International Journal of Multiphysics
Volume 18, No. 2, 2024
ISSN: 1750-9548

"Wordle" Distribution Prediction and difficulty Classification Prediction based on Deep Learning

Zheng Li[1], Zhengdong Shi[2,*], Rui Wu[3], Yan Wang[3]

[1]School of Business Administration, Xi'an Eurasia University, Xi'an, China
[2]School of mathematical information, Shaoxing University, Shaoxing 312000, China
[3]School of Economics and Finance, Xi 'an Jiaotong University, Xi'an, China
*Corresponding Author.

Abstract

This project mainly focuses on the study of word guessing games, analyzing factors such as the popularity of the game, the playability of the game itself, the difficulty of the game itself, and the game effects of the participants. Firstly, analyze the useful attributes from the words and whether these attributes have an impact on the proportion of game score results reported on that day. Secondly, clustering models are used to automatically partition words, and the difficulty is divided into two levels: simple and difficult. The machine learning xgboost model is used to evaluate the model through confusion matrix, accuracy, recall, and F1 value. The study found that the total number of reports is consistent with the trend of the number of reports in hardcore mode. The popularity rapidly increased in the early stage and gradually decreased in the later stage, which means that the loyalty of players has basically been washed away, and the remaining are mostly people with great potential to continue playing.

Keywords: Wordle guessing game, ARIMA model, regression analysis, xgboost model

1. Introduction

"Wordle" is currently a popular puzzle provided by The New York Times every day. Players try to guess a five letters word six times or less in a short amount of time to solve this puzzle, and receive feedback on each guess. After you submit the words, the color of the tiles will change. The yellow block indicates that the letters in the block are in the wrong position within the word. The green block indicates that the letters in the block are in the correct position within the word. The gray block indicates that the letters in the block are not included in the word at all. This game has different difficulty levels: regular mode and hard mode. The reason why hard mode is more difficult than regular mode is that when players guess the correct letters in hard mode, these correct letters will be used as prerequisites in the rest of the game. Many players will upload their game data on Twitter, and although the data is consistent with the guessing results, it cannot be ruled out that the results are affected by rounding, so the guessing results relying on Twitter data still need to be reviewed. .[1,2]

Regarding the issue of "Wordle", we use data generated by MCM between January 7, 2022 and December 31, 2022, and provide specific solutions using machine learning and related models.

Machine learning is a branch field of artificial intelligence, and an important concept in machine learning is generalization ability, which refers to the model's ability to perform well when processing unseen data. Machine learning is a branch field of artificial intelligence, and an important concept in machine learning is generalization ability, which refers to the model's ability to perform well when processing unseen data. As a machine learning method with strong generalization ability, ensemble learning can fully utilize the complementarity of multiple learners, avoid overfitting problems of a single learner, and improve overall performance. Through ensemble learning, the prediction results of multiple models can be combined to obtain more accurate prediction results than a single model, thereby improving the generalization ability of the model. Boosting algorithm, as a typical ensemble learning method, improves the overall performance of the model by training multiple weak learners and combining them with weights.

2. The Trend of Game Popularity over Time

2.1 Model establishment and Solution

Based on the results reported by the New York Times for the “Wordle” game, we will build a model to explain the variation in the number of reported results, and create a prediction interval. In this paper, we choose to use a time series forecasting algorithm, using the Arima model, As shown in Figure 1. where the dependent variable is the number of reported results, and the independent variables are Date, Contest number, Word of the day, Number of, reported results, and Number in hard mode.

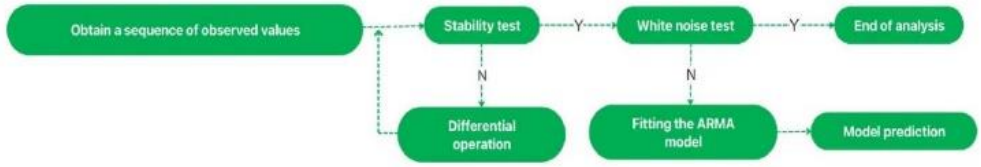


Figure 1. Analytical flow chart

ARIMA model Basics: A model built by transforming data into smooth data by differencing, and then regressing the dependent variable on only its lagged values and the present and lagged values of the random error term. It is denoted as ARIMA(p, d, q).[3,4]

ARIMA(p,d,q) parameters: where AR is autoregressive, p express autoregressive order. d express trend difference order. MA is "sliding average", q express moving average order.

First look at how the series has changed over time.

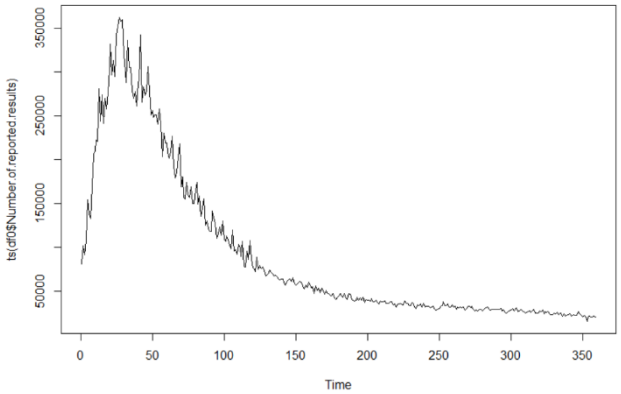
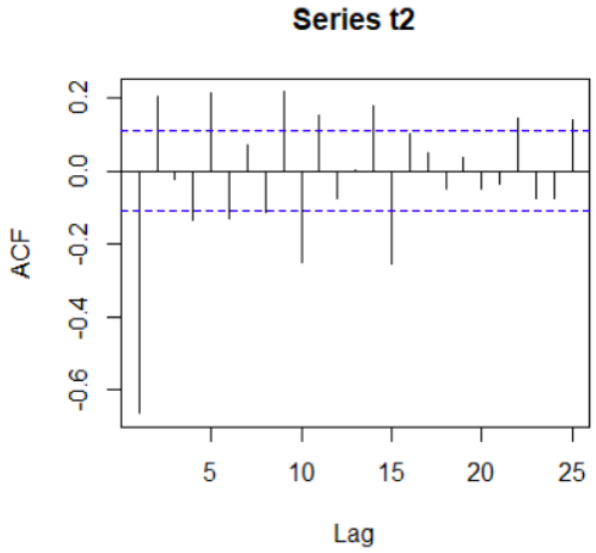


Figure 2. Number of reported results time series diagram

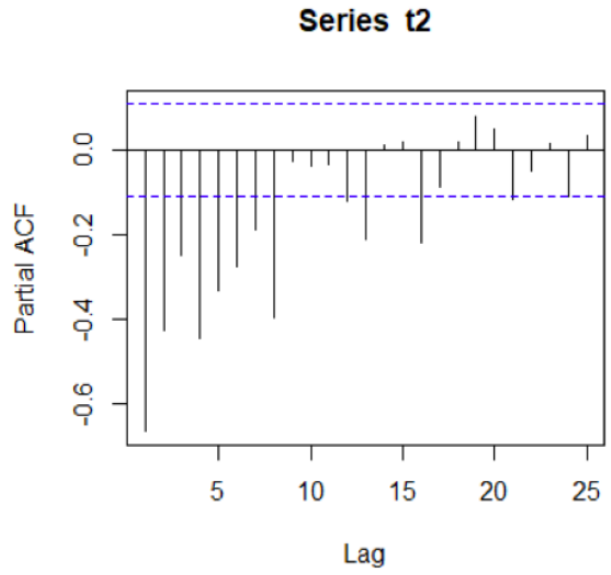
As shown in Figure 2, we can find that at the very beginning, the game's popularity rose and even rose above 350,000, but after that the image showed a very steep decrease, the game's popularity decreased and people gradually became tired of participating in the game, and then the number of participants in the game gradually

decreased, but the trend tended to level off until the end of 2022. It is foreseeable that the game's popularity will still decrease, but it will also still decrease with a smooth trend.

A time series model can be used to predict the number of game participants in the later period, and an ARIMA model is established for the prediction of the time series, with a test set of 30 samples, and a second-order difference is performed on the series to test that the series is smooth and non-white noise series, a and to view the autocorrelation and partial autocorrelation plots.



(a). autocorrelation plot



(b). partial autocorrelation plot

Figure 3 Autocorrelation and partial autocorrelation

The concluding model can be derived as shown in Figure 3(a), 3(b) above: ARIMA(8,2,4) and the coefficients at ma2 and ma3 are removed to 0. The coefficients after the model is established are. Model results as shown in Table1.

Table 1. Detail of ARIMA(8,2,4).

ar 1	**ar 2**	**ar 3**	**ar 4**
-0.4433	-0.0699	-0.0181	-0.9902
ar 5	**ar 6**	**ar 7**	**ar 8**

-0.3576	-0.1362	-0.3360	-0.3998
ma 1	**ma 2**	**ma 3**	**ma 4**
-1.1898	0	0	1.1980

The model residual series are non-white noise series. as shown in Figure 4.

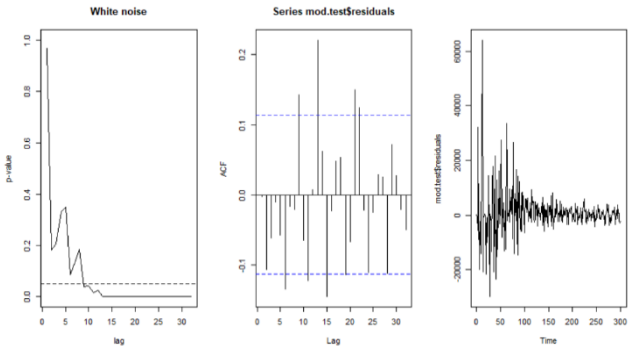


Figure 4. Residual sequence test

The AIC and BIC of the model are 6261.25 and 6305.61 respectively. At the same time, the errors of the comparison test set after 30 days are predicted as shown in Table 2.

Table 2. Model's accuracy

	ME	**RME**	**MAE**
Training	774.060	8019.418	4502.187
Test	2390.544	2972.621	2617.973
	MPE	**MAPE**	**MASE**
Training	0.752	5.676	0.738
Test	10.359	11.751	0.429

As shown in Figure 5 the original data graph, model fitting value and model prediction value of the time series model. It can be seen from the figure that the trend of the fitting sequence has great similarity with the trend of the real sequence, which shows that the fitting effect is good.

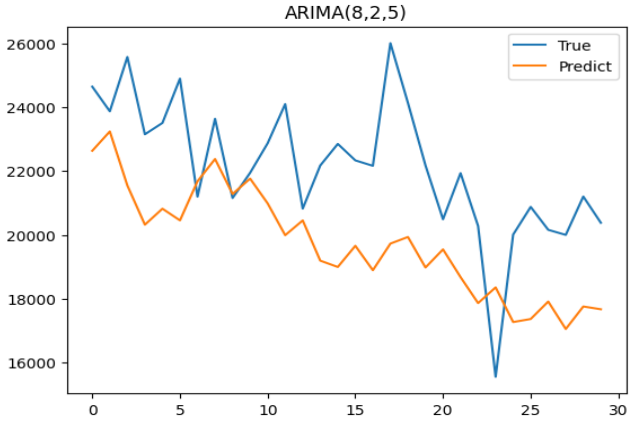


Figure 5 Prediction of ARIMA(8.2.5)

So the model is finally determined as.

$$\nabla^2 x_t = \frac{-1.3893B+0.4184B^2}{1+0.3149B+0.2522B^2+0.3165B^3+0.3013B^4}\varepsilon_t \quad (1)$$

Using the model to project backwards 60 days yields a result of 15180, which means that the total number of people reporting results on March 1, 2023 will be in a range of 15180 or less.

2.2 Parse available attributes

The percentage of each score in the hardcore mode visually expresses the difficulty of the question. The attributes of words that can be visually linked to the difficulty of a topic are first and foremost how commonly the word is used in life and the cultural trends of the time. Google's Google Books Ngram Viewer provides the Ngram data of some of the books scanned and digitized by Google Books (4% of the books published by humans), in which the frequency of words in a large number of books after 1800 is counted, which is precisely a form of life and culture, and can be searched to get the frequency of the word in a certain period of time by targeting specific time periods and words. This is a good way to show the frequency of words in common use, and solves part of the problem of how to quantify the difficulty of wordle questions.

By grabbing the packet to get the interface and parameters of the website request word frequency, by forging parameters to all words in the table can be requested to obtain the word frequency of all words, the data obtained is a very small percentage of decimal, just need to enlarge the data can be.

In addition, after playing one or two rounds of the wordle guessing game, it can be found that another factor affecting the difficulty of the questions is whether there is a repetition of letters in the words and how many repetitions, and this kind of doubt deepens the difficulty of guessing words when the words are finally found to be rare words. To address this issue, the number of repetitions in each word can be counted as an attribute of the word, and this attribute may indirectly affect the difficulty of the question.

This yields the two word attributes used for analysis.

2.3 Model establishment and analyze

After deriving the two attributes of frequency of word use and number of letter repetitions respectively, seven linear regression models were made with the number of each attempt as the dependent variable and the two words attributes as independent variables.[5,6]

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (2)$$

The following table shows the coefficients and p-values for the two attributes of the seven models mentioned above, from which a glimpse of the effect of these two attributes on the percentage of attempts can be obtained.

Table 3 Linear regression model's result

	Frequency of use		Letters Repeated Count	
	Coefficient	P value	Coefficient	P value
1 try	0.0014	<0.001	0.1760	0.038
2 tries	0.0115	<0.001	2.7358	<0.001
3 tries	0.0343	<0.001	14.6154	<0.001
4 tries	0.0357	<0.001	28.0301	<0.001
5 tries	0.0192	<0.001	24.4733	<0.001
6 tries	0.0080	0.0160	0.0080	<0.001
7 or more	0.0016	0.2410	3.6712	<0.001

The magnitude of the p-value is usually used in linear regression to reflect whether the variable is statistically significant or not. It represents the probability that the regression coefficients in the regression analysis are randomly occurring. When the p-value is less than 0.05, we can consider the variable to be statistically significant, which means that it is closely related to the significance of the results. As shown in Table 3.

The significance of the number of attempts on the frequency of word use and the number of letter repetitions can be found in the p-value. It is obvious from the above table that for the frequency of word use, excluding the first attempt based on luck alone, the fewer the number of attempts, the more significant the effect of the frequency of word use on the guessing of the question, because after the first attempt to learn the existence of the letter, everyone will go to associate the common words in life. Given that the coefficients are all positive, it can also be interpreted that commonly used words are more likely to be guessed in the first few attempts.

Similarly, the number of repetitions of letters can be interpreted in such a way that as the number of attempts increases, the more significant the effect of the number of repetitions of letters in the word on the guess, given

that the coefficient is positive, it can be understood that the more repetitions of letters in the word the easier it is to be guessed in the middle and later stages, because most of the first one or two guesses are only to determine which letters are present in the word and do not identify how many of that letter there are, so It is only after the first second or even third attempt that a range of answers can be roughly determined, whereas in the case of commonly used words, they can quickly be attempted based on this range.

3. The Trend of Game Popularity over Time

3.1 Parse available attributes

Mean Square Error and Mean Absolute Error are two commonly used measures of prediction model accuracy. They can both be used to measure the difference between the predicted and true values, but there are some differences in how they are calculated and how sensitive they are to error. The mean squared error is the average of the squares of the differences between the predicted and true values, and is calculated as.

$$MSE = \frac{1}{m}\sum_{i-1}^{n}(yi - \hat{y}i^2) \quad (3)$$

Where n is the number of samples in the data set, yi is the true value of the ith sample, and ŷi is the predicted value of the ith sample. the smaller the value of MSE, the more accurate the prediction model is. MSE has good mathematical properties and is commonly used in optimization algorithms.

The Mean Absolute Error is the average of the absolute values of the differences between the predicted and true values, and is calculated as follows.

$$MAE = \frac{1}{m}\sum_{i-1}^{n}|yi - \hat{y}i| \quad (4)$$

The smaller the value of MAE, the more accurate the prediction model is. In some scenarios, MAE is more intuitive and easier to interpret than MSE because it indicates the absolute magnitude of the mean error rather than the squared magnitude.

3.2 Model establishment and analyze

The two attributes that have an impact on the difficulty of guessing have been extracted from the words, but it is obviously not enough to infer the percentage of each attempt through these two attributes alone, as the alphabetical arrangement of the words themselves will more or less also affect the difficulty of guessing and thus the percentage of different attempts on that day. -26 for a total of 5 fields.

On this basis, assuming that the difficulty of daily guesses is completely independent of each other, and the percentage of attempts in the first few days is completely independent of the percentage of attempts in the next few days, a model can be built to predict the percentage of attempts in a certain day in the future by building a total of 7 xgboost regression models to predict the respective percentage of each kind of attempts, and finally by normalizing the sum of the percentage to 100%.

The core principle of XGBoost[7-11],An integrated learning algorithm that can be used to solve regression and classification problems, is to build a decision tree model by Gradient Boosting. Unlike traditional decision trees, each decision tree of XGBoost is composed of multiple leaf nodes, each of which corresponds to a prediction value. The prediction result of the model is the sum of the prediction results of all decision trees.

After the models were built and tuned, the MAE and MSE of the seven models were calculated separately from the test set and compared with the accuracy of the results by means of graphs and lines.

Table 4. MAE and MSE of every XGBoost model

	1try	2tries	3tries
Training	774.060	8019.418	4502.187
Test	2390.544	2972.621	2617.973
	MPE	**MAPE**	**MASE**
Training	0.752	5.676	0.738
Test	10.359	11.751	0.429

As shown in Table 4. The highest mean squared error is the one with 3 attempts, and the mean squared error is only 22.37. The comparison between the prediction results and the test set shows that the prediction results are very close to the actual values in most cases, which shows that the model is effective in non-extreme cases. The model is valid in non-extreme cases. Model prediction effect as shown in Figure 6.

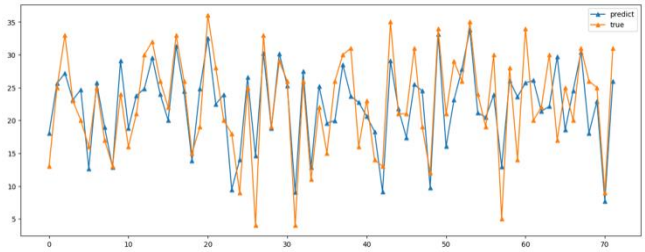


Figure6. 3tries XGBRegression model's prediction

Combine the above 7 models to predict the distribution of the percentage of attempts on the day of March 1, 2023 represented by the title "EERIE", minimize the letters of the word and extract the number of letter overlaps, then find the frequency of the word on Google Books Ngram Viewer, and finally The code for each letter in the word is passed into the model.

The predicted results are shown in Table 5. Similar to the data given in the title, given the rounding, the summation is 101%, the number of attempts is mainly concentrated in 4/5/6, and the number of people who cannot guess is relatively small.

Table 5. EERIE's distribution

	1try	2tries	3tries	4tries
Distribution	0	3	14	28
	5tries	6tries	7 or more	
Distribution	34	20	2	

4. Classify and Grade the Characteristics of Words and Predict the Difficulty of Words

The most intuitive way to express the difficulty of a question is to observe and compare the results, because objectively, the difficulty of a question is summarized and defined by all participants for the whole, as it is for "wordle" questions. The 7 different attempts given in the question represent the scores, which also represent the difficulty of the question for the guessers.

Hierarchical clustering is performed with all attempts as variables. Hierarchical clustering creates a hierarchical nested clustering tree by calculating the similarity between data points of different categories.

In a clustering tree, the original data points of different categories are the lowest level of the tree, and the top level of the tree is the root node of a cluster. Hierarchical clustering can make up for the shortage of K-means,[12-15] and hierarchical clustering does not need to determine the K value in advance, and can handle irregular scatterplots. Meanwhile, the contour coefficient is an index used to evaluate the good or bad clustering effect. It can be understood as an index describing the clarity of the contours of each category after clustering. The contour coefficient indicators are as follows.

$$S(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}} \quad (5)$$

Its value range is [-1,1], the larger the contour coefficient the better the clustering effect.The distance between clusters is calculated using the ward method with the following equation.[16,17]

$$d(u,v) = \sqrt{\frac{|v|+|s|}{T}d(v,s)^2 + \frac{|v|+|t|}{T}d(v,t)^2 - \frac{|v|}{T}d(s,t)^2} \quad (6)$$

After the model was established, the two most important indicators were output using PCA principal component analysis in order to draw a scatter plot to visualize the results of clustering classification. The difficulty was divided into three stages: easy, medium, and difficult, and the clustering profile coefficient obtained using the

ward method of hierarchical clustering was 0.3538, while the clustering results were presented in the form of scatter plots as shown in Figure7.

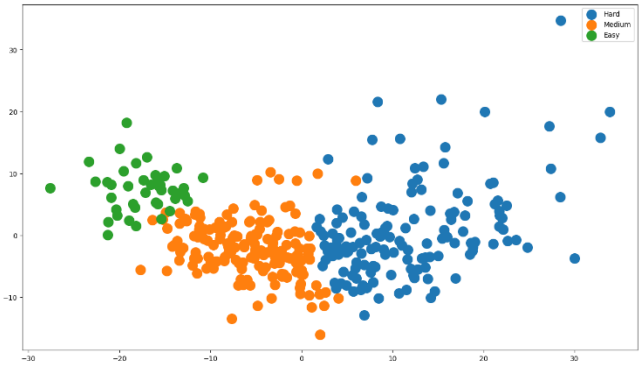


Figure 7. Scatter plot of clustering division

It can be seen that the difficulty accounts for the right half of the graph, but the percentage of easy difficulty is very small. For the convenience of classification, the following will merge easy and medium difficulty into easy difficulty, with only two modes of easy and difficult. The clustering contour coefficient obtained after re-clustering division is: 0.4199, compared to the clustering contour coefficient of three levels of difficulty, the clustering contour coefficient of two levels of difficulty has been improved, while the scatter plot is shown as follows.

The simultaneous spectral clustering diagram is as shown in Figure 8.

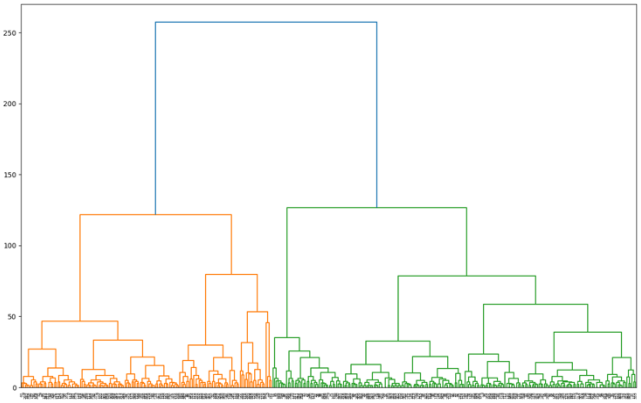


Figure 8 Pedigree cluster map

In total, the chart is divided into two categories, orange for easy difficulty and green for hard difficulty.
Finally, to prove the validity of the clustering model, it is sufficient to see whether the total percentage of attempts for the two difficulties corresponds to the characteristics expected for that difficulty. The average of the number of attempts for each of the two difficulties is found as shown in Table 6.

Table 6. Average of every tries based on different difficulties

	1try	2tries	3tries	4tries
Hard	0.3287	3.1575	15.3287	30.1095
Easy	0.6415	7.6792	27.8537	34.8632
	5tries	6tries	7 or more	
Hard	28.5958	17.3972	5.0821	
Easy	20.2311	7.4905	1.2405	

As shown in Figure 9.It can be clearly seen that the number of attempts under the simple model is significantly less than the difficult model, mainly concentrated in 3 to 5 times, and 6 times and did not guess out has been very little, try 6 times to guess out of the proportion of basic before and after 7%, the difficult model is the first few times to guess out of the proportion of less than the simple model, the proportion of late guessed and not

guessed is obviously greater than the simple difficulty. In this way to separate the degree of difficulty is a certain effect, but the difficulty is only divided into two levels in some cases is not fine.

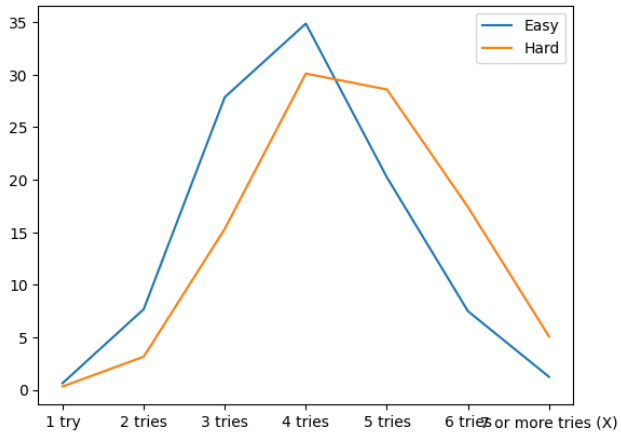
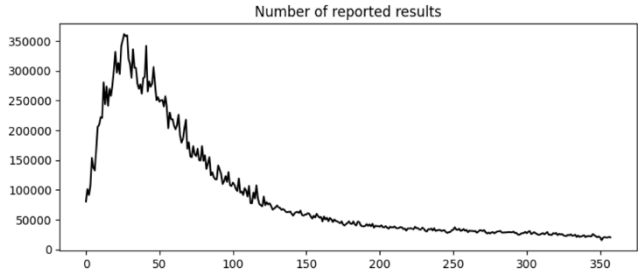


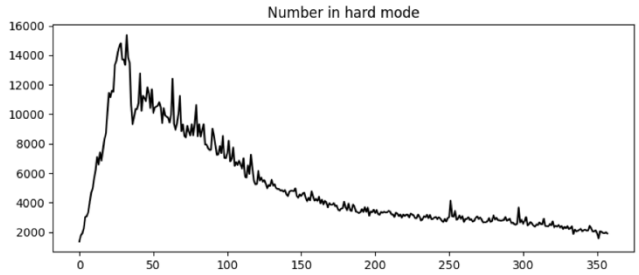
Figure9 Line chart of every tries based on different difficulties

5. Analyse the Data Provided

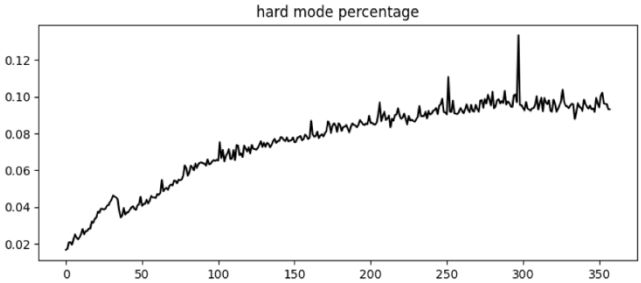
From the beginning of 2022 to the end of the year, wordle games experienced a high level of heat for several months and then a steady decrease in heat, both in terms of the total number of reports on Twitter and the number of reports in hardcore mode only in general, but in fact, the proportion of reports in hardcore mode did not drop or remain the same, twenty continued to rise and then stabilized in a certain range. As shown in figure 10.



(a) Number of reported results



(b) Number of hard mode



(c) Number of hard mode percentage

Figure 10 Number of reported results and hard mode with it's percentage

For this game, the flow is then as follows, and perhaps a complex network dynamics model can be built for this case using a similar style to the infectious disease model,[18-20] with the following transformed flow. As shown in figure 11.

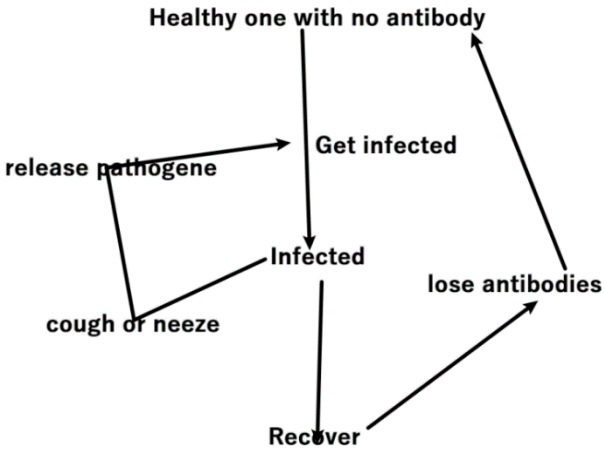


Figure 11 Infectious disease model's flow chart

6. Conclusions

First of all, for the number of reports reported every day, choose time series prediction to build a time series Arima model, or timing data sliding window conversion (data processing) + machine learning regression to predict. Or it can be used for regression prediction. Because variables are the number of reported results, independent variables can be dates, competition numbers, words of the day, the number of reported scores on the same day, and the number of players in difficult mode. It involves looking for the attributes of words, such as the length of words, word frequency, the structure of words (referring to the consistency between the pronunciation and spelling of words), etc. After the word attributes, it can be correlated to the percentage of the report. The models used include correlation analysis, regression analysis, etc.

Next, we use the xgboost model to predict the proportion of future date scores and the proportion of the word EERIE on March 1, 2023. The question requires this relevant percentage result, that is, the prediction probability obtained by each classification, and the largest prediction probability represents the classification result, so there is no need to pay attention to the classification results, but on the prediction probability of each classification, and finally, the model is advanced by confusing matrix and accuracy, accuracy, recall rate and F1 value. Evaluate, and the evaluation results represent our confidence in model prediction.

Finally, we use cluster analysis to automatically divide words, and then look at the characteristics of each attribute of the corresponding classification, so that we can summarise the attributes of each difficulty.

Acknowledgements

This research was supported by The Younth Innovation Team of Shaanxi Universities. Xi'an Eurasia University Technical Service Special Project: "Research on Dynamic Monitoring and Evaluation Process of Regional Education Informatization Development" (OYJSFW-2021004). Shaanxi Province's Education Science 14th Five Year Plan for 2021 Annual Project of Shaanxi Provincial Social Science Foundation: "Research on the Path of Digital Transformation and Development of Universities in the Digital Economy Era" (SGH21Y0382)

References

[1] Zhirui Min. A study of the number of Wordle users and experience predictions. Academic Journal of Mathematical Sciences. Volume 4, Issue 2. 2023

[2] Hontz Eric. Wordle Unlimited emerges as a word-guessing game designed to guess and create new words. M2 Presswire. 2023

[3] Zhou Binbin, Huang Jiaxin. Short-term Power Generation Forecasting in China Based on Gray-ARIMA Coupling Model. Science Technology and Industry, 2022, 12(22): 382-386.

[4] Xue Yu, Wang Changqing, Zhu Ya. Prediction of regional medical and health service volume under ARIMA gray coupling model. Health soft science, 2019, 33(11):51-56

[5] Xie T, Shang Q, Yu Y. Automated traffic incident detection: coping with imbalanced and small datasets. IEEE Access. 2022, 10:35521-35540.

[6] Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data?. Advances in Neural Information Processing Systems. 2022, 35:507-520.

[7] Md. Mehedi Hassan; Sadika Zaman; Efficient prediction of coronary artery disease using machine learning algorithms with feature selection techniques. Computers and Electrical Engineering. Volume 115, 2024. PP 109130-109138.

[8] Fellippe R.A. Bione; Igor M. Venancio; Estimating total organic carbon of potential source rocks in the Espírito Santo Basin, SE Brazil, using XGBoost. Marine and Petroleum Geology. Volume 162, Issue. 2024. PP 106765

[9] Padala Raja Shekar; Aneesh Mathew; A combined deep CNN-RNN network for rainfall-runoff modelling in Bardha Watershed, India Artificial Intelligence in Geosciences. Volume 5, Issue. 2024. PP 100073-100086

[10] Wu Renbiao, Liu Yang. Risk assessment method for key civil aviation passengers based on improved XGBoost. Journal of Safety and Environment. 2022. 2(23): 651-656

[11] Xiao Haijun, Kan Tingting. Parameter selection of xgboost based on local search bayesian algorithm. Journal of South-Central Minzu University. 2023. 3(42): 201-206

[12] Zhang Yukun. Research on e-commerce student customer segmentation based on K-Means clustering analysis. Market Modernization. 2022. 3. 33-35

[13] Li Hua, Zhao Shuying. Construction and Analysis of Financial Security Index Evaluation System Based on the Weighted Principal Component Distance Clustering. Mathematics in practice and theory. 2018. 1(48): 90-95

[14] Guo Yixin, Han Xia, Ni Guangjie. Dynamic characteristics clustering algorithm of photovoltaic power station based on improved Canopy–FCM. Computer Applications and Software. 2022. 4(39): 288-293.

[15] Zhang Nan, Li Xiang, Jin Xiaoning, et al. Joint prediction model of English words and their capitalization in neural machine translation. Chinese Journal of Information, 2019, 33 (03): 52-58

[16] Surjeet Dalal, Edeh Michael Onyema, Amit Malik. Hybrid XGBoost model with hyperparameter tuning for prediction of liver disease with better accuracy. World Journal of Gastroenterology, 2022, 28(46): 6551-6563.

[17] Hayashi Yusuke; Okazaki Saho, Development of concentration prediction models for personalized tablet manufacturing using near-infrared spectroscopy. Chemical Engineering Research and Design. Volume 199, 2023. PP 507-514

[18] Jong Hyun Lee; In Soo Lee, Hybrid Estimation Method for the State of Charge of Lithium Batteries Using a Temporal Convolutional Network and XGBoost. Batteries. Volume 9, Issue 11. 2023

[19] Myers Renée C; Augustin Florian. Using machine learning surrogate modeling for faster QSP VP cohort generation. CPT: pharmacometrics & systems pharmacology. Volume 12, Issue 8. 2023. PP 1047-1059

[20] Van der Sande Kiera. Accelerating Explicit Time-Stepping with Spatially Variable Time Steps Through Machine Learning. Journal of Scientific Computing. Volume 96, Issue 1. 2023