Word analysis prediction based on ARIMA model

Yuchen Wu [1,&],Yaomeng Zhu [2,*,&]

[1] Department of Beihang University, Beijing China

[2.]Department of economics and management, Beijing University of Technology, Beijing China

* Corresponding Author Email: zym20119206@emails.bjut.edu.cn

[&]These authors contributed equally.

Abstract. First, Data processing was first performed to judge outliers if they were outliers, and finally missing values were filled in, and for the remaining valid data, change trend graphs were drawn to judge the trend of change in ascending rates and to predict future trends, using three mathematical models, ARIMA model, principal component analysis and Lasso regression model, which were used to describe the past change trends, and predict the level of change in the number of future word predictions. A comparative analysis of the models used above was conducted to compare the accuracy of each model, to derive which model predicted the most accurately, to give a rationale for the analysis, and to compare the accuracy of each model. Second, a model was built to analyze whether the change in the number of correct word predictions was related to which attributes of the words themselves and to elaborate the relationship. The data were subjected to KMO and Bartlett's tests to determine if a principal component analysis could be performed. This is followed by an exploration of whether the attribute has an effect on the number of correct predictions, how the change is affected, a visual presentation in the form of a graph, and a conclusion. To present an opinion, the paper considered the main factor that affects the number of predictions and propose measures to reduce the scenario brought by this factor. Finally, the paper came up a constructive report with listing and describing some other interesting features of this dataset.

Keywords: ARIMA model, Prediction Model, Lasso regression model, principal component analysi.

1. Introduction

Wordle is a popular puzzle currently offered daily by The New York Times. Players try to solve the puzzle by guessing a five-letter word in six or fewer attempts, receiving feedback for each guess. For this version, each guess must be an actual English word. Guesses that are not recognized as words by the game are not allowed. Wordle is growing in popularity and versions of the game now support over 60 languages.

The Wordle instructions on the New York Times website explain that the color of the block will change after you submit the word. A yellow square means that the letter in that square is in the word but in the wrong place. A green square means that the letter in the block is in the word and in the correct position. A gray square means that the letter in the square is not included in the word at all. Players can play in either regular mode or "hard mode". Wordle's Hard Mode makes the game more difficult by requiring players to use the correct letters (yellow or green squares) in subsequent guesses once they find them in a word.

The total number of entries is given in the dataset, and the content number is proportional to the growth of the date. Therefore, the content number can be used to predict the trend of a series instead of the date, and the trend of the data is statistical every day, so it can be considered a time series. The question required that given a word, and then predict the distribution of the reported results, where the impact indicators is needed to find, the first step is to study the correlation that requires quantification, factor analysis or principal component analysis method can be used for some data visualization analysis, etc., to facilitate the discovery of data relationships. And can be combined with the results of the previous three questions to analyze. In addition to the model, there must be some visualization results, and this part can be considered to output the results after each layer of attention into the form of heat map, which can reflect the process of feature extraction of words by. The number

of Wordle reports can reflect the size of the fan base of the game, and the popularity trend of the game can be observed. Roughly in January and February the number of fans grows rapidly, open questions, and some other features are derived.

2. Model Assumptions and Symbol Description

2.1. Model Assumptions

Hypothesis 1: Remove non-5-letter words. According to the question, only 5-letter words will appear in the game.

Hypothesis 2: Remove the data whose sum of percentages of attempts is not in the interval [99%,101%], beyond which the data is calculated incorrectly.

Hypothesis 3: Assume that measurement errors in available data are not considered.

2.2. Symbol Description

Table 1. Symbol Description

Symbols	Meaning
O_i	Actual frequency
E_i	Expected frequency
γ^2	Cardinality
μ	Average value
σ	Variance
SK	Skewness
K	Kurtosis
C_v	Coefficient of variation
w	Weight values from input layer to hidden layer
b	Offset top of input layer to hidden layer
w'	Weight values from the hidden layer to the output layer
b'	Offset top of hidden layer to output layer
η	Eigenvalue contribution rate
r_{0i}	Relevance
O_i	Actual frequency

3. Modeling and solving the problem

3.1. Modeling and Solving Problem 1

3.1.1 Modeling ideas

According to the statistics, there were 258 words with no repetition letters, 83 words with two identical letters (letter repeated once), and 2 words with three identical letters (letter replace twice) in the first classification. In the second classification, there were 25 words with a word frequency of five, 53 with a word frequency of four, 92 with a word frequency of three, 133 with a word frequency of two, 36 with a word frequency of one, and 6 that were not recorded (word frequencyset to zero). We summed the reported percentage of scores for each type of word and averaged them.

3.1.2 Model Building

In the use of many time series models, such as ARMA, ARIMA, will require that the time series is smooth, so generally in the study of a period of time series, the first step is required to conduct a smoothness test, in addition to the naked eye test method, in addition to the more commonly used rigorous statistical test is the ADF test, also called the unit root test. First, using the unit root test, determine whether the time series is smooth.

1. By analyzing the t-values, it was analyzed whether it could significantly reject the original hypothesis of series non-smoothness ($p<0.05$).

2. If it is significant, it indicates that the original hypothesis that the series is not smooth is rejected and the series is a smooth time series.

3. If it is not significant, it indicates that the original hypothesis that the series is not smooth cannot be rejected and the series is an unstable time series, and differencing of the data is considered, generally not exceeding second order differencing.

The ARIMA model requires the series to satisfy smoothness, view the ADF test results, and analyze whether it can significantly reject the hypothesis that the series is not smooth ($p<0.05$) based on the analyzed t-value. View the data comparison graph before and after differencing to determine whether it is smooth (not much up and down fluctuation), while the time series is biased (autocorrelation analysis), and its p and q values are estimated according to the truncated tail. ARIMA model requires the model to have pure randomness, that is, the model residuals are white noise, view the model test table, and test the model white noise according to the P value of Q statistic ($P>0.05$), also can be combined with the information criterion AIC and BIC values for analysis (the lower the better), but also through the model residuals ACF/PACF chart for analysis according to the model parameters table, the model formula combined with the time series analysis chart for comprehensive analysis, to obtain the backward prediction of the order results.

3.1.3Model solving

Table 2. ADF test table

ADF Inspection Form							
Variables	Difference order	t	P	AIC	Threshold		
					1%	5%	10%
2 tries	0	-18.444	0.000***	1933.657	-3.449	-2.87	-2.571
	1	-9.185	0.000***	1961.841	-3.449	-2.87	-2.571
	2	-9.666	0.000***	2011.42	-3.45	-2.87	-2.571

Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively

The above table shows the results of the ADF test, including variables, difference orders, t-test results, and AIC values, which are used to test whether the time series is smooth.

The results of this series test show that based on the variable 2 tries, the significance p-value is 0.000*** at the difference of order 0, which presents significance at the level and the original hypothesis is rejected, the series is a smooth time series.

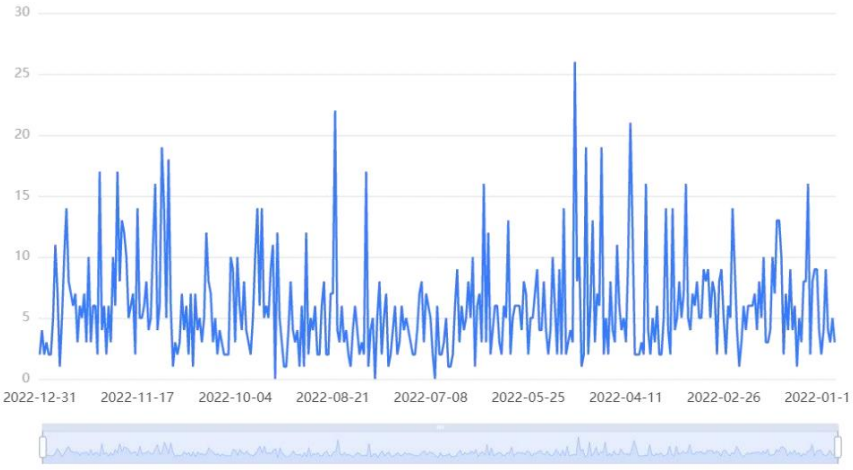

Figure 1. Original sequence diagram

The above figure shows the original plot without differencing.

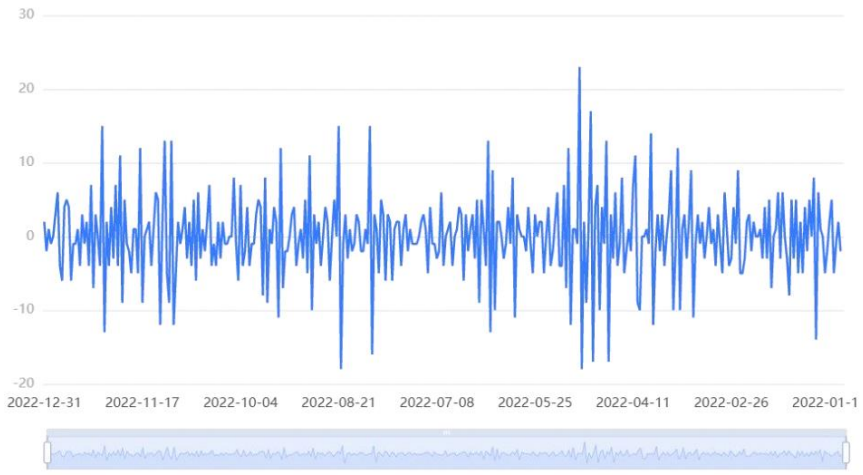

Figure 2. First-order difference diagram

The figure above shows a graph of the results of performing a first-order difference. When the time spacing is equal, the next value, minus the previous value , is used to obtain the first-order difference.

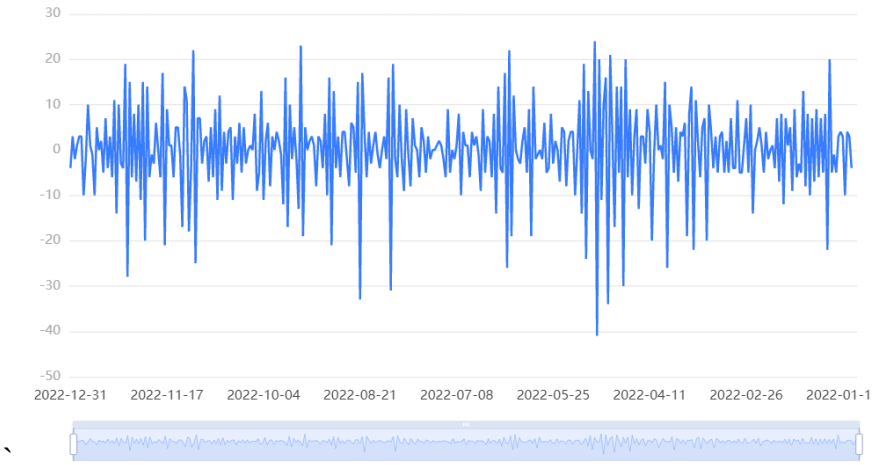

Figure 3. Second-order difference diagram

The figure above shows the result of performing a second-order difference. Doing the same thing twice, i.e., subtracting the last value from the first-order difference once more, is called "second-order difference".

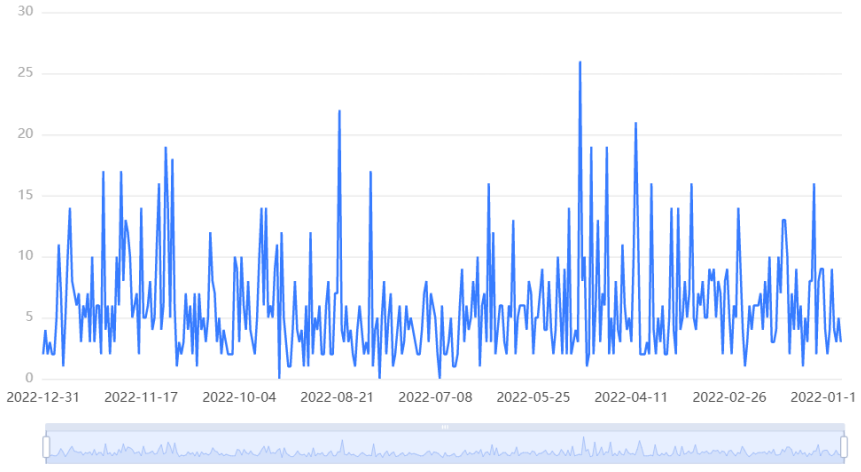

Figure 4. Diagram of the best differential sequence of the second attempt

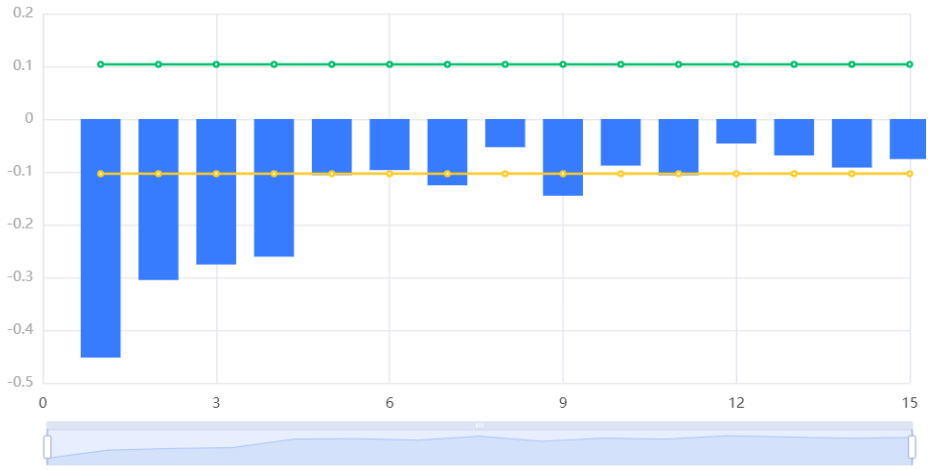

Figure 5. Final differential data autocorrelation plot

The system automatically finds the optimal parameters based on the AIC information criterion, and the model results are the ARIMA model (0,0,0) test table based on the variables: 2 tries, and from the analysis of the Q statistic results, we can get: Q6 does not present significance at the level, and the hypothesis that the residuals of the model are white noise series cannot be rejected, while the goodness of fit of the model R^2 is 0.0, and the model performs poorly, and the model basically meet the requirements.

Table 3. Table of model parameters

ARIMA model (0,0,0) test table		
item	Symbols	Value
Sample size	Df Residuals	358
	N	359
Q-statistic	Q6(P value)	0.196 (0.658)
	Q12(P value)	3.551 (0.737)
	Q18(P value)	6.674 (0.878)
	Q24(P value)	9.282 (0.953)
	Q30(P value)	14.136 (0.944)
Information Guidelines	AIC	2030.762
	BIC	2038.529
Goodness of fit	R^2	0

Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively

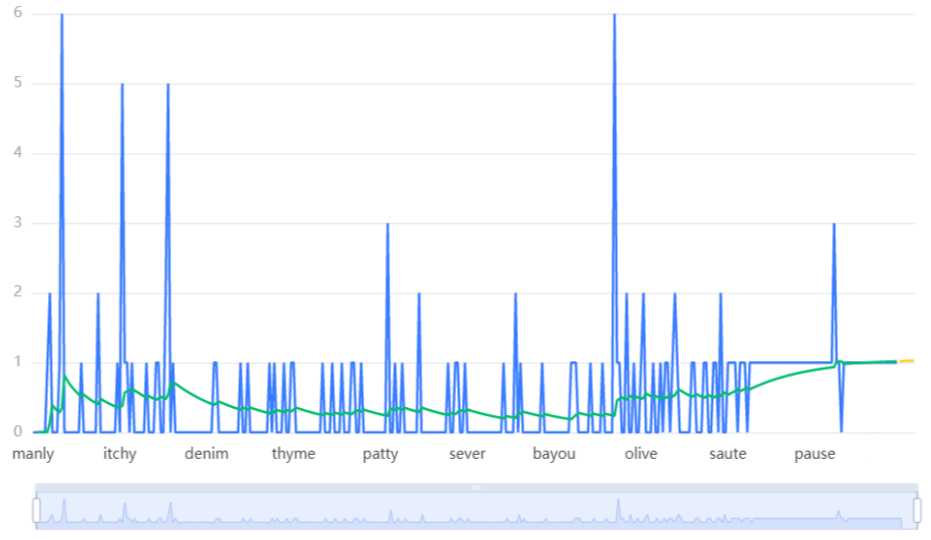

Figure 6. Time series chart

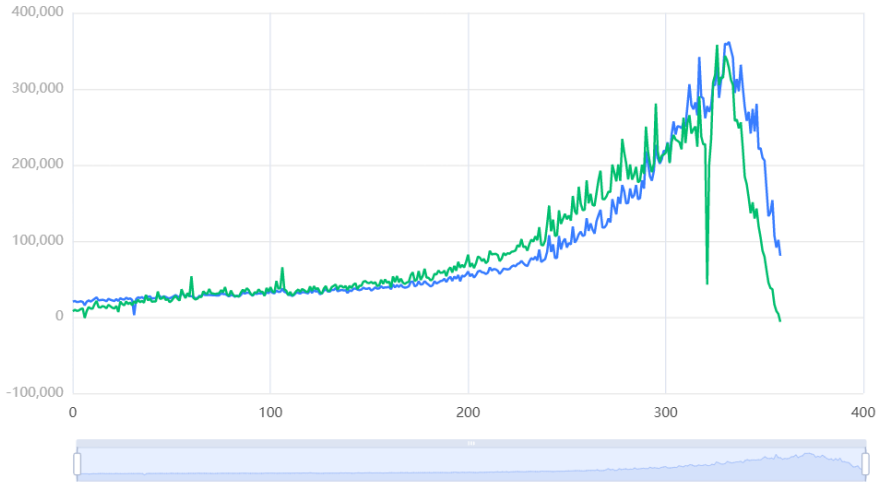

Figure 7. Fitting effect graph

The table.4. shows the correct predictions of the time series model for the next seven days for the results of the 2 attempts.

Table 4. Time series prediction table

Predicted value	
Order (time)	Predicted results
1	5.844011142061282
2	5.844011142061282
3	5.844011142061282
4	5.844011142061282
5	5.844011142061282
6	5.844011142061282
7	5.844011142061282

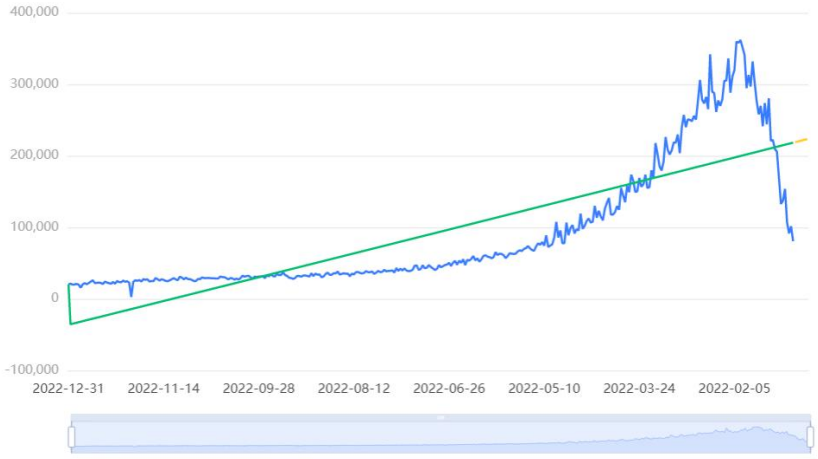

Figure 8. Result prediction chart

3.1.4 Analysis of results

Taking word characteristics into account as input to the time series prediction model, the tests were related to the attributes of the words. Words with different attributes all reach their highest percentages at a number of attempts of 4, implying that participants encounter different attributes but are most likely to guess the correct answer on the fourth guess at that point. Meanwhile, word frequency and the number of letter repetitions in a word affect the percentage of correct guesses to some extent, with high word frequency being more likely to guess correctly in fewer attempts and low word frequency vice versa. The overall trend for the three cases of letter repetition in words suggests that letter repetition may lead to an increase in the number of attempts people make.

3.2. Modeling and solving Problem 2

3.2.1Model Building

Through the cross-validation method, the value of λ was determined. λ values were chosen on the principle that they minimize the mean squared error of the Lasso model. The variables screened by the model are determined through the λ and regression coefficient plots, where the variables with zero standardized coefficients can be considered to be excluded by the Lasso regression model. The equations and prediction cases of the Lasso regression model are obtained and the retained and excluded variables are listed.

Perform dimensionless processing (homogenization, priming) for the data. Solve the gray correlation coefficient values between the parent series (comparison series) and the feature series. Solve for the gray correlation values. Ranking the gray correlation values and drawing conclusions.

3.2.2Model solving

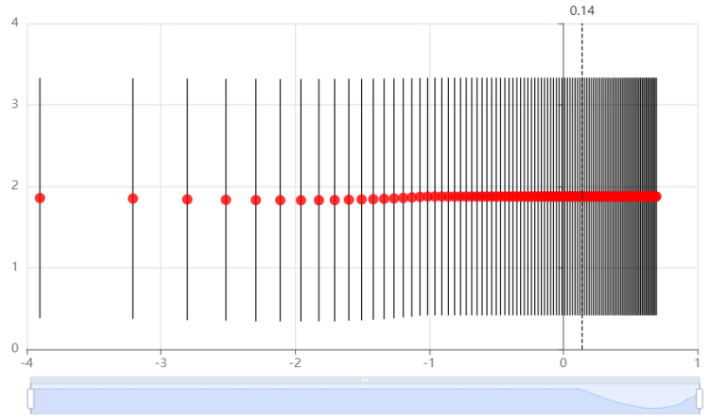

Figure 9. Lasso regression cross-validation plot

To minimize the mean square error determine λ = 0.141 and log(λ) = -1.956.

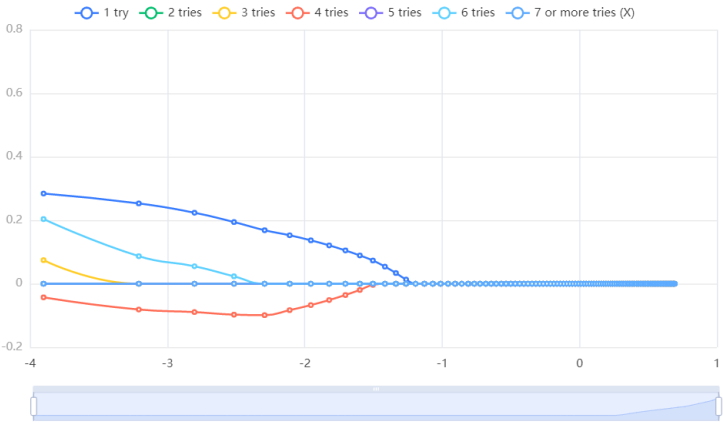

Figure 10. Plot of λ vs. model regression coefficients

Table 5. Table of model coefficients

Variable Name	Standardization factor	Non-standardized coefficient	R^2
intercept distance	-12501.516	-11931.605	0.146
1 try	1381.325	1376.849	
2 tries	63.389	58.106	
3 tries	267.947	261.462	
4 tries	81.791	77.25	
5 tries	166.392	159.079	
6 tries	306.494	302.221	
7 or more tries (X)	114.135	107.796	

The results of the Lasso regression showed that based on the standardized coefficients of the variable intercept term, 1 try, 2 tries, 3 tries, 4 tries, 5 tries, 6 tries, 7 or more tries (X), the variable intercept term, 1 try, 2 tries, 3 tries, 4 tries, 5 tries, 6 tries, 7 or more tries (X) are retained and no variables are removed.

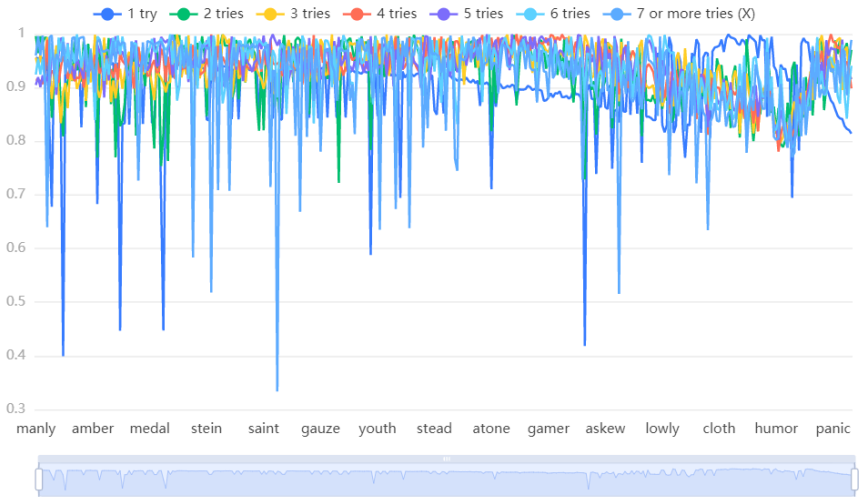

Figure 11. Correlation coefficient chart

Table 6. Relevance ranking

Relevance results		
Evaluation items	Relevance	Ranking
5 tries	0.94	1
4 tries	0.938	2
3 tries	0.937	3
6 tries	0.936	4
2 tries	0.929	5
7 or more tries (X)	0.917	6
1 try	0.903	7

The correlation coefficients were combined with the above correlation coefficients and weighted to obtain the correlation value, which was used to rank the 7 evaluation items; the correlation value ranged from 0 to 1, and the larger the value was, the stronger the correlation with the "reference value" (parent series), which means the higher the evaluation. From the above table, it can be seen that for the seven evaluation items, 5 tries has the highest rating (correlation: 0.94), followed by 4 tries (correlation: 0.938).

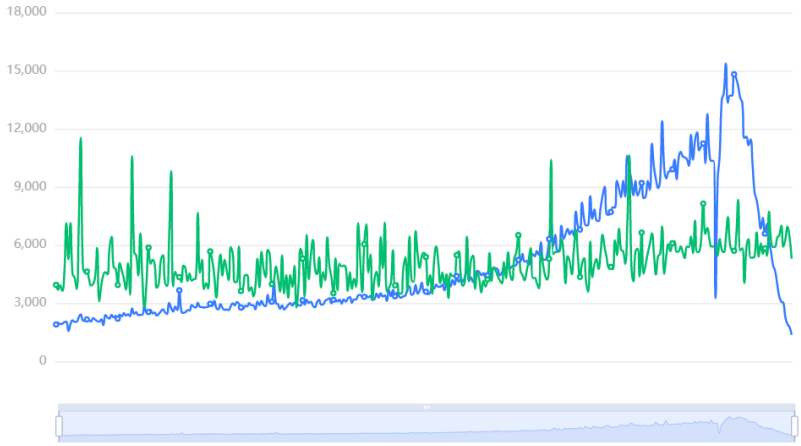

Figure 12. Model results graph

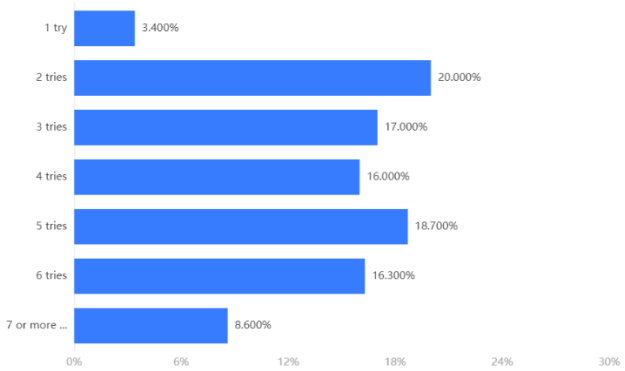

Figure 13. Percentage of feature importance correlation

Table 7. Model results predictions

Variables	Coefficient	Test value
intercept distance	11931.605	1
1 try	1376.849	
2 tries	58.106	
3 tries	261.462	
4 tries	77.250	
5 tries	159.079	
6 tries	302.221	
7 or more tries (X)	107.796	
Forecast result: 12501.516		

As can be seen above, the number of attempts needed to guess when the future date is shown above is (3.4%, 20.0%, 17.0%, 16.0%, 18.7%, 16.3%, 5.6%)

The prediction of the term EERIE on March 1, 2023, as shown in the table above, is 125.0.516, which is closer to the true value and the predicted value, indicating that the model is accurate and valid.

3.3. Modeling and Solving Problem 3

3.3.1 Model Building

First, KMO and Bartlett's tests were performed to determine whether principal component analysis could be performed. For the KMO value: on 0.8 is very suitable for principal component analysis, between 0.7-0.8 is generally suitable, between 0.6-0.7 is less suitable, between 0.5-0.6 indicates poor, and under 0.5 indicates extremely unsuitable. For Bartlett's test, if P is less than 0.05 and the original hypothesis is rejected, it means that principal component analysis can be done, and if the original hypothesis is not rejected, it means that these variables may provide some information independently and are not suitable for principal component analysis. By analyzing the variance interpretation table and the gravel plot, the number of principal components is determined The variance interpretation table mainly looks at the contribution of the principal components to the explanation of the variables (which can be interpreted as how many principal components are actually needed to express the variables as 100%), and if it is too low (e.g., below 60%), the principal component data need to be adjusted. The role of the gravel plot is to confirm the number of principal components to be selected based on the slope of the decline in eigenvalues, and the combination of these two can be used to confirm or adjust the number of principal components. By analyzing the principal component loading coefficients with the heat map, the importance of the hidden variables in each principal component can be analyzed. Based on the principal component loadings map by reducing the dimensionality of multiple principal components into two or three principal components, the spatial distribution of principal components is presented by means of quadrant plots. If 2 principal components are extracted,

the 3D loading principal component scatter plot cannot be presented, and if 1 principal component is extracted, the principal component quadrant plot cannot be displayed. By analyzing the component matrix, the principal component composition formula and weights are derived. The composite score of the principal component analysis method is output.

Clustering category variability analysis based on fields. Analysis of the frequency of each clustering category based on the clustering summary. Based on the dataset clustering annotation can know to which category each sample data is classified. The cluster center coordinates can be used to analyze the distance of each sample from the center point. An overview of the analysis is performed.

3.3.2 Model solving

Table 8. KMO test and Bartlett's test

KMO test and Bartlett's test		
KMO value		0.479
Bartlett's sphericity test	Approximate cardinality	3350.157
	df	21
	P	0.000***

Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively

The results of the KMO test showed that the value of KMO was 0.479, while the results of the Bartlett's spherical test showed that the significance p-value was 0.000***, presenting significance at the level, rejecting the original hypothesis that there is a correlation between the variables and that the principal component analysis is valid to the extent that it is highly unsuitable.

Table 9. Table of explanation of variance

Ingredients	Total variance explained		
	Feature Root	Explanation of variance (%)	Cumulative variance explained (%)
1	3.88	55.422	55.422
2	1.922	27.461	82.883
3	0.715	10.215	93.098
4	0.323	4.615	97.713
5	0.107	1.523	99.236
6	0.044	0.623	99.858
7	0.01	0.142	100

In the variance explanation table, at principal component 3, the characteristic root of the total variance explained is below 1.0 and the contribution of the variables explained reaches 93.098.

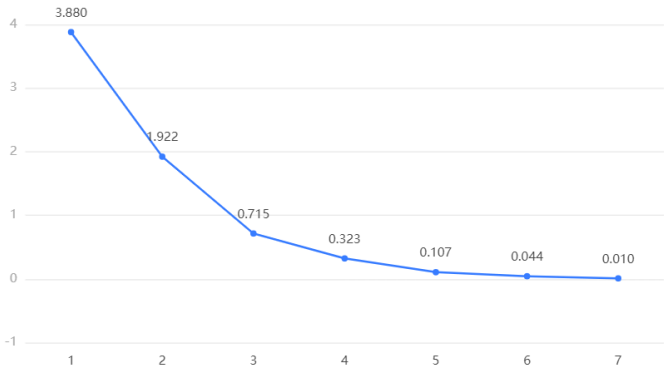

Figure 14. Gravel map

Table 10. Table of factor loading coefficients

Factor load factor table		
	Factor loading factor Main component 1	Commonality (common factor variance)
1 try	0.466	0.217
2 tries	0.813	0.662
3 tries	0.963	0.928
4 tries	0.38	0.144
5 tries	-0.846	0.716
6 tries	-0.922	0.849
7 or more tries (X)	-0.602	0.363

Table 11. Component matrix

Component Matrix	
Name	Ingredients Component 1
1 try	0.12
2 tries	0.21
3 tries	0.248
4 tries	0.098
5 tries	-0.218
6 tries	-0.238
7 or more tries (X)	-0.155

Table 12. Combined score table

Ranking	Row Index	Overall Score	Main component 1
1	train	2.9149006654696854	2.9149006654696854
2	plant	2.27405950946951	2.27405950946951
3	slate	2.1967500042410357	2.1967500042410357
4	stair	2.13201077506385	2.13201077506385
5	third	2.0610392799765616	2.0610392799765616
6	dream	1.9809406188533856	1.9809406188533856
7	rainy	1.9148765435373374	1.9148765435373374
8	tash	1.9102403888892454	1.9102403888892454
9	clen	1.9034794901482888	1.9034794901482888
10	alien	1.8659820094741104	1.8659820094741104
11	their	1.809402433861196	1.809402433861196
12	thorn	1.7564351155474676	1.7564351155474676
13	atone	1.7558212853010584	1.7558212853010584
14	treat	1.7540260497526905	1.7540260497526905
15	point	1.7531580066458896	1.7531580066458896

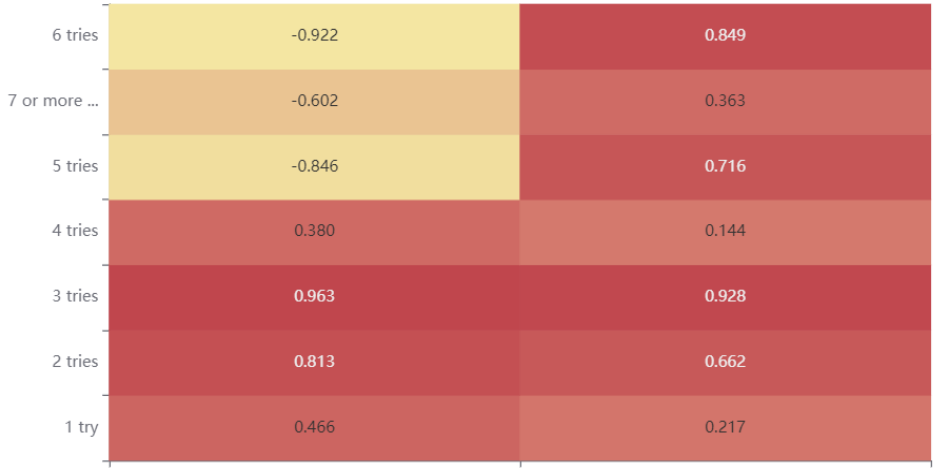

Figure 15. Heat map of the factor load matrix

Table 13. Field variability analysis

	Clustering categories (mean ± standard deviation)		F	P
	Category 1 (n=202)	Category 2 (n=157)		
1 try	0.634±0.933	0.261±0.455	21.123	0.000***
2 tries	7.901±4.156	3.197±1.831	174.65	0.000***
3 tries	28.302±4.762	15.554±4.207	700.204	0.000***
4 tries	34.703±3.98	30.643±6.003	59.022	0.000***
5 tries	19.861±3.986	28.497±4.296	387.225	0.000***
6 tries	7.347±2.465	16.981±5.294	523.232	0.000***
7 or more tries (X)	1.173±0.861	4.904±5.491	90.48	0.000***

Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively

Table 14. Clustering Summary

Clustering categories	Frequency	Percentage %
Clustering category_1	202	56.267
Clustering category_2	157	43.733
Total	359	100.0

The results of the cluster analysis showed that the clustering results were divided into 2 categories. The frequency of clustering category_1 is 202, with a percentage of 56.267%. The frequency of clustering category_2 is 157 and the percentage is 43.733%.

Table 15. Data set clustering labeling

Clustering type	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (X)
2	0	2	17	37	29	12	2
1	0	4	21	38	26	9	1
2	0	2	16	38	30	12	2
1	0	3	21	40	25	9	1
2	0	2	17	35	29	14	3
2	0	2	8	16	26	33	14
2	1	5	20	35	28	10	1
1	2	11	34	32	15	6	1
1	0	7	26	35	20	10	3
2	0	1	13	34	34	15	2

Table 16. Coordinates of clustering centroids

Clustering type	Center_value_1 try	Center_value_2 tries	Center_value_3 tries	Center_value_4 tries	Center_value_5 tries	Center_value_6 tries	Center_value_7 or more tries (X)
1	0.6336633663366338	7.9009900990099045	28.301980198019812	34.702970297	19.861386138613867	7.346534653415	1.1732673267326763
2	0.2611464968152871	3.197452229299365	15.554140127738854	30.643312101910823	28.496815286624	16.980891719745227	4.904458598726118

Table 17. Test data prediction evaluation results

Predicted results Y	Word	Predicted outcome probability_admit	Predicted outcome probability_adore	Predicted outcome probability_agape	Predicted outcome probability_album	Predicted outcome probability_alien	Predicted outcome probability_alike	Predicted outcome probability_aloud	Predicted outcome probability_alpha	Predicted outcome probability_amber	Predicted outcome probability_angry	Predicted outcome probability_aorta	Predicted outcome probability_aphid	Predicted outcome probability_apply
aptly	inert	0.0049	0.0010	0.0049	0.0049	0.0010	0.0049	0.0018	0.0049	0.0049	0.0049	0.0049	0.0049	0.0049
aptly	olivepl	0.0049	0.0010	0.0049	0.0049	0.0010	0.0049	0.0018	0.0049	0.0049	0.0049	0.0049	0.0049	0.0049
alien	antoxide	0.0040	0.0054	0.0040	0.0040	0.0054	0.0040	0.0051	0.0040	0.0040	0.0040	0.0040	0.0040	0.0040
buggy	carg	0.0042	0.0016	0.0042	0.0042	0.0016	0.0042	0.0029	0.0042	0.0042	0.0042	0.0042	0.0042	0.0042
aptly	ofoyer	0.0049	0.0010	0.0049	0.0049	0.0010	0.0049	0.0018	0.0049	0.0049	0.0049	0.0049	0.0049	0.0049
cacao	flair	0.0042	0.0016	0.0042	0.0042	0.0016	0.0042	0.0029	0.0042	0.0042	0.0042	0.0042	0.0042	0.0042
aptly		0.0049	0.0010	0.0049	0.0049	0.0010	0.0049	0.0018	0.0049	0.0049	0.0049	0.0049	0.0049	0.0049

aptly	ample	0.0049	0.0010	0.0049	0.0049	0.0010	0.0049	0.0018	0.0049	0.0049	0.0049	0.0049	0.0049	0.0049
aptly	cheek	0.0049	0.0010	0.0049	0.0049	0.0010	0.0049	0.0018	0.0049	0.0049	0.0049	0.0049	0.0049	0.0049
askew	shame	0.0044	0.0017	0.0044	0.0044	0.0017	0.0044	0.0030	0.0044	0.0044	0.0044	0.0044	0.0044	0.0044
aptly	mince	0.0049	0.0010	0.0049	0.0049	0.0010	0.0049	0.0018	0.0049	0.0049	0.0049	0.0049	0.0049	0.0049
aptly	chunk	0.0049	0.0010	0.0049	0.0049	0.0010	0.0049	0.0018	0.0049	0.0049	0.0049	0.0049	0.0049	0.0049
aptly	royal	0.0049	0.0010	0.0049	0.0049	0.0010	0.0049	0.0018	0.0049	0.0049	0.0049	0.0049	0.0049	0.0049
aptly	squad	0.0049	0.0010	0.0049	0.0049	0.0010	0.0049	0.0018	0.0049	0.0049	0.0049	0.0049	0.0049	0.0049
aptly	black	0.0049	0.0010	0.0049	0.0049	0.0010	0.0049	0.0018	0.0049	0.0049	0.0049	0.0049	0.0049	0.0049

It takes 2 times to predict the word EERIE correctly using the established model, which is efficient. The accuracy is good and the accuracy of the classification model can be obtained from the clustering summary graph.

3.4. Modeling and solving Problem 4

The overall descriptive analysis was performed for each statistical indicator of the subgroups.1. The existence of a statistically significant relationship ($P<0.05$) between XY was first tested. The correlation coefficients were analyzed for the positive and negative directions as well as the degree of correlation. The results of the analysis were summarized.

Table 18. Subgroup summary table

Word	Number in hard mode	Number of reported results	Contest number	Date
abbey	3345.000	132726.000	208.000	353.000
admit	2677.000	28994.000	463.000	98.000
adore	2398.000	25577.000	533.000	28.000
agape	3671.000	43407.000	383.000	178.000
ahead	9396.000	203730.000	258.000	303.000
album	4809.000	63380.000	339.000	222.000
alien	3395.000	38841.000	413.000	148.000
alike	2843.000	31277.000	458.000	103.000
allow	8515.000	156311.000	273.000	288.000
aloft	13708.000	319698.000	231.000	330.000
aloud	2819.000	29554.000	502.000	59.000
alpha	2706.000	29497.000	451.000	110.000
amber	2261.000	23509.000	535.000	26.000
ample	7008.000	106681.000	302.000	259.000
angry	3358.000	39667.000	395.000	166.000
aorta	2112.000	21937.000	552.000	9.000
aphid	3367.000	39086.000	397.000	164.000
apply	2150.000	22873.000	541.000	20.000
apron	4112.000	53430.000	362.000	199.000
aptly	2572.000	26498.000	499.000	62.000
aroma	10343.000	287836.000	241.000	320.000
askew	6549.000	91548.000	310.000	251.000
asset	4733.000	63188.000	341.000	220.000

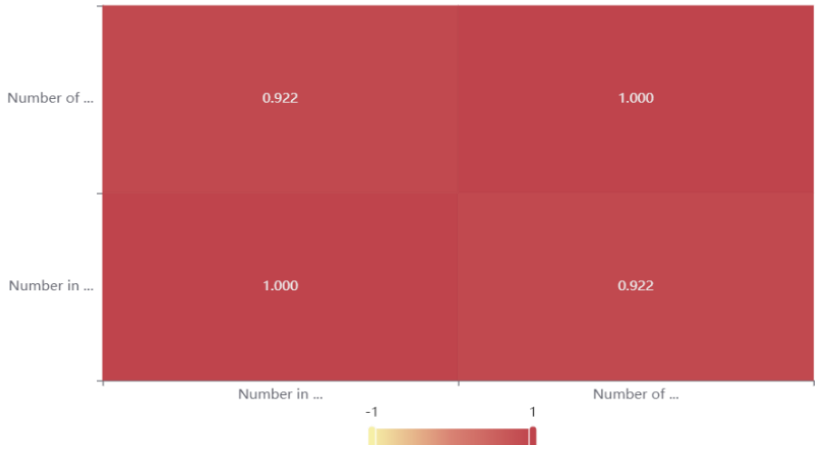

Figure 16. Heat map of correlation coefficient

Using association rules, you can find out which difficulty category a word belongs to if it has the letters E, T, or R, etc.

Based on the relevant data, the characteristics of this data were analyzed in terms of the extreme difference, distribution, anomalies, data volume and data type. The number of reports was found to have a large extreme difference of 359,339, reaching the maximum number of reports and the minimum number of reports on February 2, 2022 and November 30, 2022, respectively. The total number of users suddenly dropped precipitously on November 30, but the number of users in the difficult mode remained almost unchanged, while the number of users in the difficult mode dropped precipitously on February 13, but the total number of users did not differ much from before and after. Such phenomena may be due to incomplete data statistics or increased uncertainty in the number of reports due to social events, holidays, and other external factors.

4. Model Evaluation and Extension

4.1. Model Evaluation

Advantages of the model are as flollowing.

1. Using the ARIMA model, the iterations are repeated by varying the number of iterations to make the model solution results more accurate and at the same time to check the sensitivity of the results.

2. Under the condition of limited data, a principal component analysis model was used to select the appropriate indicators, and the indicators selected using this model had a more significant impact on the results.

3. A gray correlation prediction model is used to determine the correlation between each indicator, which reduces the correlation between the values of each indicator.

Disadvantages of the model.

1. Some theories in the gray correlation prediction model are not perfect, and some deviations may occur when making predictions.

2. The selection of indicators is made through numerical values in mathematics, without taking into account the actual situation of various indicators appearing in real life.

4.2. Promotion of the model

The gray correlation model is usually used to analyze the degree of influence of each factor on the results, and it can also be used to solve comprehensive evaluation problems that change over time.

References

[1] Wang Yan. Applied time series analysis [M]. Beijing: People's University of China Press 2005.

[2] Chen M.X., Liu SH, Sun D.X., Liu Z.Y.. Optimization algorithm of temperature distribution reconstruction based on principal component analysis [J]. Instrumentation Technology and Sensors, 2021(03): 102-105.

[3] Key Su, Jun Chen, Jie He. Principal component analysis and its application [J]. Light Industry Science and Technology, 2012, 28(09): 12-13+16.

[4] Zhao Qiang. A review of principal component analysis methods [J]. Software Engineering, 2016, 19(06): 1-3.

[5] Ke, Zheng-Lin. Application of Lasso and its correlation method in multiple linear regression models [D]. Beijing Jiaotong University, 2011