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

Riddle of Wordle : Predicting and Analyzing the Popularity and Players' Performance of the Puzzle

Summary

Wordle is a five-letter English word guessing game that has become popular worldwide due to its simplicity, fun, and challenging nature. However, how to provide better marketing strategies based on the game's features and user behavior remains a question. In this study, we propose a framework to analyze the game's popularity, user performance, word difficulty, and data set features to provide insights into effective marketing strategies.

Firstly, to analyze the popularity and patterns of the game, we propose an **Game Popularity and Mode Prediction Model**. This model can predict the number of users and the number of users choosing the Hard Mode for the game on a future day, and can provide a prediction interval at a certain confidence level. By removing time factors that affect mode selection, we can analyze the relationship between the game's difficulty mode and word features.

Secondly, we analyze the performance of users in the Wordle game using an LSTM neural network model based on the normal distribution characteristics of the data.We built a **Player Behavior Prediction Model** based on the normal distribution characteristics of the data. This model trains the relationship between word features and the percentage of answer attempts and can predict the distribution of user guesses for a given word on a specific day. We provide factors associated with model uncertainty and calculate the model's uncertainty using **MC dropout**.

Next, we use an MLP to predict the difficulty coefficient of the word and provide a **Word Hard Classification Model**, categorizing words into easy, medium, and difficult categories. To analyze the attributes of words associated with each category, we use kPCA to extract the main influencing factors for each category. Our analysis shows that EERIE is a hard word with a high difficulty coefficient.

Finally, we provide interesting features of the dataset in terms of **total number of people**, **user game strategy**, and **word features**. Based on our model's conclusions, we reported the current popularity of the game and provided marketing suggestions for future promotion to the puzzle editor of The New York Times,such as strengthening promotion based on the game's popularity and patterns and increasing the difficulty level by selecting less frequently used words or increasing the number of repeated letters in a word.

Keywords: ARIMAX; Confidence Interval; LSTM ; MLP; kPCA

Contents

1 Introduction

1.1 Problem Background

Wordle is an online word guessing game that is both simple and entertaining. In each game, a five-letter target word is provided, and the player must guess the word within six attempts. After each guess, the game provides feedback on which letters of the guessed word are correct. The difficulty of the game is dependent on various factors such as the player's vocabulary, logical thinking ability, and luck. Due to its engaging gameplay, Wordle has gained immense popularity, attracting a vast number of players. This game not only provides a fun and enjoyable experience but also facilitates the improvement of players' vocabulary and language skills while fostering communication and interaction between people.

Recently, the puzzle editor of the New York Times has a keen interest in conducting an in-depth analysis of the existing data of Wordle. The study aims to explore the change pattern of the number of Wordle players over time and investigate players' game performance and strategy choices based on the vocabulary difficulty ratings of different words. By doing so, the study seeks to provide valuable insights and guidance for subsequent marketing strategies and player experience optimization.

1.2 Clarification and Restatement

In this issue, we were provided with a dataset consisting of the Wordle Word of the Day and player scores. The New York Daily News aims to analyze this dataset in order to gain insights into player demand and market operations, with the goal of optimizing the game and supporting its future development. The dataset includes information such as the date, the solution word, the total number of scores, the number of scores achieved in Hard Mode, and the percentage of answers provided within different time frames.

- Explain how the number of reported results changes over time and make projections for March 1, 2023. Explore whether word attributes have an effect on the number of people who choose difficult patterns.
- Predict the associated percentages of (1, 2, 3, 4, 5, 6, X) for a future date and delve into the uncertainty factors of the model. This model is then used to predict the EERIE words for March 1, 2023, and to determine the accuracy of the results.
- The solution words are classified by difficulty, and the attributes of the relevant words in each classification are given. This model is applied to determine the difficulty of EERIE, and the accuracy of the model is analyzed.
- Combined with the results of the above model study,describe some interesting features of this dataset.
- Write a letter to the Puzzle Editor of the New York Times by summarizing the analysis and results.

1.3 Data Cleaning

After observing the provided data, we have identified some errors and have made the following changes to ensure data accuracy and reliability:

1. **Elimination of data with incorrect word length:** According to Wordle's game rules, the solution word should have a length of 5, and data that do not conform to this rule have been removed. For example, Contest number 525 has a solution word of length 4, "clen," which is not acceptable.

2. **Removal of anomalous player data:** Data with total scores or Hard Mode scores that are not in the same order of magnitude as the number of scores for each date in the week before and after or with a large discrepancy have been removed. For example, the total number of scores on November 30, 2022, was 2569, which is much lower than the scores on Hard Mode for the week before and after, which ranged from 22000 to 28000. This information is considered anomalous.

3. **Elimination of abnormal answer percentage data:** Due to rounding, the sum of different percentages of answers should fluctuate between 98% and 102%, which is a reasonable range. However, the sum of the percentages of different answer times for Contest number 281 is 126%, which is considered an outlier and has been removed.

1.4 Our Work

Our problem-solving framework is presented in Figure 1.

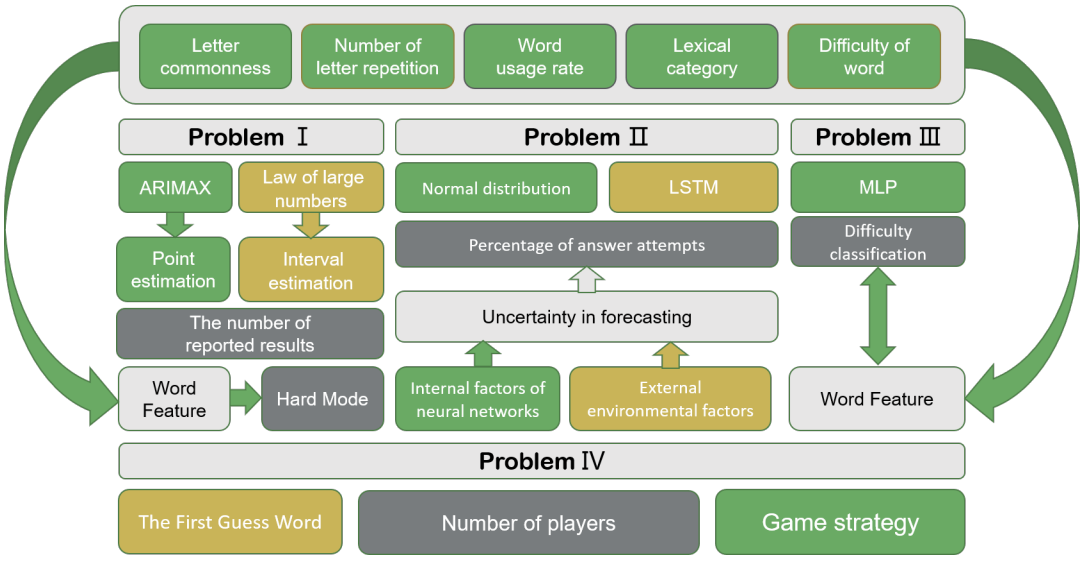


Figure 1: Overview of our work

2 Assumptions & Nomenclature

To simplify our modeling, we make the following assumptions:

Asm.1 The majority of users who share on Twitter account for a large portion of the total game users, ensuring that the reported results can serve as characteristics of the overall user population.

Asm.2 The scores shared on Twitter are users' actual scores, and they do not know the answer in advance, ensuring the accuracy of our classification of word difficulty.

Asm.3 There are no external factors, such as significant advertising and media promotion, or competition from similar games that significantly affect user numbers and characteristics, making user number fluctuations relatively stable and analyzable.

Asm.4 Wordle's words are completely random and do not cater to fixed holidays or current events, making our calculations of word frequency more accurate.

Asm.5 The game mode and rules will not change.

In this work, we use the nomenclature in Table 1 in the model construction. Other nonefrequent-used symbols will be introduced once they are used.

Table 1: Notations used in this literature

Symbol	Description
$letter\_fre_j$	Frequency of each letter in a word
$initial\_fre_j$	Frequency of each letter in the initial of a word
$word\_letter_i$	Commonness of each letter in a word
$letter\_repeat_i$	Number of letters repeated in a word
$word\_fre_i$	Word usage rate in daily life
$word\_class_i$	Lexical category
$word\_guesstimes_k$	Number of average times players solving the puzzle
$word\_precentage_k$	Percentage of each times
$word\_difficulty_i$	The difficulty of a word being guessed
y_t	Number of reported results per day
z_t	Number in Hard Mode per day
per_t	Percentage of Number in Hard Mode per day

3 Game Popularity and Mode Prediction Model

3.1 Analyze Word Features

Define the word feature as a five-dimensional vector as follows:

$$\gamma_i = (word\_letter_i, letter\_repeat_i, word\_fre_i, word\_class_i, word\_gesstimes_i) \quad (1)$$

where:

- $word\_letter_i$: Letter commonness
- $letter\_repeat_i$: Number of letter repetition
- $word\_fre_i$: Word usage rate
- $word\_class_i$: Lexical category
- $word\_difficult_i$: Word difficulty

1. Letter commonness

$word_letteri$ assesses the difficulty level of a word based on the frequency of its letters, to determine the complexity of a given word. The initial letter of a word can greatly affect the accuracy of a word guess, for example, when the initial letter has been guessed correctly, it is easier to guess the spelling of the entire word.Therefore, we increase the weight of initial letter commonness

$$word\_letter_i = \sum_{j=1}^{5} letter\_fre_j + initial\_fre_j \tag{2}$$

where:

- $letter\_fre_j$: Frequency of occurrence of each letter in the word
- $initial\_fre_j$: Frequency of each letter appearing in the first letter

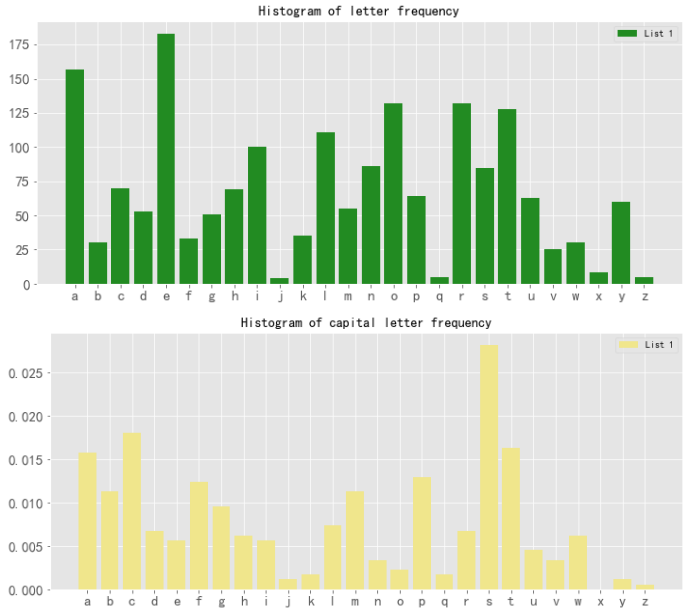


Figure 2: Histogram of letter freguency and capital letter frequency

2. Number of letter repetition

The frequency of letter repetition can affect a word's recognizability. For instance, the word "mummy" appears on October 23, 2022, where the letter "m" repeats three times while the other letters occur only once, these repetitions may occupy a larger portion of the word, thereby making the remaining letters more difficult to recognize.The definition of letter repetition is shown in the following table.

Table 2: The definition of $letter\_repeat_i$

Letter repetition	$letter\_repeat_i$	Example
No repetition	1	manly
Repeat one letter twice	2	goose
Repeat two letter twice	2	vivid
Repeat one letter thrice	3	mummy

3. Word frequency distribute

To determine the relevance of words to daily life, we have gathered a COCA word frequency table that lists the most commonly occurring words and phrases in contemporary American English. The table is organized in descending order of frequency, and we have categorized words into five levels of frequency based on this table.

Table 3: The definition of $word\_fre_i$

Word Ranking on COCA	$word\_fre_i$
0~3000	5
3000~6000	4
6000~9000	3
9000~12000	2
12000+	1

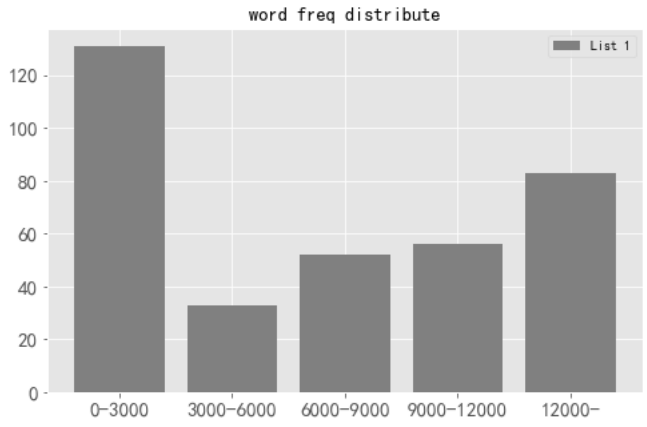


Figure 3: Word frequency distribute

4. lexical category

Furthermore, we have classified words based on their lexical properties, such as nouns, verbs, adjectives, adverbs, and other words. The values of $word\_class_i$ correspond to these different lexical properties, as shown in the table below.

Table 4: The definition of $word\_class_i$

lexical category	$word\_class_i$
Nouns	1
Verbs	2
Adjectives	3
Adverbs	4
Other words	5

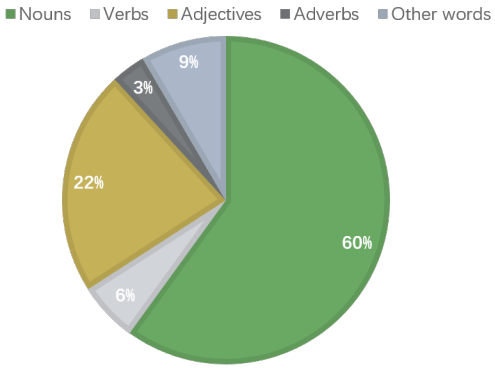


Figure 4: Lexical category distribute

5. Word difficulty

We denote the number of attempts required for a user to guess a word as the difficulty level, ranging from 1 to 7. Let $word\_precentage_k$ be the corresponding weight for each difficulty level. Then, we define the word difficulty score as

$$word\_difficult_i = work\_times_k \times word\_precentage_k \tag{3}$$

3.2 Point estimation of ARIMAX-based prediction

To better explain the variation in the number of reported players over time, we need to analyze the time series of known changes in player numbers and predict their future values. To accomplish this, we will build a time series forecasting model. First, we test the smoothness of the time series using the Augmented Dickey-Fuller test (ADF) and fit a p-order autoregressive (AR) model as follows:

$$r_t = \sum_{i=1}^{P} \alpha_i r_{t-i} + w_t \tag{4}$$

The number of roots of this equation is represented by the p_value, and the presence of a unit root is determined by checking whether p_value is less than the 0.05 confidence interval. The time series in question is subjected to the ADF test and the resulting p_value is 0.03, which indicates that the time series is smooth and does not have a unit root.

To predict future values for the smooth time series, we can use an Autoregressive Moving Average (ARMA) model, where the stationary series Y_t is described as ARMA(p,q) with white noise σ_t.[1]

$$Y_t = c + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} + \epsilon_t \tag{5}$$

While the ARMA model is efficient for simple forecasts, to improve its flexibility, we apply the autoregressive integrated moving average (ARIMA) model. This model can be adapted to different time series by adjusting the difference order, as well as the AR and MA orders, making it highly generalizable. The stationary series Yt is said to be ARIMA(p,q,d) if

$$Y_t = c + \phi_1 (Y_{t-1} - \mu) + ... + \phi_p (Y_{t-p} - \mu) + \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} \tag{6}$$

Since Wordle is an entertainment game, the number of players tends to increase during holidays and weekends, making it an important variable to consider. Including holidays in the model is likely to improve the prediction performance. Thus, we propose to extend the ARIMA model into an ARIMA model with explanatory variables (X), known as the ARIMAX(p,d,q).

$$Y_t = c + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} + \beta_1 X_{t-1} + ... + \beta_m X_{t-m} + \epsilon_t \tag{7}$$

If day t is a holiday, X_t is set to 1, otherwise it is set to 0.ϕ and θ are the autoregressive and moving average coefficients in the ARIMA model, and β is the external factor coefficient in the ARIMAX model.The residual sequence is shown in the Figure 5.

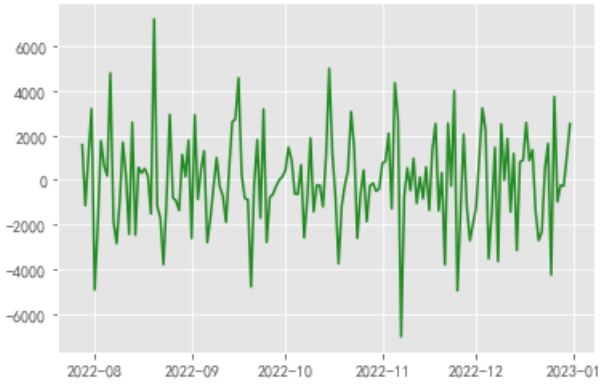


Figure 5: Residual series plot

As the construction of our model above, there are three tunable parameters i.e.,p,d,q. Figure 6 depicts the predicted and true value curves with the parameter settings of p = 4 ,d = 1 and q = 2.The total number of reported results predicted with ARIMAX are similar to the true values, therefore ARIMAX can be considered as a good fit.

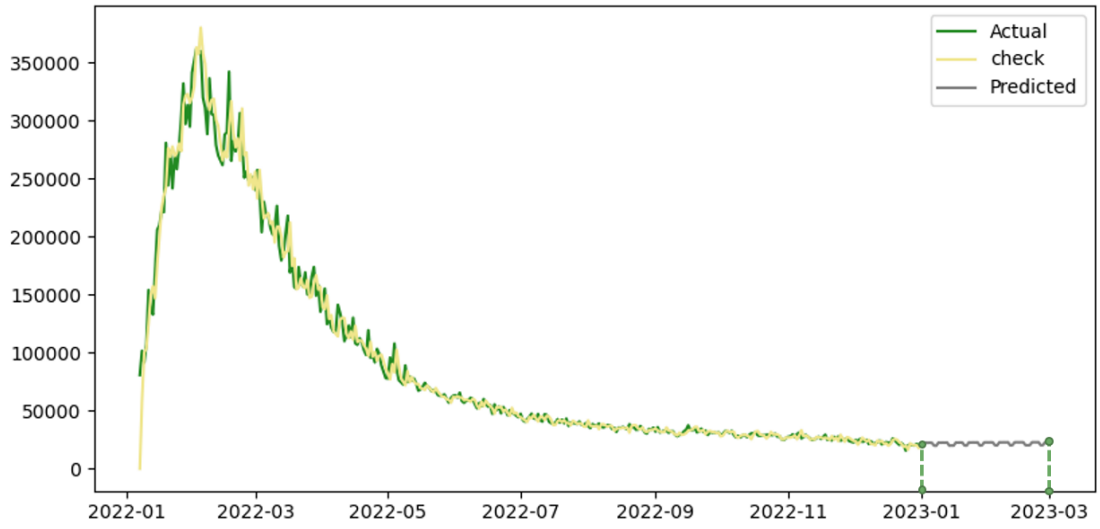


Figure 6: the Number of Reported Results Forecasting

3.3 Interval estimation based on the law of large numbers

In the above model of ARIMAX, we get a point estimate of the forecast value, denoted by y'_n, which corresponds to the forecast value y_t of the ARIMAX model at a particular date. To obtain the interval estimate of the predicted value and the corresponding confidence level, we give the interval corresponding to the predicted value at different confidence levels by deriving the distribution satisfied by the predicted value and using the mathematical properties of the normal distribution.

$y'_1, y'_2, \ldots, y'_n$ represent the individual possible values of the predicted total number of reported results at a particular date. When the time series sample is large enough, by the law of large numbers, the error ϵ_t of the past predicted values can be considered to have the same distribution as $y'_n - y_t$, as a normal distribution. By the central limit theorem, the possible values of the number of future Wordle players y'_n follow a normal distribution with mean y_t and variance $\sigma^2 = s^2(\varepsilon_t)$.

$$y'_s \sim N(y_t, \sigma^2) \tag{8}$$

For a variable y'_s that obeys a normal distribution, for a determined confidence level α, the confidence interval

$$[\mu - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}, \mu + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}] \tag{9}$$

shows y'_s will fall within the interval with the probability of α. where $z_{\alpha/2}$ denotes the quantile of the standard normal distribution. The choice of $z_{\alpha/2}$ instead of z_α is because this ensures that the two bounds of the confidence interval correspond to an area of (1 - α)/2, respectively, thus ensuring that the confidence interval is covered with probability α.

Set $\alpha = 90\%$,$\alpha = 95\%$,$\alpha = 99\%$ respectively, and get the prediction interval as Table 5.

Table 5: Confidence intervals at different confidence levels

α	$z_{\alpha/2}$	Lower Limit Lower	Upper Limit Upper
90%	1.645	22480.6636	23050.0756
95%	1.96	22426.1454	23104.5938
99%	2.575	22319.7052	23211.034

3.4 How Word Features Affect Hard Mode Numbers

Since Hard Mode can only be selected before the start of the game, the number of hardmode users is not related to the word feature and difficulty of the day. We believe that users' selection of Hard Mode on a given day may be influenced by the word features of the previous day.

To better investigate the relationship between the selection of Hard Mode and word features, we need to exclude the effect of the following factors on the number of people selecting Hard Mode.

- The change in the total number of users
- The change of time

For the first factor, the percentage of people choosing Hard Mode to the total number of people can be estimated by calculating the percentage of people choosing Hard Mode on that day, denoted as per_t:

$$per_t = \frac{z_t}{y_t} \times 100\% \tag{10}$$

For the second factor, the time trend of the percentage of people choosing Hard Mode can be calculated by the ARIMAX model in previous sub-question to obtain the value $\hat{per}_t$ predicted by the time trend only The residual series thus obtained:

$$\epsilon'_t = per_t - \hat{per}_t \tag{11}$$

excludes the effect of the above two factors. By plotting the heat map of the correlation between ϵ_t and the features of the words of the previous day, we can conclude that

- An excessive number of repeated letters in the words of the previous day leads to a significant decrease in the number of people choosing Hard Mode on the following day
- Less common words or words with lower letter frequency on the previous day will lead to a slight decrease in the number of people choosing Hard Mode on the following day
- The word nature of the previous day's words hardly affects whether users choose Hard Mode or not on the following day.

The obtained heat map is displayed in Figure 7.

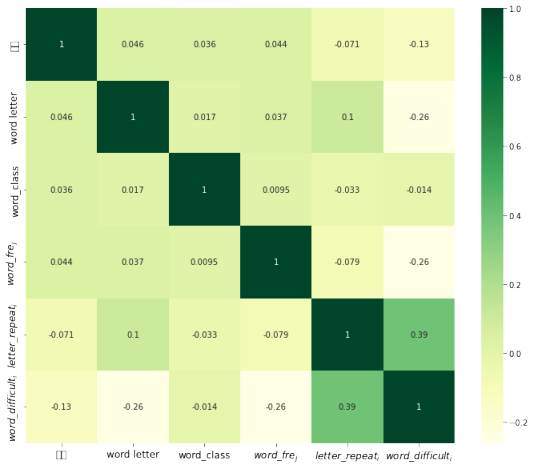


Figure 7: Correlation heatmap between word features and the percentage of Hard Mode players

4 Player Behavior Prediction Model

In this section, we investigate the relationship between word features and user behavior in the popular online game Wordle. Our analysis is based on daily data from Wordle, which has at least 20,000 users per day.

We used a hybrid approach to predict the probability distribution of the number of times a user solves a puzzle, combining statistical and machine learning, specifically, a normal distribution model and an LSTM neural network. And we give the percentages of (1, 2, 3, 4, 5, 6, X) for the word EERIE.We also analyzed the uncertainties of the model and calculated the confidence of the model with MC dropout.

4.1 Distribution Prediction Model Based on Distribution and LSTM.

By observing the daily distribution of (1, 2, 3, 4, 5, 6, X), we conjecture that it may obey a normal distribution. To verify the conjecture, the Kolmogorov-Smirnov normality test was applied to the historical data, and more than 80% of the p-values were obtained to be less than 0.05, so the original hypothesis was accepted and it was considered to obey a normal distribution with parameters (μ_t, σ_t^2).

$$\mu_t = \sum_{l=1}^{7} P_l \times l \tag{12}$$

$$\sigma_t^2 = \sum_{l=1}^{7} P_l \times (l - \mu)^2 \tag{13}$$

where l denotes the number of successful user attempts and takes values from 1 to 7. P_l denotes the percentage of the total number of people who succeeded on the lth answer.

For using LSTM and combining it with normal distribution to predict the word distribution, the specific steps are as follows:

1. Use LSTM to train the relationship between word features $word\_letter_i$,$letter\_repeat_i$, $word\_fre_i$, $word\_class_i$ and words (μ_t, σ_t^2) in historical data

2. Calculate the word features of a given word and predict its mean and variance using LSTM

3. Use Eq.
$$P(Y = l) = \frac{1}{\sigma \times \sqrt{2\pi}} \times exp(-(l - \mu)^2/(2\sigma^2)) \tag{14}$$
The predicted normal distribution probability function is discretized and the probability of the l guess of the word is calculated

4. Normalize (1, 2, 3, 4, 5, 6, X) so that their sum is 1

For LSTM neural networks, the specific principles are as follows:

Long Short-Term Memory (LSTM) introduces the memory cell, a unit of computation that replaces traditional artificial neurons in the hidden layer of a network. The memory unit works as shown in Figure 8, where X_t is the input of the current time step. h_{t-1} is the output of the previous LSTM unit, and C_{t-1} is the "memory" of the previous unit. For the output, h_t is the output of the current network, and C_t is the memory of the current cell. Then, multiple such memory cells are cascaded to achieve prediction of the data.[2]

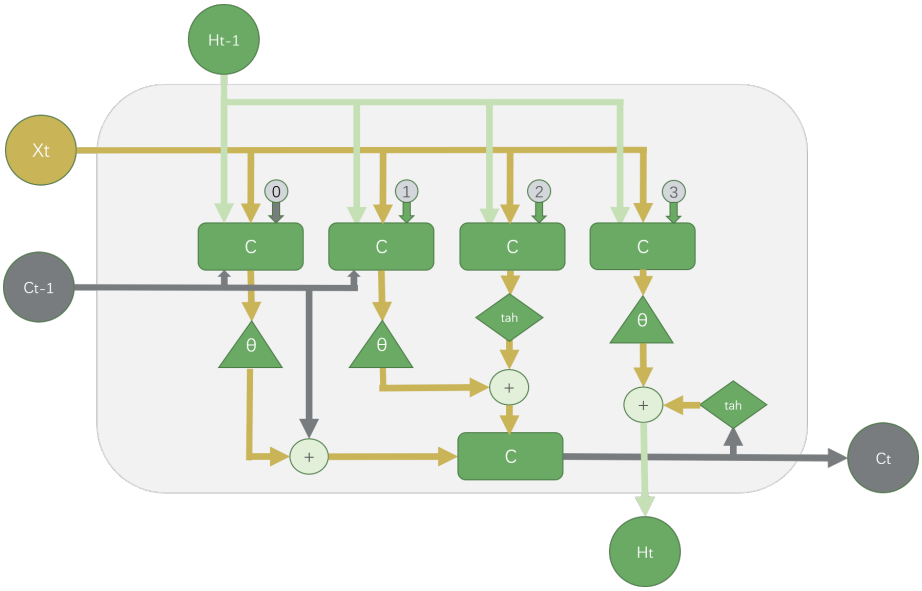


Figure 8: The memory unit of LSTM

With these memory cells, the model is able to grasp the structure of the data dynamically over time, with high predictive power.In our LSTM model for percentage prediction, A sequence is defined as y_t,z_t,$word\_letter_i$, $letter\_repeat_i$,$word\_fre_i$,$word\_class_i$.This describes the feature representation of words with sequential learning characteristics.

Our model consists of (1) a single input layer with the same number of memory units as the number of sequence learning features a sequence may hold, followed by (2) multiple LSTM layers and (3) a dense layer and (4) an output layer,used to output(μ_t, σ_t^2).

The above model was used to predict the word "EERIE" given by the question. We first used the ARIMAX model to predict $\hat{y}_t = 22765.37$ and $\hat{z}_t = 0.963$ and obtained the distribution of successful player answer attempts (1, 2, 3, 4, 5, 6, X) as (1.91%, 6.45%, 14.74%, 22.90%, 24.18%, 17.35%, 8.48%). As is shown in the Figure 9.

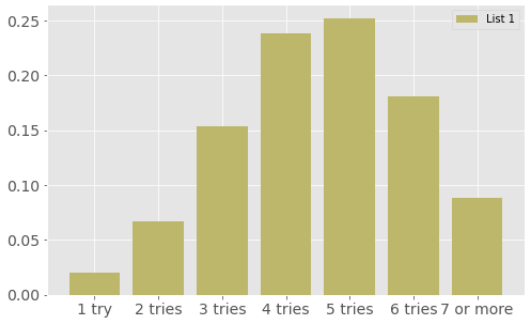


Figure 9: Percentage distribution histogram

4.2 Uncertainties Analysis

Considering a fixed number of users, daily keywords and successful attempts, the uncertainty of the model is divided into internal and external factors.

For the neural network LSTM used in this model, there are several factors that may increase the uncertainty of the model as follows:

1. **Data loading:** Neural networks need to load data from the training dataset during training, and usually random sampling is used to select samples, which can make the model more generalized. However, random sample selection can also bring uncertainty in the model.

2. **Weights and biases:** The weights and biases of a neural network are the key parameters of the model, and they control the expressiveness and learning ability of the model. During the training process, these parameters are randomly initialized and optimized and updated by the back propagation algorithm. The values of the weights and biases are usually somewhat random due to differences in the way the random initialization is performed and the optimization algorithm.[3]

3. **Data augmentation:** Data augmentation is a common data preprocessing technique that can be used to increase the amount and diversity of training data by randomly rotating, flipping, cropping, scaling, etc. This randomness can increase the robustness and generalization ability of the model.

At the same time, some external factors of the model may also directly or indirectly affect the percentage of successful attempts.

1. **Media Promotion:** When media promote the game, more novices tend to join the game. This may lead to a lower percentage of words being guessed in a smaller number of answers.

2. **Weather Factors:** When the weather is nicer, people tend to play casual games indoors and have more time to slowly deduce the game. This may result in a higher likelihood of the word being guessed that day.

3. **Current Political Factors:** When significant news events occur in society, people tend to check Twitter for news updates and share their Wordle scores of the day. This can result in an increase in the total number of people sharing on Twitter for that day.

4.3 Sensitivity Analysis

We use Monte Carlo Dropout (MC Dropout) to estimate the confidence of the neural network. specifically, during testing, we make multiple predictions with the model and randomly drop a certain percentage of neurons in each prediction. The mean of all predictions is used as the final prediction, while the variance of all predictions is used as the prediction uncertainty.

In the LSTM model, we set the number of hidden layers to 6, and the computed mean and variance are denoted as μ and σ^2, respectively.

The smaller the variance of the model, the smaller the uncertainty. In the sensitivity analysis, we analyze the uncertainty of the model when the parameters of the hidden layers have different values. Specifically, we can understand the uncertainty as follows: for a predicted value $\hat{y}$, the probability that the true value of $\hat{y}$ falls within the prediction interval with a confidence level of $1-\alpha$ is α.[4]

$$\hat{y} \pm z_{\alpha/2} \cdot \sqrt{\sigma^2 \cdot (1+\frac{1}{n}) + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \tag{15}$$

where $\hat{y}$ is the predicted value of the model, $z_{\alpha/2}$ is the $\alpha/2$ percentile of the standard normal distribution, σ^2 is the variance of the model, n is the sample size of the training data, x_i is the eigenvalue of the input data to be predicted, and $\bar{x}$ is the input mean eigenvalue of the training data.[5]

5 Word Difficulty Classification Model

5.1 Calculation of word difficulty coefficient

In the model establishment for the above-mentioned problem, we assume that the average number of user guesses, $word\_guesstimes_i$, represents the difficulty of the word, $word\_difficulty_i$. However, for a word that has not appeared in the historical records, we cannot calculate its difficulty based on the number of successful guesses. Therefore, we need to predict the word difficulty based on the word features, as a groundwork for the subsequent word difficulty classification.

In this section, we use the multilayer perceptron (MLP) to classify the difficulty of words. MLP is a simple artificial neural network commonly used in logistic regression and nonlinear

classification problems, consisting of an input layer, an output layer, and multiple hidden layers, each of which contains several nonlinear unit nodes (neurons). This is a forward structure artificial neural network (ANN), which can map a set of input vectors to a set of output vectors, overcoming the weakness of perceptron's inability to recognize linearly inseparable data.[7]

First, the network is initialized. The weight and bias values of each neuron are initialized using random initialization.The equations for initializing the weights and biases are as follows.

$$W_{i,j}^{(0)} \sim N(0, \sigma^2) \tag{16}$$

$$b_i^{(0)} = 0 \tag{17}$$

The weights $W_{i,j}$ connect the j-th neuron to the i-th neuron and are initialized according to a normal distribution with mean 0 and variance σ^2, denoted by $\mathcal{N}(0, \sigma^2)$. Additionally, b_i represents the bias value of the i-th neuron.

Then, the forward propagation is performed, and the input data starts from the input layer, passes through multiple hidden layers, and finally reaches the output layer. In each neuron, the weighted sum of the input signal needs to be calculated, and then it is transformed nonlinearly through an activation function to obtain the output value. The output value serves as the input value of the next layer of neurons. [8]The activation function introduces nonlinearity into the output of the neuron. For each neuron in the hidden layer, the output is calculated through the weighted sum and the activation function. For the neurons in the output layer, the output is also calculated through the weighted sum and the activation function,where the output $y_{i,j}$ of the jth neuron in the ith layer is calculated as

$$z_i = \sum_{j=1}^{n} W_{i,j} x_j + b_i \tag{18}$$

$$h_i = f(z_i) \tag{19}$$

In this process, $W_{k,j}$ represents the weight that connects the kth neuron in the $i-1$ layer to the jth neuron in the ith layer, and $y_{i-1,k}$ denotes the output of the kth neuron in the $(i-1)$th layer. b_j is the bias of the jth neuron, and f denotes the activation function. Typically, the activation function has the following types.

Table 6: Some Activation Functions

Activation Function	Function Formula
Sigmoid	$f(x) = \frac{1}{1+e^{-x}}$
Identity	$f(x) = x$
ReLU	$f(x) = max(0, x)$
Tanh	$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

Regarding the data of this problem, after multiple experiments, we created a new activation function called Tanh-ReLU to combine their prediction performance on words with different levels of difficulty. The mathematical formula of the Tanh-ReLU activation function is defined as follows:

$$f(x) = \begin{cases} x & x \geq 0 \\ \frac{e^x - e^{-x}}{e^x + e^{-x}} & x < 0 \end{cases} \tag{20}$$

In the subsequent sensitivity analysis, we will analyze the strengths and weaknesses of these activation functions, and explain the reason why we created the Tanh-ReLU activation function.

After forward propagation, the loss function is calculated using the mean squared error to measure the difference between the model's predicted output and the actual output. The formula for calculating the loss is:

$$L(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} l(y_i, \hat{y}_i) \tag{21}$$

where N represents the number of samples, y_i represents the actual output, and $\hat{y}_i$ represents the target output.

Next, during the backpropagation process, the gradient of the output layer error is calculated first, and then the gradient of the hidden layer error is calculated layer by layer in a forward manner. Finally, the weights and biases of the model are updated based on the gradient.

$$W_{i,j}^{(t+1)} = W_{i,j}^{(t)} - \eta \frac{\partial L}{\partial W_{i,j}} \tag{22}$$

$$b_i^{(t+1)} = b_i^{(t)} - \eta \frac{\partial L}{\partial b_i} \tag{23}$$

where h_i is the output of the i-th neuron, z_i is the input of the i-th neuron, $l(y_i, \hat{y}_i)$ is the loss for the i-th sample in the loss function, δ_i is the error term for the i-th neuron, $\frac{\partial L}{\partial h_i}$ is the partial derivative of the loss function with respect to h_i, and $\frac{\partial h_i}{\partial z_i}$ is the partial derivative of the activation function with respect to z_i.

Finally, iterative training is carried out by randomly dividing the training set into several small batches, and performing operations such as forward propagation, loss function calculation, backpropagation, and parameter update for each batch. Finally, the model for judging word difficulty coefficients is obtained.

In order to better categorize words by their difficulty, we divided the difficulty scores into three levels by observing the range of the given data for $word\_guesstimes_i$.

Table 7: Word difficulty Classification

Degree of difficulty	Range of $word\_guesstimes_i$
Easy	[2, 3.5]
Medium	[3.5, 4]
Hard	[4, 6]

We can compute their word features $\gamma_i = (word\_letter_i, letter\_repeat_i, word\_fre_i, word\_class_i)$ for a given word, and then use the neural network mentioned above to predict its difficulty level $\widehat{word\_difficulty_i}$. We can then categorize the word into one of three classes: easy, medium, or hard. For example, for the word "EERIE", we can calculate a difficulty score of 4.92, which would classify it as "hard".

5.2 Word attributes associated with difficulty classification

In order to obtain word attributes associated with each difficulty category, we performed Kernel Principal Component Analysis (KPCA) on the word features of each difficulty level. Standard PCA is a linear transformation that can only handle linearly related data. However, in this study, word features may have non-linear relationships, which requires non-linear PCA, i.e. KPCA. In KPCA, we first map the original data to a high-dimensional feature space, and then perform PCA in that feature space to obtain the eigenvalues and eigenvectors of the feature matrix.

The dataset in this study includes a total of $m = 355$ samples, $x_1, x_2, \ldots, x_m$. Each sample has $n = 4$ features: $word\_letter_i, letter\_repeat_i, word\_fre_i, word\_class_i$. We define a kernel function $k(x, y)$ to map the original data to a high-dimensional feature space. We then calculate the kernel matrix K that includes the similarity information of the original data in the high-dimensional space and center it, i.e. $K \leftarrow HKH$.

$$K_{ij} = k(x_i, x_j) \tag{24}$$

$$H = I - \frac{1}{m}11^T \tag{25}$$

In this approach, we use the unit matrix I and the all-ones vector 1 to eliminate translation and scaling effects in the kernel matrix K. After centering the resulting kernel matrix, we perform an eigenvalue decomposition to obtain the eigenvectors $v_1, v_2, \ldots, v_m$ and corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_m$. These eigenvectors form the principal components of the data in the high-dimensional feature space, with the corresponding eigenvalues indicating the importance of each principal component.

We select the top $k = 3$ eigenvectors to form a new transformation matrix $V_k = [v_1, v_2, \ldots, v_k]$ and map the original data x_i to a k-dimensional space.

$$\phi(x_i) = [v_1^T x_i, v_2^T x_i, \ldots, v_k^T x_i] \tag{26}$$

The mapped data is a vector composed of k eigenvalues, with each eigenvalue corresponding to an eigenvector that describes the main direction and magnitude of the data's variation in that direction.

By observing the differences in feature vectors of word groups with different difficulties, we summarize three word attributes associated with difficulty as follows:

Table 8: Words features associated with each difficulty level

Associated features	Easy	Medium	Hard
Primary associated feature	$word\_letter$	$word\_fre$	$word\_repeat$
Secondary associated feature	$word\_repeat$	$word\_letter$	$word\_letter$

For example, for the word EERIE, we can see that the features of $word\_repeat$ and $word\_letter$ are particularly significant and are related to "Hard" classification. The result is consistent with the findings of the second sub-question, indicating a certain level of reliability of our model.

5.3 Model Accuracy Evaluation

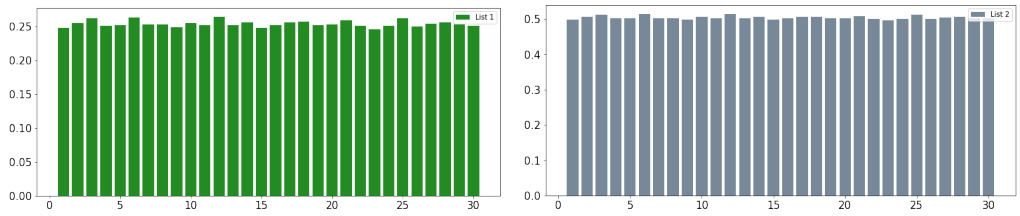
To evaluate the performance of the MLP model, the cross-validation method can be used. The data set is divided into several subsets, one for testing while the others are used to train the neural network. This process is repeated several times to obtain estimates of multiple performance metrics, and these estimates can be used to calculate the average performance of the neural network.

In this study, we define k = 5 for k-fold cross-validation. We first select 80% of the data for training the neural network, and then use the trained neural network to predict the remaining 20% of the data. The Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are calculated for the predicted values.

$$MSE = \frac{1}{n} * \sum (\widehat{y} - y)^2 \tag{27}$$

$$RMSE = \sqrt{\frac{1}{n}} * \sum (\widehat{y} - y)^2 \tag{28}$$

By repeatedly changing the dataset and recalculating the MSE and RMSE, the results shown in the figure indicate that both the MSE and RMSE are within the range of 0.24-0.26, indicating good model accuracy.



(a) MSE of the predicted value (b) RMSE of the predicted value

Figure 10: Predictive Value Evaluation

6 Interesting Features

6.1 Which word to choose as the first guess

Based on the data set,we counted the occurrences of the 26 letters in 5 positions and obtained the distribution of letters. If we start with the first letter and choose the letter with the highest frequency at each position, and choose the next lower frequency letter if the chosen letter appears in previous positions, we can form a 5-letter word "space". We can reasonably assume that if "space" is chosen as the first guessed word, more information can be obtained more quickly, assuming the word corpus used in Wordle remains the same.

6.2 What impact the acquisition of the New York Times bring about

Based on the dataset, it can be seen that the number of Wordle users increased rapidly in January and February 2022, reaching its peak. It was in January of that year that the New York Times acquired Wordle, and perhaps the New York Times, with its traffic and media promotion, attracted

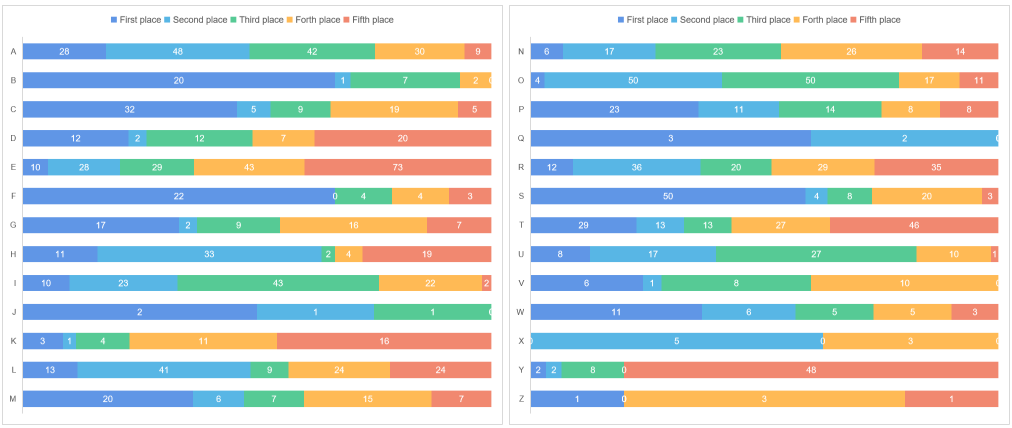


Figure 11: The frequency of letters in each position

a large number of Wordle players. However, the number of users rapidly decreased after February, possibly due to the addition of ads or fees after the acquisition by the New York Times, resulting in a significant loss of original users.

6.3 How players apply their intelligence

After analyzing the data of successful guesses with a success rate of over 1%, we compared the characteristics of these words with those of ordinary words in terms of $(word\_letter_i, letter\_repeat_i, word\_fre_i)$. The results show that words with higher values of $word\_letter_i$ and $word\_fre_i$ and lower values of $letter\_repeat_i$, i.e. higher Letter commonness and usage rate with fewer repetitions of letters, are easier for players to guess.

Next, we analyzed the characteristics of words that had a success rate below 10% after six or more attempts. The results show that words with lower values of Letter commonness and usage rate and higher values of letter repetition are harder to guess. In addition, we observed that the percentage of players who selected the Hard Mode gradually increased. This suggests that Wordle's loyal players, who are becoming more familiar with the game, are challenging themselves by selecting Hard Mode. Therefore, increasing the difficulty of Wordle is a trend in the future.

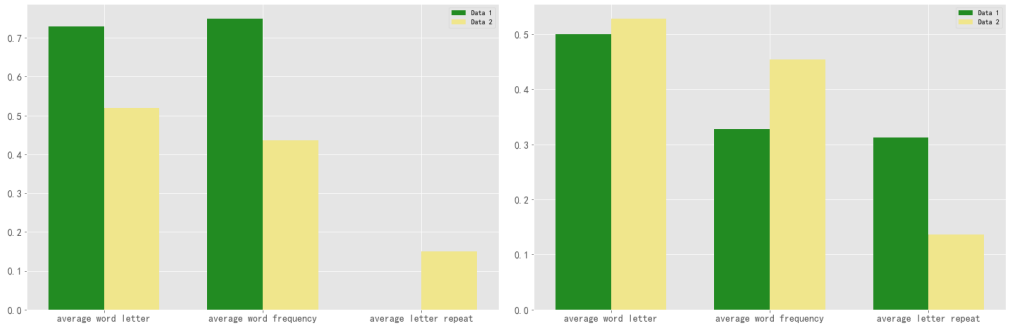


Figure 12: Comparison of word features

7 Sensitivity Analysis

1. ARIMAX model We set the three parameters p, d, and q of ARIMAX to (4,1,2), (2,1,0), (2,1,2), (7,1,4), and (7,1,2), and obtained their corresponding numerical results.

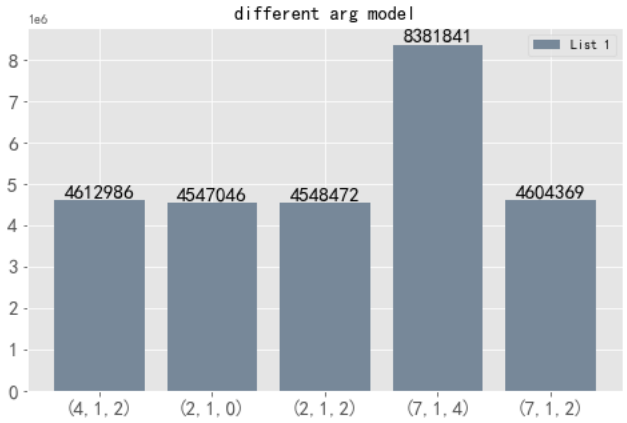


Figure 13: Sensitivity test results of ARIMAX with different parameters

2. LSTM model We used the dropout method to test it, and changed the number of layers in the LSTM neural network to 2-6 layers, and obtained their mean and variance as follows.

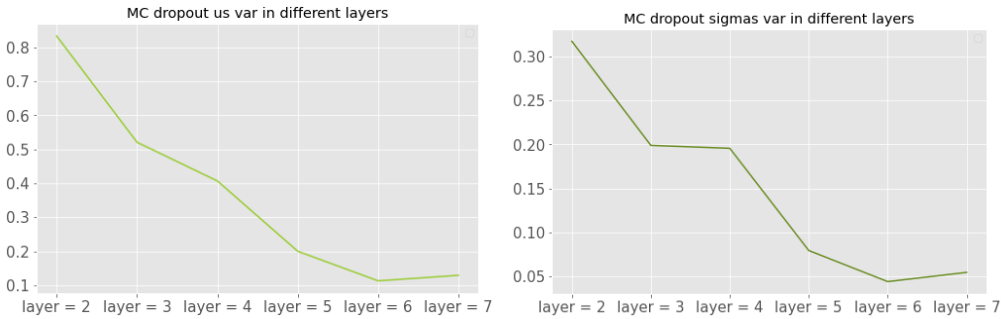


Figure 14: Mean and variance of dropout for different LSTM layer numbers

3. MLP model For the four candidate activation functions and the self-designed activation function of our model, we calculated the MSE and RMSE for their predictions on words of different difficulties, in order to compare their performance in predicting words of different difficulties.

   Five functions show different advantages in the errors of words with different difficulties. For simple words, tanh has the smallest error and the best performance. For difficult words, ReLU has the smallest error and the best performance. For all words, Tanh-ReLU has the smallest error and the best performance.

8 Strengths and Weaknesses

8.1 Strengths

1. **Proper use of methods.** We adapted the methods appropriately based on the characteristics of the problem. When using ARIMA, the ARIMA was extended to an ARIMAX model. When using LSTM, the mathematical distribution was predicted directly. When using WLP, the Tanh-ReLU activation function is created to reduce the error. When using PCA, the nonlinear characteristics of words are taken into account.

2. **Multiple evaluation methods.**For different models, we employed diverse evaluation methods to assess their effectiveness. For instance, we analyzed the uncertainty of LSTM using the dropout layer and examined the accuracy of WLP using cross-analysis.

3. **Detailed sensitivity analysis.**This model conducted detailed sensitivity analysis on the parameters of ARIMAX, the number of layers in the LSTM neural network, and the activation functions of MLP, demonstrating the model's strong robustness and mild sensitivity.

4. **Combining with mathematical principles.**This model integrates various statistical principles such as the law of large numbers, correlation, point estimation and interval estimation of parameters, and provides derivations and explanations of confidence intervals and model determinacy.

8.2 Weaknesses

1. **Insufficient feature extraction for words.**This model did not extract features based on whether a word letter belongs to a vowel or consonant, nor did it consider the impact of common letter combinations on the difficulty of guessing.

2. **The contribution of principal component analysis is not sufficient.**Due to the lack of distinct features among words, the contribution rate of the principal components decomposed by kPCA is not significant.

3. Due to the small size of the dataset, the neural network may suffer from **overfitting.**

9 Conclusions

Nowadays, the popularity of Wordle has been increasing rapidly. After The New York Times' acquisition of Wordle, a deep analysis of player behaviors and word difficulty design is of great significance for its subsequent marketing strategy and game operation.

In the paper, we conducted a detailed data cleaning process on the dataset. We then built an ARIMAX time series forecasting model a percentage prediction model based on the normal distribution and LSTM and a difficulty prediction model based on MLP. We also used kPCA to analyze the relationship between word attributes and difficulty. These help us to explore the important features and some interesting insights in the Wordle data, and we hope our findings could be of reference value for The New York Times to promote the better development of Wordle in the future.

10 A Letter to the the Puzzle Editor of the New York Times

Dear Editor,

We are delighted to seize such an unprecedented opportunity to share with you some of our group's research on Wordle, which is a word puzzle prevailing over the world. Based on the data provided by The New York Times, our research group has established a series of mathematical models to conduct a detailed and comprehensive analysis of various aspects of the Wordle game, resulting in a number of conclusions that we hope can be helpful to your work.

1. **The number of players remains relatively stable.**

   The number of players participating in the game is expected to increase slightly in the future, but overall there will not be significant changes. As the current popularity of Wordle is not as high as when it first appeared, those who continue to play the game are a subset of people who truly love it.The percentage of the total number of people playing the Hard Mode will increase slightly, probably because some players are used to the Hard Mode, and some players keep improving their level in the easy mode and seek higher challenges in the Hard Mode, and the percentage of the Hard Mode will also increase when the total number of people playing only increases slightly.

2. **The difficulty of the words was affecting the number of people choosing Hard Mode.**

   The effect of the word nature of the previous day's words on the percentage of people who chose the difficult mode on the following day was not significant, but the number of repetitions of letters in the words had a great effect on that percentage on the following day. It is possibly because if there were words with many repeated letters in the word the day before, which means such words are more difficult, it will lower the win rate and also motivate more people to challenge the more difficult mode.

3. **The difficulty of subsequent words can be determined based on the distribution of player response counts.**

   For the word "eerie", our model predicts that it has a 26% chance of being guessed on the fifth time, a 23% chance of being guessed on the fourth time, only a 1% chance of being guessed on the first time, and an 8% chance of not being guessed. This is because people tend to guess a word with different letters for each letter the first time to maximize the number of letters present or absent, so eerie, with its 3 e's, resulted in few people being able to guess it the first time. Also, it is often not easy to guess words that contain repeated letters, because once people get a letter present in the word through the first few tries, they tend to ignore its repetition and are apt to guess other letters, so this also leads to a high difficulty in the word "eerie", which a large percentage of people cannot guess.

If this the conclusion of our research are helpful to your marketing management or blueprint planning, we would be really glad. And if you have any questions, please feel free to contact us.

Sincerely

Team #2308240 in 2023 MCM

References

[1] Kongcharoen, Chaleampong & Kruangpradit, Tapanee. (2013). Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX) Model for Thailand Export.

[2] https://medium.com/mlreview/understanding-lstm-and-its-diagrams-37e2f46f1714

[3] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human Trajectory Prediction in Crowded Spaces," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, jun 2016, pp. 961–971.

[4] T. Fushiki, "Estimation of prediction error by using k-fold crossvalidation," Statistics and Computing, vol. 21, no. 2, pp. 137–146,2011.

[5] J. Schlechtriemen, A. Wedel, J. Hillenbrand, G. Breuel, and K.-d.Kuhnert, "A lane change detection approach using feature ranking with maximized predictive power," in 2014 IEEE Intelligent Vehicles Symposium Proceedings. IEEE, jun 2014, pp. 108–114.

[6] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. arXiv preprint arXiv:2102.05918, 2021.

[7] J. Schlechtriemen, A. Wedel, J. Hillenbrand, G. Breuel, and K.-d.Kuhnert, "A lane change detection approach using feature ranking with maximized predictive power," in 2014 IEEE Intelligent Vehicles Symposium Proceedings. IEEE, jun 2014, pp. 108–114.

[8] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.