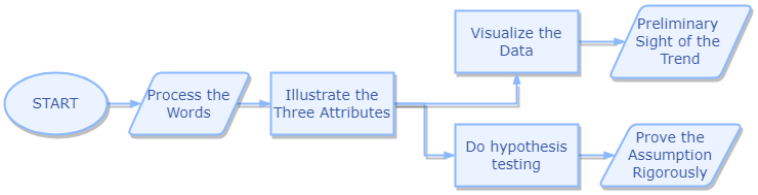


Figure 8: Flow Chart

4.2 Model Construction

4.2.1 Model Assumption

We introduce three character for you in this section as three main attributes of a word.

They are:

- The first one describe the rate one letter happens to show up at a certain slot with the support of all words in corpus, the complete potential word set. We name it Appearance Rate (AR).
- Every word has its sentiment value(SV) in the human view. Some negative like "bitter", some positive as "happy" and some with even affections like simple object nouns. We quantify this emotion and take it into consideration to see whether it can effect the frequency people choose the hard mode.

- Part of speech (POS) is the characteristic or property of a certain word, crudely be divided into Verbs, Nouns, Numerals, Adjectives, Adverbs, Pronouns, Articles, Prepositions, Conjunctions, Interjections and so on.

4.2.2 Main Model

- For
-
- The words in the full dataset were selected and each word lexical property, so called POS, was classified using python NTLK, from which we import the pos_tag to compute specific value.

  Among the multiple data obtained by classification, most types of words have outcome data below 10. To improve the robustness and reliability of the data, we selected data sets which has a value above 10 for concentration, including nouns (NN) adjectives (JJ) verbs (VB).

4.3 Result

5 Predict the Distribution

5.1 Problem Analysis

- Development of the model for future dates future scenario words and reported its results. Gives percentages of (1, 2, 3, 4, 5, 6, X) for "EERIE" on March 1, 2023
- uncertainty of the model prediction
- the confidence of the model prediction

This is the second question mapping to the requirement we dig more data attributes

1. The combine_score

When two word appears to be a combination in a word, people are prone to choose the one their more familiar with. It is known as the "preferred-candidate effect" or "lexical bias," which refers to the tendency of people to select the more familiar word when encountering a combination of two words. Once a study found that the familiarity of the constituent words affected the processing of compounds in the brain, with more familiar words being processed faster.[Schriefers, H., & Teruel, E. (2000).]

Due to the five-word limit, we only extract the two-word combination in English Letter Frequency. With data found in `http://norvig.com/mayzner.html`, we score each of the words by identity the two words combination inside them.

For example, we got the word "intro". Put it into the word pool, we have "in" stands for possibility of 2.43, "ro" stands for possibility of 0.73, "nt" stands for possibility of 1.04, finally the sum reaches 4.30, which is the value of the combine_score of "intro".

2. The green_score

Follow the principal of the game, when a letter reaches it right position in a word, Wordle will appear green in the corresponding place. It gives a good perspective to identify the importance of a word, so that we decided to score a word by the possibility it appears correct on its position. An important attribute for a word is the way it arranged. And the position it occupied has its possibility influence on the possibility we guess right. Here is the way we design to quantify this characteristic.

As far as we've concerned, the overall letter may be shown on Wordle is 2315, which is selected by creator Wordle in the 12972 valid words [B. J. Anderson and J. G. Meyer].

For every letter in this database, we denote it as a_i and divide it into five parts by site. $a = [x_{11}, x_{12}, x_{13}, x_{14}, x_{15}]$

Here is the notation for the subsequent algorithm.

Table 3: Notation

name	illustrate	name	illustrate
D	letter database with a size of 2315	a_i	a word in D
$\#a_i$	number of a_i	x_{ij}	a letter in a_i
P_i	its possibility$_s core$	p_i	one certain letter on certain position ei. s is the start letter
SCOPE	the enlarged dataset	EXTEND	the constant 500000

The following we show the connections of the above notations.

$$p_i = \frac{\Sigma\{some\ certain\ letter\}}{\Sigma_1^{2315}}$$

$$P_i = \Sigma_1^5 p_i$$

To make a Monte Carlo simulation, we need to enlarge the database. Follow the possibility_score, we generate more data with the below function.

$$\frac{\#(a_i)}{\#(a_j)} = \frac{P_i}{P_j}$$

$$\Sigma_{i=1}^{2315} a_i = EXTEND$$

Finishing all the preparatory work, we can get to the Algorithm part to calculate green-score.In the following code, x means every word we put in.

```
foreach x do
  define count, true_times,random-database is D
  do p_x*50000 times:
```

```
random-a-word(p_x)
while(word doesn't matched)
  narrow the random-database
  count++
The final word adds one point to the identical word's second line of D
if(count<6)true_times++
```

Here we define two useful number to illustrate this green_score. It depends on the algorithm to choose one of the them as the final green_score.

- correctness:

$$correctness = \frac{true\_times}{\#a_i}$$

- leave_possibility: the possibility a number will be returned by the algorithm

3.yellow_score

4.entropy

And we found that the distribution of the data set given by the question, i.e., the correlation percentage of future dates (1,2,3,4,5,6,X), is a fixed order data, all have some correlation, so we use the method of plotting histograms and fitting the data.

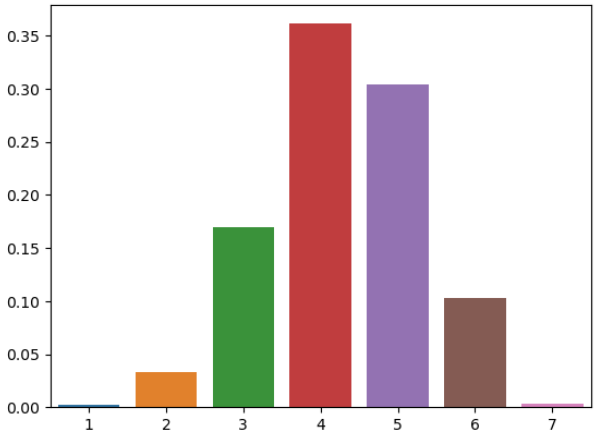
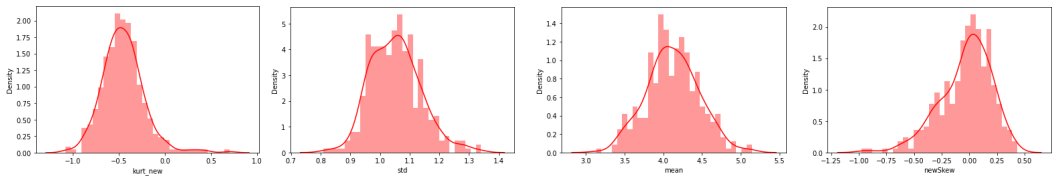


Figure 9: Histogram

The correlation percentages are interpreted as the mean, standard deviation, skewness, and kurtosis. The approximate distribution of these four values is as follows

Through correlation analysis, we found that skewness and kurtosis are basically uncorrelated with word attributes, and the mean and standard deviation correlations are low enough to be analyzed as two independent indicators.

For the consideration of the time direction, since we used the ARIMA model in the first question to give the relationship between time and the number of reporters, in this question, we used the



(a) (b) (c) (d)

Figure 10: Predicts

number of reporters instead of time to influence the distribution of the reported results. Thus, by regressing the random forest regression model, we use the five dimensions of the indicators (1234+number) to map the mean and standard deviation of the two indicators, and fit the mean and standard deviation to the distribution function. to predict the distribution of the reported outcomes.

5.2 Model Construction

5.2.1 Model Principle

Random can handle very high dimensional data and does not have to do feature selection (because the feature subset is chosen randomly), so for our multiple features, the random forest model works well. And to better observe the uncertainty, random forest can give which features are more important, and changing important features is more likely to perturb the model to determine the sensitivity and uncertainty of the model.

5.2.2 Model Assumption

1. it is assumed that the time parameter only affects the number of reports per day, independent of the difficulty of the questions, the quality of the participants and the accumulated game experience.

2. The standard deviation of the report distribution is assumed to be independent of the kurtosis and word properties.

3. It is assumed that the noise caused by chance does not mask the characteristics of the mean and skewness

5.2.3 The Four New Variants

5.3 Result

5.3.1 General Result

5.3.2 Specifically for "EERIE" on March 1, 2023

Give a specific example of your prediction for the word EERIE on March 1, 2023.

5.4 Sensitivity Analysis

6 Classify Words Model

6.1 Problem Analysis

Flowchart

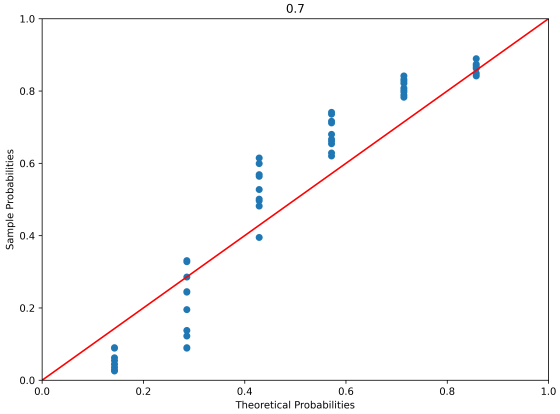


Figure 11: QQ distribution

6.2 Model Construction

6.2.1 Model Principle

6.2.2 Model Assumption

6.2.3 Main Model

6.3 Result

6.3.1 General Result

6.3.2 Specifically on March 1, 2023

prediction interval for the number of reported results on March 1, 2023

6.4 Sensitivity Analysis

7 interesting features

According to the problem, we could learn from the fact that the dataset given by this problem was taken from Twitter, but Twitter's data was not comprehensive since users have the right to choose whether to report their results. Following the rules of Wordle, users have six chances to get the right answer, Otherwise he/she will fail the game. This game was a little bit difficult for most of us. However, after integrating the percentage of each try times in the dataset, we find most of the users completed the game within 5 trials, and only few (2.8%) failed the game, which is not in

line with our tuition, **so we take the assumption that users who reported their results are usually those who completed the game and use less chances.

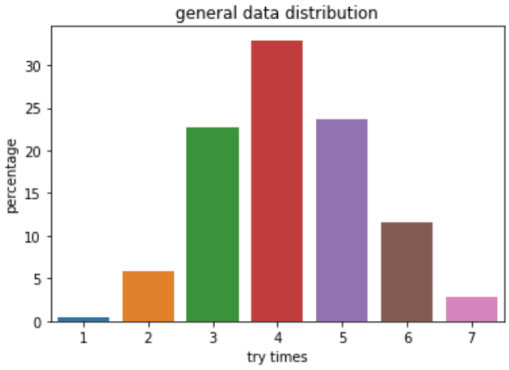


Figure 12: The bar chart

In order to verify our assumption, we developed a model to simulate the process of how people guess words in ***Wordle***. To simplify this model, we assume people choose words which come to their mind first, which has a high correlation with word frequency in English. So we continuing using the word frequency data taken from ***wolftram***, distributing probabilities to each word according to their frequency, and choose them randomly. the algorithm flow are as follows:

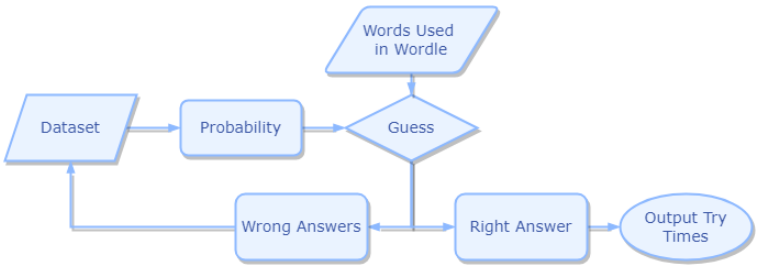


Figure 13: The Algorithm Flow

We ran each word 20 times to find the average expectation, compared it with the expectation of the known words of the topic, and made a correlation analysis with hypothesis testing, as shown in the following figure.

We calculated each word for 20 times, getting their average try time and compare it with the true average try time, seen from the line chart , the predict data was much higher than real data. and the Mean expected difference of these (359 words) was 6.925 (11.046-4.121),which was related to our assumption.

then we need to test the correctness of our algorithm model. we calculated Pearson correlation coefficient, and did hypothesis testing of these 2 data the result of which are as follow:

Table 4: Relevance

	test	predict
Pearson Relevence	.530**	1
Sig. two-sided hypothesis	.000	

Since p=0.000<0.01, we are 99% sure that these 2 data have significant correlation, and the Pearson correlation coefficient is 0.53. We can conclude our model is useful and correct to some extent.

To make further analysis, we integrated all predicted data and real data, getting its distribution and drew its histogram. As can be seen from the image, users who report their data are only the tip of the iceberg of the whole data. Behind the displayed data set, there are a large number of users who fail to guess the word successfully. Most users who share their data on Twitter had figure out the puzzle and had good game scores.

Based on this model, we try to figure out in what kind of situation would users like to report their grades: we divide real value by predict value and rescaling it, and draw its line chart

Within the allowable range of errors, we find that, unlike the expected results, the proportion of sharing data does not show a monotonically decreasing trend, but first increases and then decreases, and reaches a peak at x=4. Although generally speaking players with better results are more likely to post their grades, the percentage declines when the results are too good (the number of attempts is close to one). We reasoned that such groups guessed words based more on probability than on the amount of information in the question. Rapid success does not make users feel the fun of the game, while users in the interval of 3 to 5 guess data more rely on the information of ideas and topics constantly updated. The stronger the sense of achievement brought to users by guessing words in this interval, it is in line with the inference and expectation of psychology.

8 Strengths and weaknesses

8.1 Strengths

- **Applies widely**
  This system can be used for many types of airplanes, and it also solves the interference during the procedure of the boarding airplane,as described above we can get to the optimization boarding time.We also know that all the service is automate.

- **Improve the quality of the airport service**
  Balancing the cost of the cost and the benefit, it will bring in more convenient for airport and passengers.It also saves many human resources for the airline.

-

References

[1] Morning Consult. (2022, January 20). The top 10 mobile games in the US: January 2022.

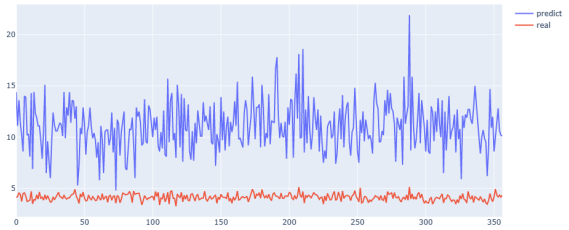


Figure 14: The Prediction Result

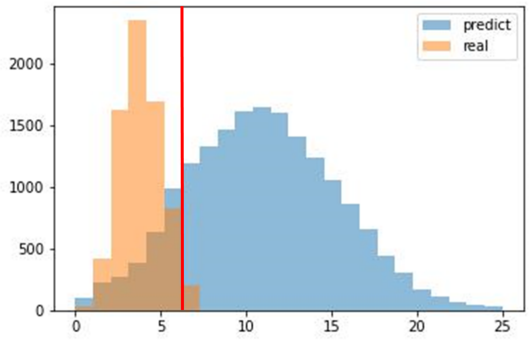


Figure 15: Data Mountain and its Knap

https://morningconsult.com/form/top-mobile-games-us-january-2022/

[2] "Wordle Stats." Twitter, January 1, 2022.

[3] OnePoll. (2022). The Wordle survey. https://www.onepoll.us/the-wordle-survey/

[4] Oxford English Dictionary. (2022). Entry. In Oxford English Dictionary Online. Retrieved February 20, 2023, from https://www.oed.com/

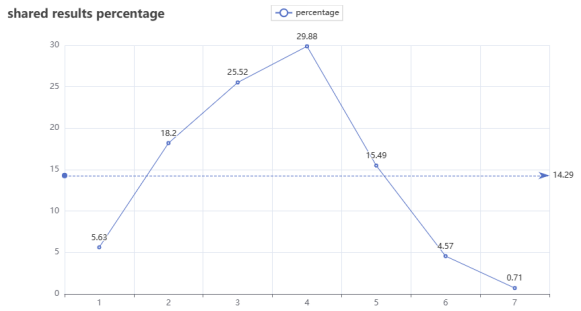


Figure 16: The Line Chart

Dear, Mr. Alpha Chiang

After processing the data to a more accurate version, we do a prediction of number of people sharing on Twitter in the upcoming few days. It is the Variation towards time model.

Furthermore, we discover a connection between Word Attributes and peoples' specific casting preferences in the hard mode.

In the fifth section, we forecast how individuals would do given a specific phrase and predefined time.

We created a Classify Words Model after extensively learning from existing models. It can give a profit of a word from the result side.

What's more intriguing is that, in contrast to the surveys mentioned above, we focused on the interesting features of this game, from which we created new neural network learning algorithms to abstract "a changing throng" and analog the distribution of the actual number of players when playing this game, regardless of whether they tweeted their results or not. There after, We discovered a correlation between the frequency with which people solve this challenge and their desire to share it on Twitter from the Analog Gamer Mass, which is in line with psychological studies.

Sincerely yours,

Your friends

Appendices

Appendix A First appendix

In addition, your report must include a letter to the Chief Financial Officer (CFO) of the Goodgrant Foundation, Mr. Alpha Chiang, that describes the optimal investment strategy, your modeling approach and major results, and a brief discussion of your proposed concept of a return-on-investment (ROI). This letter should be no more than two pages in length.

Here are simulation programmes we used in our model as follow.

Input matlab source:

```
function [t,seat,aisle]=OI6Sim(n,target,seated)
pab=rand(1,n);
for i=1:n
    if pab(i)<0.4
       aisleTime(i)=0;
    else
        aisleTime(i)=trirnd(3.2,7.1,38.7);
    end
end
```

Appendix B Second appendix

some more text **Input C++ source:**

```
//============================================================================
// Name        : Sudoku.cpp
// Author      : wzlf11
// Version     : a.0
// Copyright   : Your copyright notice
// Description : Sudoku in C++.
//============================================================================

#include <iostream>
#include <cstdlib>
#include <ctime>

using namespace std;

int table[9][9];

int main() {

    for(int i = 0; i < 9; i++){
        table[0][i] = i + 1;
    }

    srand((unsigned int)time(NULL));
```

```
    shuffle((int *)&table[0], 9);

    while(!put_line(1))
    {
        shuffle((int *)&table[0], 9);
    }

    for(int x = 0; x < 9; x++){
        for(int y = 0; y < 9; y++){
            cout << table[x][y] << " ";
        }

        cout << endl;
    }

    return 0;
}
```

(1)

$$a^2 \tag{1}$$

$$\begin{pmatrix} *20ca_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \frac{Opposite}{Hypotenuse} \cos^{-1}\theta \arcsin\theta$$

$$p_j = \begin{cases} 0, & \text{if } j \text{ is odd} \\ r!\,(-1)^{j/2}, & \text{if } j \text{ is even} \end{cases}$$

$$\arcsin\theta = \oiiint_{\varphi} \lim_{x\to\infty} \frac{n!}{r!\,(n-r)!} \tag{1}$$

Theorem 8.1.

Lemma 8.2.

Proof. The proof of theorem. □

8.1 Assumption

-
-
-
-