|  |  |  |
| --- | --- | --- |
| **Problem Chosen** C | **2024 MCM/ICM Summary Sheet** | **Team Control Number** 2401509 |

**Your Paper's Title**

**Summary**

1.开头：1.简单交代背景

2.所做事情

3.解决问题的实际意义（可选）

2.中间段：应用了什么方法，解决了什么问题，得到了什么结果

对问题的求解思路，应用了什么模型，注意紧扣题目背景描述，以及通过这个方法/模型得到的结果

若为数值答案，直接摆答案。模型中有重要参数时，可以做灵敏度分析（做了灵敏度分析后，得到的结果是。。。），若概率统计相关可以考虑加上置信区间，若为预测类或数值计算类，可以考虑做误差分析。

若为开放性的评价/建议类问题，写主要结论，一定要明确。若完整答案比较长，只要写最主要的部分，加一句话引导读者到正文或附录中查看完整结果。只要写明确的结果即可。能加上一些具体的数值结果一定也要摆上来。

We discovered that...

3.结尾段：（可选）总结，介绍论文两点，或者对类似问题的适当推广，千万不要写主观的内容和评价！

不要写废话，要紧扣背景，有理有据，令人信服。。。

关键词 4-6个，使用的主要模型（3-4？）。论文中出现次数较多的重要词（1-2？）大部分在摘要里找

**Keywords:** keyword1; keyword2; keyword3; keyword4

Contents最后记得更新整个目录

[1 Introduction 3](#_Toc157187213)

[1.1 Problem Background 3](#_Toc157187214)

[1.2 Restatement of the Problem 3](#_Toc157187215)

[1.3 Our Work 4](#_Toc157187216)

[2 Assumptions and Justifications 4](#_Toc157187217)

[3 Notations 5](#_Toc157187218)

[4 The Data 6](#_Toc157187219)

[4.1 Data Description 6](#_Toc157187220)

[4.2 Data Cleaning and Visualization 6](#_Toc157187221)

[5 Predictive Model of Daily Report Number Variation 7](#_Toc157187222)

[5.1 Winter's Multiplicative Model 7](#_Toc157187223)

[5.2 Model Solution and Testing 8](#_Toc157187224)

[5.3 Report Number forecast Based on the Predictive Model 9](#_Toc157187225)

[6 Predictive Model of Report Result Constitution Based on Word Attributes 10](#_Toc157187226)

[6.1 Selection of Index Attributes and Pre-processing 10](#_Toc157187227)

[6.2 Multiple Linear Regression Analysis 11](#_Toc157187228)

[6.3 BP Neural Network for Result Constitution Prediction 14](#_Toc157187229)

[7 The name of model 3 14](#_Toc157187230)

[8 Sensitivity Analysis 14](#_Toc157187231)

[9 Model Evaluation and Further Discussion 14](#_Toc157187232)

[9.1 Strengths 14](#_Toc157187233)

[9.2 Weaknesses 14](#_Toc157187234)

[9.3 Further Discussion 14](#_Toc157187235)

[10 Conclusion 14](#_Toc157187236)

[References 15](#_Toc157187237)

[Appendices 16](#_Toc157187238)

# Introduction

## Problem Background

Wordle, the wide-loved web world puzzle game, was first invented by software engineer Josh Wardle, and received its shocking rise in popularity during the COVID-19 pandemic. Each day the game generates a five-letter word from a compact word list of 2,309 commonly recognized words. The game requires players to guess the word within six tries, each try followed by hints given in the form of colored tiles, where a green tile indicates that the letter is included in the word and is in the correct location. a yellow tile indicates that the letter is included but not in the correct place, and a grey tile means the letter is not included by the word at all. Players can make dynamic change to their strategy according to the given information. The game also sets a hard mode, where players are required to include the correct letters they have found in a word in their next try.

|  |
| --- |
| A four-row grid of white letters in colored square tiles, with 5 letters in each row, reading ARISE, ROUTE, RULES, REBUS. The A, I, O, T, and L are in gray squares; the R, S, and E of ARISE, U and E of ROUTE, and U and E of RULES are in yellow squares, and the R of ROUTE, R and S of RULES, and all letters of REBUS are in green squares. |

Figure 1. Wordle puzzle providing hints in the form of colored tiles.

While the word for each day is randomly generated, the reported number and portion of people succeeding in guessing the word in each try and people fail to guess the word can vary under complicated factors, including certain intrinsic attributes of the word, players' favored playing habits and puzzle-solving routines, their willingness of reporting, etc. Therefore, the daily results reflected by the reported data obtained from Twitter might well manifest some interesting characteristics. Among all the factors mentioned, we are most interested in the relation between the words' certain attributes and people's daily results. So, what are words like that are easy to solve? How is it related to the varying daily reports of results from people? These form the primary goal of our research.

## Restatement of the Problem

In our research, we aim at analyzing the statistical characteristics of the game's daily results from January 7, 2022 to December 31, 2022, based on the data obtained from users' reports on Twitter. Meanwhile, we want to make predictions about future results reported based on the analysis, and discover some regular features of the game, including the correlation between the words' intrinsic attributes and the daily results, and the difficulty level of different words. Considering the given requirements and provided data set, the problems are summarized as follows.

* Develop a model to explain the variation of the total number of reported results daily, and exploit the model to create a prediction interval for the number of reported results on March 1, 2023. Then find out and explain the possible correlation between the attributes of the word and the percentage of scores reported that were played in Hard Mode.
* Develop a model that predicts the distribution of reported results on a given future date with a given solution word. Analyze the uncertainties and reliability of the prediction based on a specific prediction for the word EERIE on March 1, 2023.
* Develop and summarize a model that classifies words by difficulty according to different word attributes. Identify the difficulty class of the word EERIE and evaluate the accuracy of the classification model.
* Discuss some other interesting features of the data set.

## Our Work

# Assumptions and Justifications

Our model is established on the basis of the following assumptions:

* **The provided data is valid, which means people reporting in the data keep true to their real scores and the mode they chose to play, and the game runs strictly under its current rule.**

The validity of our model analyses and prediction is based on the authenticity of the data provided. Considering that fake reports can be sorted out in the data collecting process, for example, only include reports that have attached images of their playing results, we decide that fake reports are not given consideration.

* **No major social change that can significantly impact the number of players occurs during the period.**

Social events like the COVID-19 pandemic can have great influence on the daily number of game participants. However, the provided data only covers the time period ranging from January 7, 2022 to December 31, 2022, which is over one year after the pandemic. Considering that social factors are not the main research object in our study, social environment is also given minor consideration in the model.

* **There is no pervasive answer sharing among players in each day's game. In other words, players do not know the answer before they solve the puzzle.**

We assume that most players solve the puzzle on their own and those who get answer from people who have already played the game count for an insignificant part of the data. Usually answer communication is likely to occur in puzzles that require a longer solving process, which makes people feel tortured by being unable to reach a solution and become eager for an answer. Because Wordle requires people to finish their guessing within only six tries, we assume that people will unlikely go to search for an answer from others.

* **Most players possess some initiative in solving the word. They have a tendency to make better choices.**

Our evaluation method of word difficulty is to a great extent dependent on the assumption that most players participating in each day's game have a desire to succeed, and they opt an intuitively better choice in the word they are going to use in their next attempt according to the hints, though the option might not be optimal. It is reasonable because it is the nature of people to have a desire to get better results when they decide that they are in for a game, else they would better not participate. Even though there are those who play casually, making merely random choices, this type of people is unlikely to take up the majority.

# Notations

The key mathematical notations used in this paper are listed in Table 1-2.

Table 1: Notations in Model 1

|  |  |
| --- | --- |
| **Symbol** | **Description** |
|  | Frequency of seasonality. |
|  | Level smoothing parameter of Winter's Multiplicative Model. |
|  | Trend smoothing parameter of Winter's Multiplicative Model. |
|  | Seasonal smoothing parameter of Winter's Multiplicative Model. |
|  | Current date. |
|  | Observed number of scores reported on the  day. |
|  | Expected level of report number on the  day. |
|  | Expected variation trend of report number on the  day. |
|  | Expected seasonal variation of report number on the  day. |
|  | Days ahead of current date of the predicted date. |

Table 2: Notations in Model 2

|  |  |
| --- | --- |
| **Symbol** | **Description** |
|  | Information quantity. |
|  | Possibility of the random event . |
|  | Information entropy. |
|  | The  letter in the word . |
|  | The original set of suggested answer word list. |
|  | The set of suggested answer word list after one guessing. |

# The Data

## Data Description

In this research, we use the data set of daily results of the Wordle game from January 7, 2022 to December 31, 2022, collected from reports of players of their scores on Twitter with data mining. The data includes the date, number of contest, word of the day, the number of people reporting that day, the number of people on hard mode, and the percentage of players succeeding in guessing the word in different tries. A failure in solving the problem is denoted by *X*.

## Data Cleaning and Visualization

During our preliminary analysis of the data set, we noticed abnormal values at the date November 30, 2022 in the data of all scores reported and February 13,2022 in the data of scores reported played in hard mode. The data were modified with linear interpolation. Visualization of the cleaned data is presented below.

|  |
| --- |
|  |

Figure 2. Visualization of fluctuation of total number of scores reported.

|  |
| --- |
|  |

Figure 3. Visualization of fluctuation of total number of scores played in hard mode.

The visualized data indicates a transformation in the variation pattern from a steep rise during the first month to a rapid decrease from February to May, and enters a subsequent stable decreasing process. We selected this part of data covering the period from May 22 to December 31 for the establishment of our predictive model. The included data shows a decreasing trend in the long run and a periodic fluctuation. Intuitively, the decreasing trend presents a linear character, and the periodic fluctuation is relatively regular and repeats every approximately seven days. This makes sense because most people follow a weekly life cycle. Weekends allow them more time and leisure to enjoy puzzles, while they hardly have time to be spent on such games on weekdays. Therefore, we made the preliminary assumption that the variation of both number of all scores reported and number of scores reported played in hard mode obey a one-week cycle.

# Predictive Model of Daily Report Number Variation

## Winter's Multiplicative Model

We first applied time series analysis and regression to describe variation of the total number of scores reported daily. As is described in former analysis, the data manifests a combined feature of a long-term linear decreasing trend and a weekly fluctuation. Therefore, we selected the Winter's Multiplicative model to describe its variation process. The model is an extension based on single exponent smoothness model. It shows efficiency in prediction data that involves both long-term trend and seasonal variation.

We identified the seasonal variation of time series to follow a cycle of 7 days, which is denoted by seasonality frequency *m*. Based on the model, the variation of the total number of scores reported daily is denoted by following equations.

The level equation is described as:

|  |  |
| --- | --- |
|  | () |

where  describes the expected level of the total number of scores reported on the  day. The level smoothing parameter  decides the weight of past data in estimating the result level of the next day.  is the actual observed value on the  day.

The expected trend of variation in report number  is described as:

|  |  |
| --- | --- |
|  | () |

where the smoothing parameter  is the weight of past level and trend components in estimating the trend of the next day.

The seasonal component  is described as:

|  |  |
| --- | --- |
|  | () |

where the seasonal smoothing parameter  decides the weight that past level, long-term trend and seasonal component should be considered in estimating the seasonal variation of the

report number of the next day. Seasonal component varies with a cycle of 7 days.

The estimated value of future number of scores reported is influenced by the superimposed effects of the trend and seasonal components. In Winter's Multiplicative Model, the aggregational effect is their product.

The predictive equation of Winter's Multiplicative Model is presented as follows. It gives the predicted value on the *t+hth* day, which is *h* days ahead of the current date *t*.

|  |  |
| --- | --- |
|  | () |

## Model Solution and Testing

We picked out the data during the period from May 22 to December 1 as training group,

while data after December 2 as testing group. Fitting result is presented in the diagrams below.

The smoothing parameters are decided under maximized fitting goodness. When  reaches maximum value, ,  and  take values of 0.223, 0.058 and 0.025.

The result of Q-test with residual shows a significance of over 0.05. Residual ACF and PACF indicates that the autocorrelation coefficient of residual is not significant at all lag orders. Therefore, we contended that the model ideally describes the variation of the number of scores reported daily.

|  |
| --- |
|  |

Figure 4. Residual ACF and PACF of the predictive model.

|  |
| --- |
|  |

Figure 5. Fitting result of report number variation with Winter's Multiplicative Model

We used the model to predict the testing group data and compared the prediction with observed value. The result is presented below.

|  |
| --- |
|  |

Figure 6. Model prediction result of the testing group.

Generally, the model is fairly effective in depicting the long-term downward trend and weekly fluctuation of future number of scores reported. The result demonstrates ideal predictive ability of the model established.

## Report Number forecast Based on the Predictive Model

We used our predictive model to create a prediction interval for number of reported results on March 1, 2023. We let the parameters  vary within a 50% confidence interval. The result shows that the predicted upper bound of the number of reported results on March 1,

2023 is 19280.1, while the lower bound is 3457.5.

|  |
| --- |
|  |

Figure 7. Prediction interval within 50% confidence interval.

# Predictive Model of Report Result Constitution Based on Word Attributes

## Selection of Index Attributes and Pre-processing

To describe the correlation between certain features of words and players' results, it is necessary to find out the index attributes of words that have decisive effects on the guessing process. We considered the attributes of words that could possibly affect the constitution of reported results, which summed up as five index attributes.

* **Number of syllables**. People tend to think of words with fewer syllables first when they make an attempt in guessing. Fewer syllables are likely to be correlated to success in fewer attempts. We used the Python pronouncing library to help count the syllables of every word. For words that are not included in the library, the data were manually set.
* **Word usage frequency**. Words that are more frequently used in daily life are more likely to pop into people's mind while they are in a game. If the word of the day is a frequently used word, people may also get higher scores.
* **Information entropy**. Information entropy evaluates the expected quantity of information a random event could possibly bring. For each possible result of every attempt, the colored tiles provide a different amount of information. The information entropy of a word applied in guessing is the expected information it could provide in a guess. The appliance of information entropy is based on the rational man supposition of Wordle players.

The amount of information is calculated by the following expression:

|  |  |
| --- | --- |
|  | () |

Here *p* is the possibility of the random event, and *I* denotes the amount of information, which is measured by *bit*. A bit of information identifies a possible result with a possibility of 50%. Therefore, we calculated information entropy with the equation presented below.

|  |  |
| --- | --- |
|  | () |

where *x* traverses all possible situation after the guess.

The information entropy included in each attempt equals the difference between total information entropy of the suggested answer words before and after the attempt. Here we assumed that the possibilities of occurrence of all words included in the suggested answer list in the next guess are equal. However, the expense of directly calculating information entropy of every word turned out to be unacceptable. To simplify the process, we calculated the information entropy of words by letter. For a guess using the word *x* in the suggested answer list, the information entropy of the *ith* letter is:

|  |  |
| --- | --- |
|  | () |

where *H(x[i])* is the information entropy of the *i*th letter in the word *x*. *wordlist* is the suggested answer list before the guess, while *wordlist'* is the list after the guess. The total information entropy of the word *x* equals:

|  |  |
| --- | --- |
|  | () |

* **Occurrence of particular letters**. While on one hand words with less frequently used letters like *z* may not become the prior choice of players, such letters provide a considerable amount of information when they appear as green or yellow tiles, which significantly narrows down the scope of suggested answers. Words with less frequently used letters might indicate a larger proportion of failures in all reports and meanwhile a larger percentage of results with fewer steps in all successful results.

We preprocessed the data and find letters whose occurrence frequency is closely related to the constitution of daily results. Among all the selected letters, *w, y, g, j, z* are letters with least occurrence frequency, and *a, e, h, i, o, s, r, t* and *n* has highest occurrence frequency.

* **Occurrence of repeated letters**. If the five-letter word contains repeated letters, one hint given by a green or yellow tile of the letter can provide more information.

For each word of the day, we counted the repeated letter pairs contained. A letter repeated once counts as 1 letter pair, while a letter repeated twice counts as 3 pairs.

## Multiple Linear Regression Analysis

The composition of daily reports depends largely on the overall difficulty of words. For quantified evaluation of word difficulty, we calculated the expected contribution of each guess under all possible word of the day in the Wordle wordlist. The less the expected contribution of a guess, the more difficult the word is estimated to be.

We used multiple linear regression to estimate the respective contribution of the five index word attributes. We have six variables in the regression model: number of syllables, word usage frequency, information entropy, occurrence frequency of 9 common letters *a, e, h, i, o, s, r, t*, *n,* occurrence frequency of 5 uncommon letters *w, y, g, j, z,* number of repeated letter pairs.

The former 334 words of the day out of the data set were selected as training group data, while the remaining were left as testing group. Preliminary regression result of OLS and variable regression coefficients are presented in Table 4 The regression coefficients are regarded as quantified distribution of factors to the guessing process.

**Table 4. OLS result of multiple regression**

|  |
| --- |
|  |

We noticed that the regression coefficient of word usage frequency is close to zero. To reduce possible error, the regression coefficients were modified after standardization of word usage frequency. The modified coefficients are presented below.

**Table 5. Modified coefficient of OLS**

|  |
| --- |
|  |

Visualization of residual shows ideal regression result, as is presented in the diagram below. The residuals stay in acceptable range and basically fits normal distribution.

|  |
| --- |
|  |

Figure 8. Residual distribution of OLS regression.

|  |
| --- |
|  |

Figure 9. Residual bar chart indicating normal distribution.

We investigated possible existence of heteroscedasticity in the regression. Both BP test and White test results indicate that the model does not include significant heteroscedasticity in any form with a confidence interval of 95%.

Then, we tested VIF to check for multicollinearity. VIF results reflect that variables show no significant signs of multicollinearity, apart from the virtual variables of existence of 1 repeated letter pairs and 2 repeated letter pairs. Considering that multicollinearity has minor effect on the prediction of the model, we decided to ignore this factor.

We used the regression model to estimate the contribution of testing group data. The residuals are controlled within [-0.01, 0.01]. It shows ideal regression effects.

|  |
| --- |
|  |

Figure 9. Scatter chart of testing group residuals.

## BP Neural Network for Result Constitution Prediction

# The name of model 3

# Sensitivity Analysis

# Model Evaluation and Further Discussion

## Strengths

## Weaknesses

## Further Discussion

# Conclusion

# References

[1]: Wikipedia contributors. "Wordle." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 22 Jan. 2024. Web. 23 Jan. 2024.

[2]: Bonthron, Michael. "Rank One Approximation as a Strategy for Wordle." *arXiv preprint arXiv:2204.06324*  (2022). Print.

# Appendices

|  |
| --- |
| Appendix 1 |
| Introduce: 这里放上附录1的介绍 |
|  |

|  |
| --- |
| Appendix 2 |
| Introduce: 这里放上附录2的介绍 |
|  |

本部分是附录部分，美赛对于附录不是特别看重，今年还限制了论文的页数（从第二页开始编号，不能超过25页）。

\*包括支撑材料的文件列表

一般新起一页列出附录。

在不超过页数限制的条件下，附录中可以包括下面内容：

* 你们写的代码；
* 某一问题的详细证明或求解过程；
* 自己在网上找到的数据；
* 比较大的流程图；
* 较繁杂的图表或计算结果。