Is Wordle Puzzle a True Puzzle?

Summary

In 2021, Wordle became a trendy online word game. Players enjoy this challenge of guessing the five-letter word in six attempts or less. Now, we need to analyze the variation in the number of scores and predict this variation and distribution of the number of tries.

Firstly, we conducted an exploratory analysis of the dataset and found that the number of people who pass the game at 1, 2, and 3 tries is negatively related to the number of people who pass the game at 4, 5, 6, 7, or more tries. Therefore, we can consider four tries as a dividing line. People below this line probably master some skills in this game. Meanwhile, analyzing the number of attempts to pass this game, we develop a word difficulty classification model based on the topsis-entropy weight method. Finally, about 3/4 of the words are of medium difficulty.

Besides, we built a continuous prediction model based on the ARIMA model to predict the number of reported results on March 1, 2023. For more accurate predictions, we use the average time series sliding window value of 10 as the basic sequence. We observed that the goodness of fit of the finally constructed ARIMA (2,1,0) is about 0.97. Therefore, if the New York Times does not improve, the reported results will fall to 22661-23978 **on March 1, 2023**. At the same time, we use one-hot encoding to extract the attributes of the words. Then, we use the least squares method to regress and find that the F test of the proportion of the number of hard modes and the attributes failed, which means there is no relationship between them.

Next, we use the relevant percentage of (1,2 3,4,5,6, X) as output and word attributes as input to construct LGBM regression models and multiple output regression models through the regression chain function provided by sklearn. The MAPE of the model in the training set is 0.53%, and the MAPE performance in the test set is 0.41%. Its uncertainty is 0.53%, and we are 99.59% confident that the model has no problem. The model-fitting effect is excellent. According to the model, if the solution word on March 1, 2023, is ERRIE, the relevant percentage of (1,2 3,4,5.6, X) are 0.23, 3.52, 18.34, 31.25, 27.51, 15.74, 2.87.

Finally, we use the word difficulty coefficient and divide the index into easy, medium, and difficult as the dependent variable and the attribute of the word as the independent variable. Based on the Decision Tree model, we construct a model to identify the difficulty level of a given the word. The F1 of the model reaches 1 after approximation; therefore, it is excellent. Based on the model, the probability of the word ERRIE being easy is 0.08, being medium is 0.75, and being difficult is 0.17. Hence, this word belongs to the medium difficulty.

Keywords: ARIMA, LGBM, Decision Tree, MAPE.

Contents

1. Introduction

1.1 Problem Background

As we live in a fast-paced world, people have less and less time to learn how to access and play a very complex game. Meanwhile, hard-to-play games also mean that only a minority of people play them. Therefore, the market for little quick games is void. Jonathan Feinberg found this gap and launched the game Wordle at about the end of the year 2021,

Wordle is a simple but addictive game where the player has six chances to guess a five-letter word. Each time the player enters a guess, the game gives feedback on how many letters are in the correct position and how many letters are correct but in the wrong position. Wordle soon became a viral sensation among people of all ages. At the same time, Wordle has also sparked social media trends. People share their results and strategies on platforms like Twitter. Because of its popularity, some users have even created tools and scripts to help them guess the correct word more efficiently.

Of course, creating tools and scripts to help guess the daily solution word results in losing the purpose and happiness of playing this game. Nevertheless, it is worth making a model to explain the number of reported results variations and the distribution of the number of tries on each date. Therefore, for this Wordle Puzzle, our strategy is based on the time series to analyze and forecast data points collected over time.

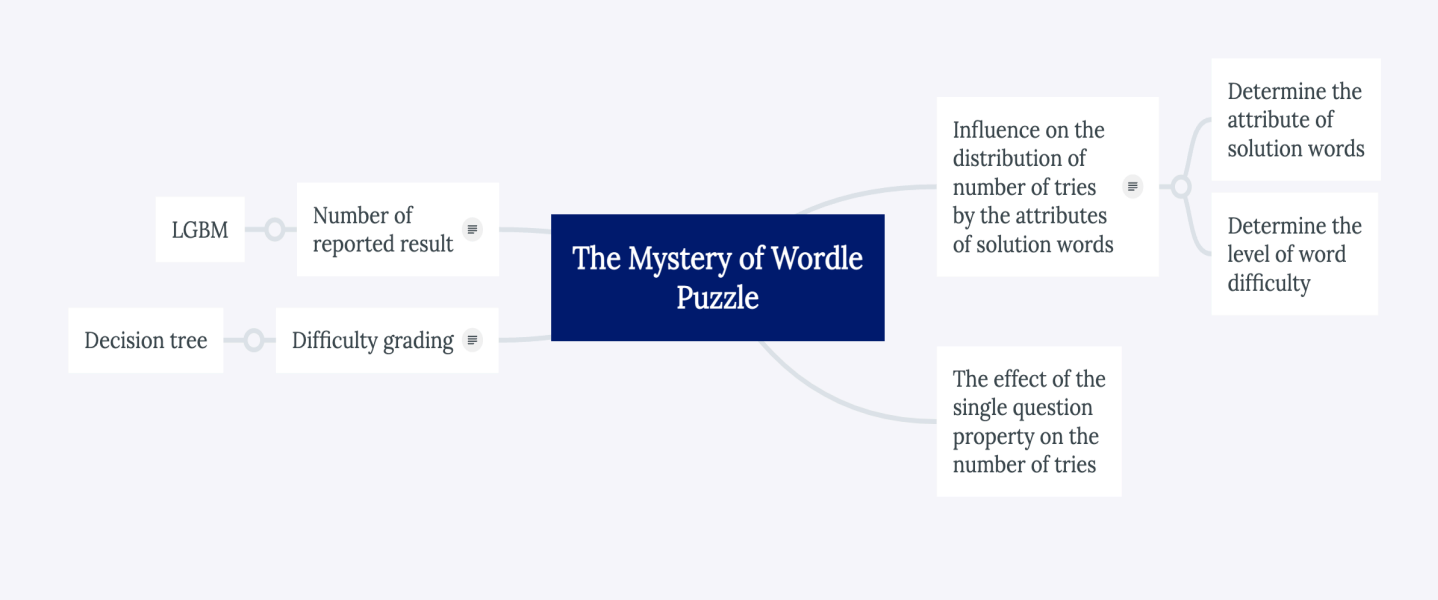
1.2 Restatement of the Problem

Considering the background information and the rule of the game, there are problems we need to solve:

- Given the data including date, the solution word, number of reported results, the number of scores on hard mode, and the percentage of players in 1, 2, 3, 4, 5, 6, 7 or more tries in each date, using ARIMA to create a model to explain the variation of daily scores and therefore to make predictive the scores on March 1, 2023.
- Given the attributes of word in each date, whether it will influence the percentage of scores played on hard mode.
- Given a future solution word (e.g., ERRIE), create a model to predict the distribution of 1, 2, 3, 4, 5, 6, 7 or more tries. Demonstrate its performance.
- Create a model to classify the difficulty of solution word (e.g., ERRIE)

1.3 Flow Chart

The flow chart is shown in Figure 1.



2. Assumptions and Justifications

Assumption 1: The reported results are independent and identically distributed.
Justification: In the real world, weather, exam, or any emergent things may result in people needing more time to play Wordle or worse performance in guess word. Therefore, to simplify the circumstances, we assume that no external factors influence the number of reported results, the number of hard mode scores, and the distribution of 1, 2, 3, 4, 5,6, 7, or more tries distribution,

Assumption 2: Assume Wordle will not be influenced by any social factors and will keep running.
Justification: In the real world, everything is unpredictable. Therefore, it is essential to assume that no external factors could influence the regular operation of Wordle.

Assumption 3: Assume the existence of a data error and revise it.
Justification: According to the rule of Wordle, players need to solve the puzzle by guessing a five-letter word in six tries or less. Therefore, we can conclude that the solution words whose length is not equal to five are data entry errors. Hence, we revise them by either searching for the solution word at that date or as what they fit most.

3. Notations

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

Table 1: Notations used in this paper

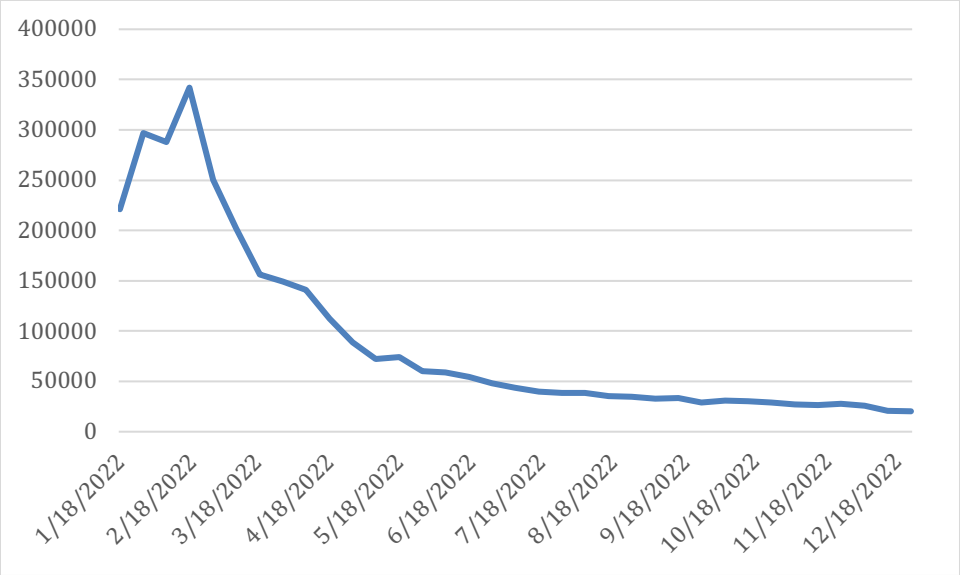
Symbol	Description
MAPE	Mean absolute percentage error
P	Significance
S-W	Shapiro-Wilk Test
p	Order of the autoregression model
q	Order of moving-average model
d	degree of differencing

4. Data Preprocessing

4.1 Average value of sliding window of reported results

To begin with, we made a graph of the reported results of each date. However, for more accurate predictions, we use the average value of the time series sliding window of 10 as the basic sequence. Therefore, we can scale the data (Dataset 1) to a smoother line by getting the average value every ten days, shown in Figure 2.

Figure2: Number of Reported Results



4.2 Analysis of Attributes of Solution Words

To facilitate the classification of word difficulty and the correlation impact analysis of words on the number of people in difficult mode, we must extract the attributes of words, such as the common degree of words, spelling, and part of speech. However, some feature information

extraction is complicated, so we choose a one-hot encoding for analysis. We constructed the vector of every word; every vector dimension represents one word. For example, the word list [“happy,” “sad,” “angry”] after doing one-hot encoding is [100, 010, 001].

Based on this method, we did one-hot encoding for every given the word (Dataset 2)

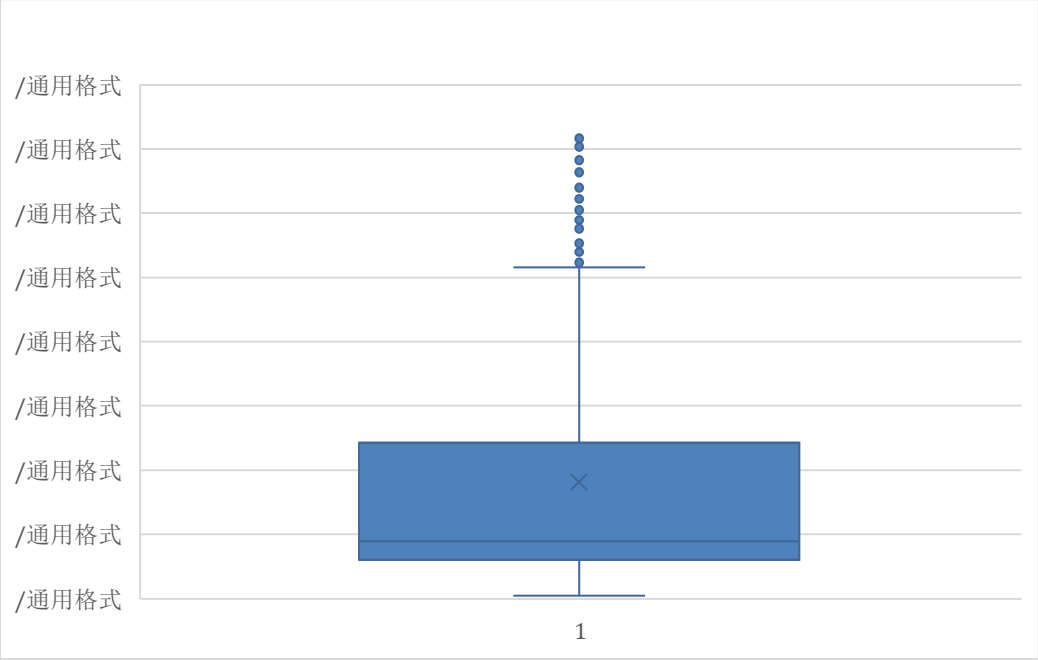
Dataset 2: One-hot Encoding for Every Word

	a	b	…	x	y	z	Word
0	0	0	…	0	0	0	slump
1	1	0	…	0	0	0	crank
2	0	0	…	0	0	0	gorge
3	0	0	…	0	1	0	query
4	0	1	…	0	0	0	probe
	…	…	…	…	…	…	…
355	0	0	…	0	0	0	chord
356	1	0	…	0	0	0	taper
357	1	0	…	0	0	0	slate
358	0	0	…	0	0	0	third
359	1	0	…	0	0	0	lunar

4.3 Exploratory Analysis

Figure 2 we made above shows the number of reports changing over time. It was found that the game continued to decline after reaching its peak in February and March. Finally, the data converged, gradually stabilizing at around 20,000-30,000. The boxplot below, Figure 3, also shows the relatively discrete distribution between 50,000 and 150,000.

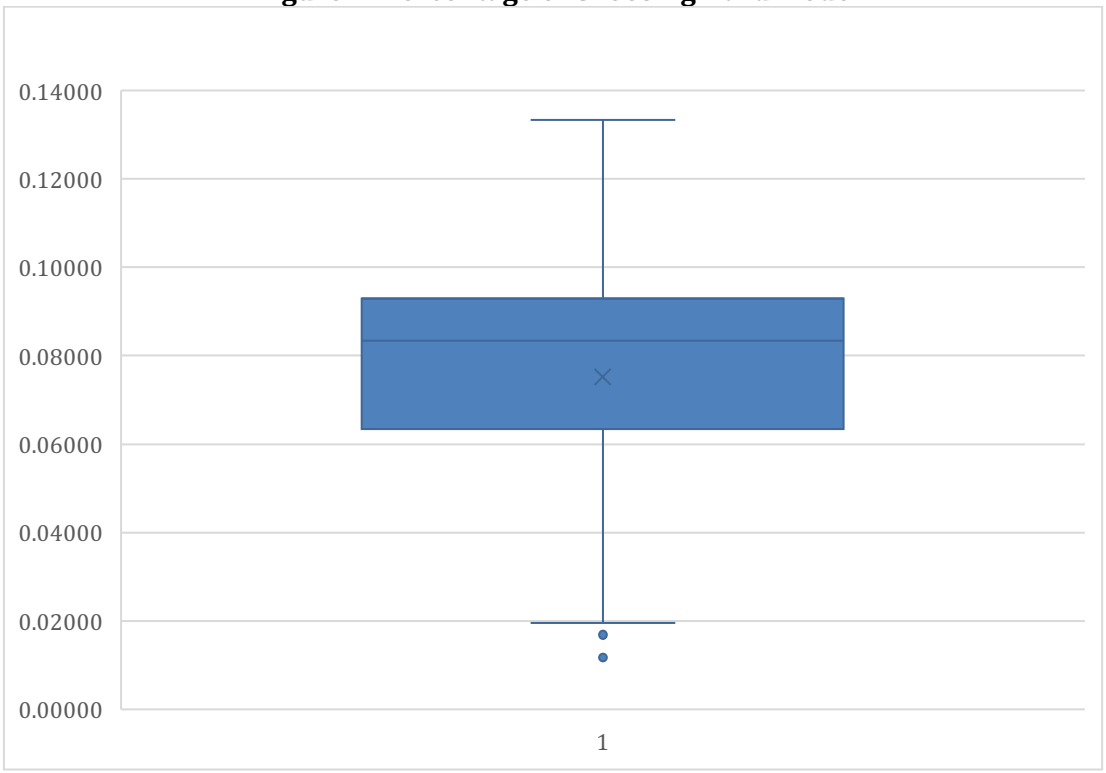
Figure 3: Number of Reported Results



4.4 Analysis of the Distribution of Hard Mode

After dividing the number of people choosing hard modes by the number of reported results, we got the percentage of hard mode and made a boxplot to demonstrate its result. However, we found one outlier, the percentage of hard mode over the number of reported results. After removing this outlier, we get the boxplot of the percentage of choosing hard mode shown in Figure 4. Finally, we observed that the maximum is about 0.137 to 0.138. The minimum number is about 0.015 to 0.016. The medium is about 0.081 to 0.082. Its IQR is between 0.061 to 0.093.

Figure 4: Percentage of Choosing Hard Mode



4.5 Topsis-Entropy Weight Method to Evaluate Difficulty Level

4.5.1 Spearman's Rank Correlation Coefficient

To evaluate the difficulty level of solution words, it is possible to achieve it by analyzing the distribution of the relevant percentage of 1, 2, 3, 4, 5, 6, 7or more tries. In the table below, we showed the result of Spearman's rank correlation coefficient. We can see that the distribution of 1, 2, 3, 4, 5, 6, 7, or more tries significantly correlated. However, since the 4 tries, it shows a significant negative correlation. We need to find a line to differentiate between positive and negative indicators. Therefore, for this model, we take 1-3 as negative indicators; a higher percentage of negative indicators means the word is easier. By contrast, we take 4-X as positive indicators; a higher percentage of positive indicators means the word is harder. The result is shown below in Table 2.

Table 2: Spearman's Rank Correlation Coefficient

	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (X)
1 try	1(0.000***)	0.54(0.000***)	0.36(0.000***)	-0.34(0.000***)	-0.42(0.000***)	-0.22(0.000***)	-0.11(0.025**)
2 tries	0.54(0.000***)	1(0.000***)	0.843(0.000***)	-0.116(0.028**)	-0.84(0.000***)	-0.67(0.000***)	-0.49(0.000***)
3 tries	0.364(0.000***)	0.843(0.000***)	1(0.000***)	0.264(0.000***)	-0.91(0.000***)	-0.91(0.000***)	-0.76(0.000***)
4 tries	-0.34(0.000***)	-0.12(0.028**)	0.26(0.000***)	1(0.000***)	-0.05(0.340)	-0.51(0.000***)	-0.62(0.000***)
5 tries	-0.42(0.000***)	-0.84(0.000***)	-0.91(0.000***)	-0.05(0.340)	1(0.000***)	0.78(0.000***)	0.56(0.000***)
6 tries	-0.23(0.000***)	-0.68(0.000***)	-0.91(0.000***)	-0.51(0.000***)	0.78(0.000***)	1(0.000***)	0.91(0.000***)
X tries	-0.12(0.025**)	-0.50(0.000***)	-0.76(0.000***)	-0.62(0.000***)	0.555(0.000***)	0.906(0.000***)	1(0.000***)

P.S. ***、**、*means 1%、5%、10% significance level

X tries means 7 or more tries

4.5.2 Topsis-Entropy Weight Method Creation and Solution

Our first step is to construct normalization matrix after trending original data. Then, we calculate the difference of evaluation objects by calculating the difference between the optimal vector and the worst vector.

Construct n x m matrix X_{ij}, which means that value of the index j of the i object.

$$Zij = \frac{X_{ij}}{\sqrt{\sum_{i}^{j}(X_{ij})^2}}$$

Calculate the distance between each rating index and the best and worst vectors.

$$D_i^+ = \sqrt{\sum_{j=1}^{m} \omega_j \ (Z_j^+ - z_{ij})^2}\ , D_i^- = \sqrt{\sum_{j=1}^{m} \omega_j \ (Z_j^- - z_{ij})^2}$$

Calculate the gap between each evaluation index and the best and worst vectors.

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-}$$

Next, we normalize each factor by the number of each option and calculate the entropy value of the j-index. Finally, calculate the difference of information entropy redundancy and the weight of each indicator in Table 3.

Table 3: Entropy Method

	e	d	Percentage (%)
1 try	0.998	0.002	8.05
2 tries	0.995	0.005	21.48
3 tries	0.990	0.010	48.30
4 tries	0.993	0.007	3.34
5 tries	0.990	0.01	5.85
6 tries	0.996	0.004	12.98

Finally, we get the overall rating of words and classify them by tri-sectional quantiles. Some of the results are shown in Table 4.

Table 4: Overall Rating

Solution Words	（D+）	（D-）	Overall Rating	Index
manly	0.55326555	0.54193324	0.49482637	104
molar	0.62795729	0.48310338	0.43481278	169
havoc	0.54587539	0.55508398	0.50418207	92
poise	0.78729515	0.28877189	0.26835865	322
aorta	0.66148694	0.41733996	0.38684608	223

5. Number of Reported Results Prediction

5.1 Establishment and solution of time series prediction model based on ARiMA

Time series is to record the process of random events changes based on chronological order. Observing, researching, and discovering the rule of time series transformation and predicting its future trends is the key to time series analysis. The time series contains three models:

AR(p) model:

$$\begin{cases} X_t = \emptyset_0 + \emptyset_1 x_{t-1} + \emptyset_2 x_{t-2} + \cdots + \emptyset_p x_{t-2} + \varepsilon_t \\ \emptyset_p \neq 0 \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = {\sigma_\varepsilon}^2, E(\varepsilon_t, \varepsilon_s) = 0, s = t \\ Ex_s\varepsilon_t = 0, \forall s < t \end{cases}$$

MA(q)model

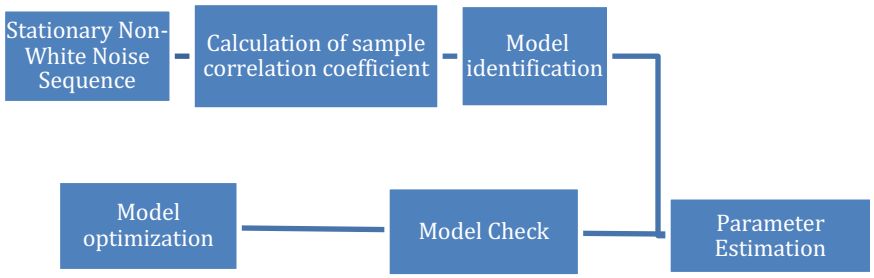
$$\begin{cases} X_t = \boldsymbol{\mu} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} \ldots - \theta_q\varepsilon_{t-q} \\ \theta_q \neq 0 \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = {\sigma_\varepsilon}^2, E(\varepsilon_t, \varepsilon_s) = 0, s = t \end{cases}$$

ARMA (p, q) model

$$\begin{cases} X_t = \emptyset_0 + \emptyset_1 x_{t-1} + \emptyset_2 x_{t-2} + \cdots + \emptyset_p x_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} \ldots - \theta_q\varepsilon_{t-q} \\ \theta_q \neq 0, \emptyset_p \neq 0 \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = {\sigma_\varepsilon}^2, E(\varepsilon_t, \varepsilon_s) = 0, s = t \\ Ex_s\varepsilon_t = 0, \forall s < t \end{cases}$$

A model with the above structure is called a p-order autoregressive model, recorded as ARMA (p, q). The steps of stationary sequence modeling are as Figure 5

Figure 5: Stationary Sequence



Then calculate the sample correlation coefficient:

$$\widehat{p} = \frac{\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t-1}^{n}(x_t - \bar{x})\text{^}2}$$

Sample Partial Autocorrelation Coefficient:

$$\widehat{\emptyset_{kk}} = \frac{\widehat{D_k}}{\widehat{D}}$$

Basic Principles of Model Identification:

$\widehat{p_k}$	$\widehat{\emptyset_{kk}}$	Model Chosen
smear	P-order smear	AR(p)
Q-order smear	smear	MA(q)
smear	smear	ARMA (p, q)

Approximate Distribution of Sample Correlation Coefficients:
Barlett:
$\widehat{p_k} \sim \mathrm{N}(0,1/\mathrm{n}), \mathrm{n} \to \infty$
Quenouille:
$\widehat{\emptyset_{kk}} \sim \mathrm{N}(0,1/\mathrm{n}), \mathrm{n} \to \infty$

Then, because A good fitting model should be able to observe almost all sample-related information in the value sequence in advance, that is, the residual sequence should be a white noise sequence.

For this reason, we made Hypothesis:

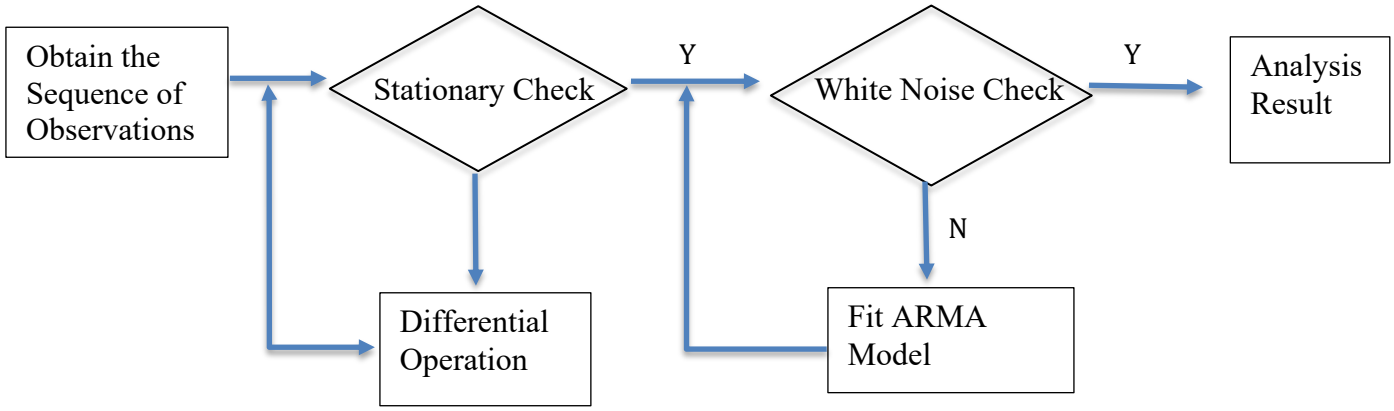
$$H_0: \ \rho_1 = \rho_2 = \cdots = \rho_m = 0, \forall m \geq 1$$
$$H_1: \ at\ least\ exists\ a\ \rho_k \neq 0\ , \forall m \geq 1, k \leq m$$

$$LB = n(n+2)\sum_{k=1}^{m}\left(\frac{\widehat{p_k}^2}{n-k} \sim x\text{^}2(m)\right.$$

Besides, parameter significance test is required. We are to test whether each unknown parameter is significantly nonzero, and deleting insignificant parameters is the most streamlined model structure. Then, we suppose:

$$H_0: \ \beta_j = 0,$$
$$H_1: \ \beta_j \neq 0$$

$$T = \sqrt{n-m}\,\frac{\widehat{\beta_j}-\beta_j}{\sqrt{a_{jj}Q(\tilde{\beta})}} \sim t(n-m)$$



5.2 Explanation of the Model:

ARIMA has three steps to the solute. Firstly, require the sequence to satisfy the stationarity, check the ADF test results, and analyze whether it can significantly reject the hypothesis that the sequence is unstable according to the analysis t value.

Secondly, analyze the comparison chart of the data before and after the difference to judge whether it is stable (the fluctuation range is not large), and at the same time, perform partial (autocorrelation analysis) on the time series and estimate its p and q values according to the censored situation.

Thirdly, the ARIMA model requires the model to have pure randomness; that is, the residual error of the model is white noise. Check the model test table and test the simulated white noise according to the P value of the Q statistic.

Based on these steps, we made Table 5 a Model Evaluation Table.

Table 5: Model Evaluation Table

	Coefficient	σ	t	P>|t|	0.025	0.975
Constant	-4195.17	7356.47	-0.58	0.57	-18613.58	10233.25
AR. L1	1.32	0.179	7.391	0	0.97	1.67
AR. L2	-0.702	0.185	-3.802	0	-1.063	-0.34

Hence, the Model Equation is:

$$y(t) = -4195.168 + 1.32 * y * (t-1) - 0.702 * y * (t-2)$$

Next, we created 6-time units and made predictions, See Table 6.

Table 6: Time Series Prediction

Time	Prediction Values
1	22697
2	22887
3	23036
4	23154
5	23247
6	23320

Therefore, the number of reported results on March 1, 2023, is about 23320 and the prediction interval is [23971, 22667].

6. Word Attributes and the Number of Hard Mode

6.1 Regression Based on Least Square Method

We extract all the encoding attributes at that time through word attribute analysis through the hot-encoding method. Then we can use the percentage of people who signed up for the hard mode as the dependent variable and the encoding attribute of the word as the independent variable to perform linear regression analysis.

The Least Square method minimizes the sum of squared errors between the actual and predicted values. Through Linear Regression, we made Table 7 to demonstrate the analysis result. To evaluate this model, we use the F test to determine the existence of a significant linear relationship and R2 to determine the overall fitness of this model.

Table 7: Result of Linear Regression

	t	P	VIF	R^2	Adjusted R^2	F
Constant	0.214	0.830	-			
a	0.208	0.835	39.663			
b	0.234	0.815	11.102			
c	0.263	0.793	24.357			
d	0.831	0.407	18.596			
e	0.274	0.784	46.157			
f	0.294	0.769	14.47			
g	0.234	0.815	19.315			
h	0.052	0.959	22.086			
i	0.279	0.780	28.809			
j	0.426	0.670	2.422			
k	0.219	0.827	13.561			
l	0.194	0.846	38.624			
m	0.235	0.814	20.982	0.067	-0.006	F=0.92 P=0.58
n	0.131	0.896	26.61			
o	0.093	0.926	39.846			
p	0.209	0.835	20.224			
q	-0.215	0.830	2.879			
r	0.131	0.896	32.935			
s	0.432	0.666	27.421			
t	0.502	0.616	36.289			
u	0.67	0.503	21.658			
v	0.422	0.673	10.103			
w	0.194	0.847	10.491			
x	0.377	0.706	3.83			
y	0.89	0.374	19.563			
z	0.548	0.584	2.802			

Based on the F-test, the P-value is 0.58 which is greater than 0.05. Therefore, we cannot reject the null hypothesis, which means that the attributes of solution words have no relationship with the percentage of people who register hard mode.

7. Solution Words Analyzation

7.1 Index Creation

After the model is trained, the entire test set can be predicted, and then the predicted number of reported results can be compared with the actual number of reported results. Based on the observed predicted and actual values, we used MAPE (Mean Absolute Percentage Error) as the evaluation index.

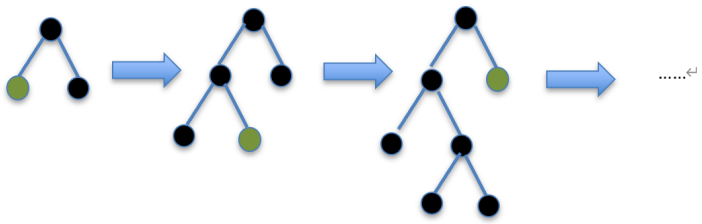
$$MAPE = \frac{1}{N}\sum_{i}^{N}\left|\frac{(y' - y_i)}{y_i}\right|$$

y' is the predicted value, y_i is the actual value, and N is the number of samples. Based on this equation, we can conclude that with a smaller MAPE, the error between the predicted number of reported results and the actual value of the number of reported results is smaller. Therefore the performance of this model is better.

7.2 Solution Words Analyzation Based on LGBM

We chose to construct a model based on the decision tree algorithm of LightGBM (See figure 5). A decision tree algorithm is a method of machine learning. It is a tree structure (either a binary tree or non-binary tree) in which each internal node represents a judgment on an attribute, and each branch represents the output of a judgment result. Finally, each leaf node represents a classification result.

In our model, we decide to discretize the continuous point eigenvalues into n integers and make a histogram. When traversing the date, the statistics are accumulated in the histogram according to the discretized values. Then, according to these discrete values of the histogram, traverse to find the optimal segmentation point. **Figure 5: Leaf-Wise Tree Growth**



Using the LGBM algorithm supports categorical features, which is one essential characteristic that most other algorithms can't support. LightGBM also adopts the leaf-wise growth strategy, and each time finds a leaf with the largest split gain (generally the largest amount of data) from all the current leaves, and then splits, and so on. Therefore, compared with Level-wise, Leaf-wise can reduce errors and obtain better accuracy when the number of splits is the same. The disadvantage of Leaf-wise is that it may grow a relatively deep decision tree, resulting in overfitting. Therefore, LightGBM adds a maximum depth limit Leaf-wise to prevent overfitting while ensuring high efficiency.

7.3 Explanation of the Model

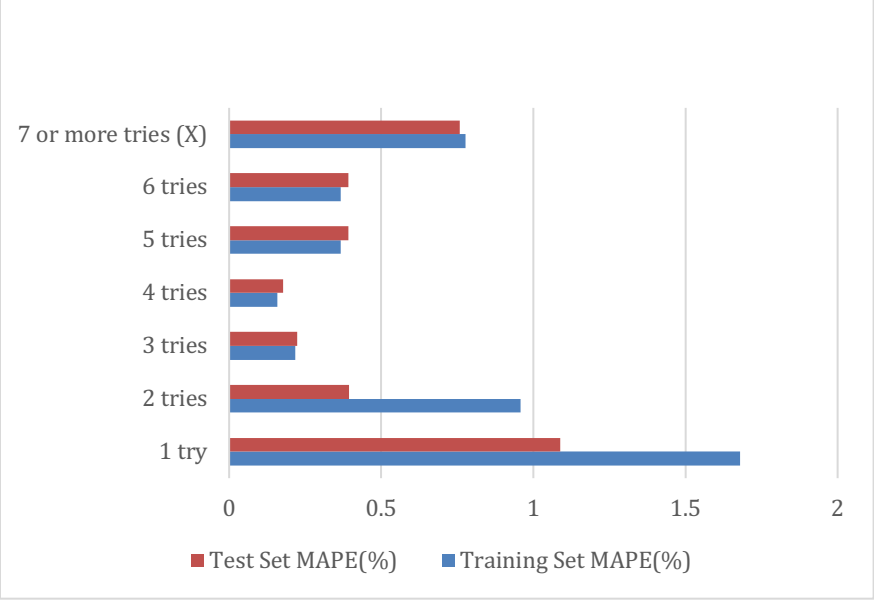
Before we start to solve this model, we prepare the parameters of LGBM.

Table 5: Parameters of LGBM

Parameter	Number
Learning Rate	0.1
N_estimtors	10
Depth_max	5
Lea_num	31
Chi_sam_min	20
Sum_hess_min	0.001
Bag_frac	1
Bag_frq	0
Bin_max	20

After training this model, we can the following results:
See Figure6.

Figure 6: Comparison of LGBM Model



According to the distribution comparison of the prediction results of the LGBM model, the following table shows the evaluation results of the model. The average MAPE of the training set is 0.646%, and the average MAPE of the test set is only 0.489%. Therefore, the performance of the model is relatively excellent.

Table 6: Results of LGBM model

Number of Tries	Training Set MAPE (%)	Test Set MAPE (%)
1 try	1.6789	1.0883
2 tries	0.9584	0.3942
3 tries	0.2175	0.2245
4 tries	0.1583	0.1774
5 tries	0.3672	0.3926
6 tries	0.3675	0.3928
7 or more tries (X)	0.7776	0.7572
Average Value	0.646485714	0.489571429

7.4 Conclusion

Based on the one-hot encoding we summarized earlier (see 4.2 for details), we get the one-hot encoding of ERRIE is:

[0, 0, 0, 0, 3, 0, 0, 0, 1, 0, 0 ,0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0]

Though the LGBM model we constructed, the relevant percentages of (1, 2, 3, 4, 5, 6, 7 or more tries) of the solution word ERRIE on March 1, 2023 is shown in Table 7.

Table 7: Different Tries Distribution on Mar 1, 2023

Try	Prediction
1 try	0.2302
2 tries	3.5132
3 tries	18.326
4 tries	31.263
5 tries	27.500
6 tries	15.751
7 or more tries (X)	2.8610

8. Classification of Word Difficulty

8.1 Solution Words Difficulties Analyzation

For this question, we decided to use the Decision Tree model, which is based on the Genetic Algorithm, to analyze the difficulty level of solution words.

To solve this problem, we designed eight layers of code and the best solution for each parameter.

Besides, in 4.4, we have already classified the difficulty level of solution words and trained our model based on that in Table 8.

Table 8: Parameter Table

Parameters	Optimal Value
depth	5.00
samsplmin	2.00
samleamin	1.00
weifrac-lea	1.00
impmin	0.10
Alp	0.20
learate	0.14
leanodmax	0.00

Finally, we can create the Decision Tree Model and get the result that the probability of ERRIE being easy is about 0.076, the probability of ERRIE being medium is about 0.755, the probability of ERRIE being difficult is about 0.18,

9. Summary

Wordle became a trendy online word game. Players enjoy this challenge of guessing the five-letter word in six attempts or less. Recently, we analyzed NYT's currently offered game, Wordle. Please allow me to share what we have found.

To begin with, we did some exploratory analysis of the data we had. The reported results started to proliferate in January and peaked in February. However, since the beginning of Mar 1, 2022, the number of reported results has been decreasing, and according to our prediction, on Mar 1, 2023, it will decrease from about 23971 to 22667. At the same time, the percentage of people who would like to challenge themselves and therefore chose the hard mode is always about 8% on each date. Therefore, we recommend Wordle make some improvements and not focus more on hard mode because of that. Besides, we also suggest keeping Wordle free. After investigation, we have seen much news that after the New York Times acquisition, Wordle will one day not be free and paywalled. Hence, it is more profitable to keep Wordle free to attract more readers of NYT and increase customer engagement to advertise other NYT games.

Next, to analyze the user experience in playing Wordle, we also construct a model to understand the percentage distribution of (1, 2, 3, 4, 5, 6, 7 or more tries). Taking ERRIE as an example, its distribution of the percentage is (0.23%, 3.51%, 18.33%, 31.26%, 27.50%, 15.75%, 2.86%) and we are 99% confident on it and only 1% uncertainty. After that, we also want to provide some suggestions. To improve user experience, we recommend utilizing this model to ensure that the solution word of every day is manageable and accessible. However, as a marketing approach, NYT can sometime increase the percentage of making 5-7 or more tries to stimulate Twitter or other social media reports about 1-2 times (s) every month.

Besides classifying the difficulty level of different solution words, we made a Decision Tree based on Genetic Algorithm. Again taking ERRIE as an example, the probability of being easy is about 0.076, the probability of being medium is about 0.755, and the probability of being difficult is about 0.18. Hence, it is a very nice tool to discern the difficulty level of solution words and ensure they are not too hard or too easy to guess.

Finally, we always believed that Wordle is a very excellent game, which also stimulated us to do this report. Wordle can also add group mode to reverse the fewer and fewer reported results. In markets, a group of people's engagement is always better than an individual's, and it is also much more fun. Besides, using our classification of difficulty level, Wordle can also set an easy-medium-hard mode to enable customers to adapt to what fits them most.

10. Reference

[1] Khandelwal, E. (2020, March 27). Light GBM vs XGBOOST: Which algorithm takes the Crown. Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/

[2] Liu, S. yin, Huang, J., & Xu, longqin. (June 2020). Combination prediction model of air temperature in poultry house for lion head goose breeding based on PCA-SVR-ARMA. *Transactions of the Chinese Society of Agricultural Engineering* (p. 36).

[3] Yan, Q., Ma, R., Ma, Y., & Wang, J. (Aug 2021). An Adaptive Simulated Annealing Particle Swarm Optimization Algorithm. *JOURNAL OF XIDIAN UNIVERSITY* (p. 48).

[4] Zhang, L., Wang, T., & Zhou, H. (2021). Research Progress of SVR Parameter Optimization Based on Swarm Intelligence Algorithm. *Computer Engineering and Applications* (p. 50).