Research On Lexical Data Set Analysis Based on Decision Tree Model

Yufei Du *

School of Finance, Dongbei University of Finance and Economics, Dalian, Liaoning, 116025

* Corresponding Author Email: duyufei2001@163.com

Abstract. In order to predict the future number of users and development prospects of "World" game, this paper analyzes the terminology used in game data sets. Firstly, descriptive statistical analysis was used to analyze the existing data, and ARIMA model was used to estimate the data during the forecast period, so as to obtain the interval estimation and point estimation of the number of results. Secondly, EXCEL is used to calculate the percentage of each word, and analysis of variance model is used to get the attribute influence between words. Thirdly, the training set and test set were analyzed through machine learning, and the mapping model was established. "EERIE" was input as the word vector to obtain the prediction results. Finally, the relationship between the decision tree model and the actual expected break time is established, and the difficulty of EERIE is evaluated. The results show that with the word vector of EERIE as input, when the model accuracy is 55.56%, the prediction result is (1,2,3,4,5,6,X)-(6,12,21,33,25,3,0). The prediction difficulty was medium, and the accuracy of the model was 80%.

Keywords: Arima, Decision-tree model, Data process, Data visualization.

1. Introduction

“World”, a word game that became popular in the United States and even around the world in a short period of time, was featured in the New York Times, showing the game's influence and even a cultural phenomenon [1]. Each player around the world eliminates and locates five bytes a limited number of times a day, eventually working together to get a word given by the system. Because the popularity of word play partly reflects the fact that people today do not reject the incorporation of local culture into entertainment, but actively support it. Therefore, this paper uses the data link behind the game to get the support of the model, which can not only predict the number of users and even development prospects of the game in the future, but also reflect people's demand for cultural entertainment [2], which has important reference significance for future game development.

2. Materials and methods

2.1. Data source and analysis

This paper uses 2023 mathematical contest in modeling certificate contest C (https://www.contest.comap.com/undergraduate/contests/mcm/contests/2023/problems/) for analysis and research. A descriptive breakdown of the data presented suggests that world still has a solid user base and is not a passing cultural phenomenon.

2.2. Data preprocessing

2.2.1 Data Description

We plot the current data in chronological order: visualize the daily reported results in Python during the period 2022-2023 using a moving window, and we can see through Figure 1 that the total player score (average player score * the total number of players) peaked in February 2022 (one month after the game was launched), indicating a high level of overall market acceptance, even though the total score gradually declines after the peak (here we think the precipitous decline will not be due to a significant drop in the average player score, but rather to a higher loss of players), the tail is

relatively stable, indicating that Wordle still has a stable user base and is not a fleeting cultural phenomenon. (Note: the word "study" is located in line 330, and its reported results we think are too different from other words, in order to ensure the accuracy of the subsequent modeling predictions so the "2569" is changed to "25690 ") The details are shown in Figure 1.

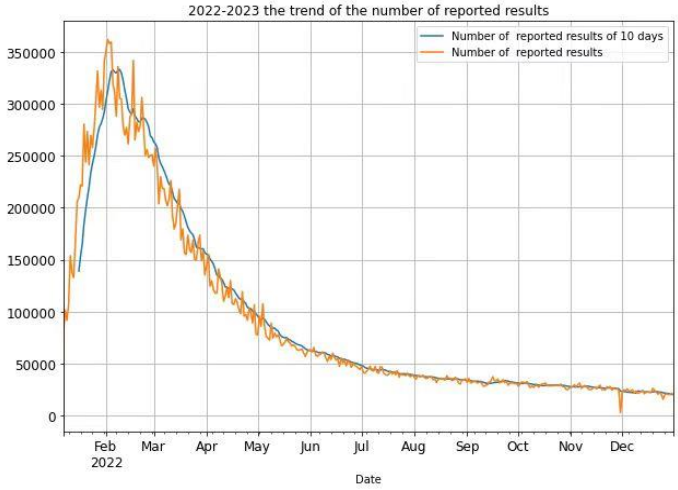

Figure 1. The current data in chronological order

2.2.2 ARIMA

From Figure 1, we can see that the number of reported results is changing dynamically from day to day, and we build the model to account for this change and predict the reported results for March 1, 2023 (the total player score on that day).

"ARIMA"——"AR" Autoregressive model [3,4,5]: The current daily reported result is a time series analysis problem. Predicting the result for March 1, 2023 requires the use of the given historical time data, and the autoregressive model is used to describe the relationship between the current and historical values. Equation (1) establishes a standard time-series relationship between current and historical values, which can well describe the trend of the total daily player score over time.

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t \tag{1}$$

"ARIMA"——"MA" Moving average method: Random perturbations in the error term in the autoregressive model can affect the accuracy of the model. In addition to time, there are still many factors affecting the results reported on that day in the current data, and the moving average method based on the standard autoregressive model can effectively eliminate the random fluctuations in the results forecast for March 1, 2023.

$$y_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \tag{2}$$

"ARIMA"——"I" Differencing: But the autoregressive model must satisfy the "smoothness" requirement. At the end of the figure 1, we find that the sample data seems to satisfy the "smoothness" characteristic, and the use of differencing makes the data smoother.

Combining equation (1) (2) we get the ARMA model, where the coefficient d determined by the difference method we can get through the diagram (detailed on the next page), so that we have a complete ARIMA (p,q,d) model.

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i \, y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \tag{3}$$

2.2.3 The Solution of ARIMA

For the solution of the ARIMA model, the most important thing is to determine the three parameters p,q,d. Here we choose the second order difference. The details are shown in Figure 2.

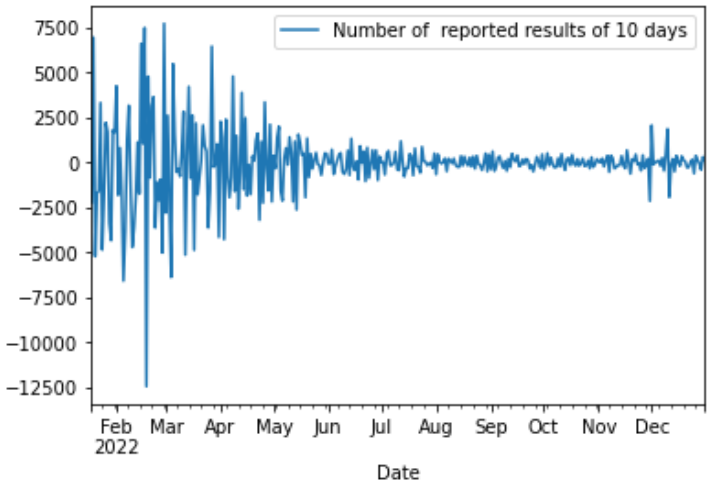

Figure 2. The results of the second-order difference——d=2

The establishment of p,q requires the use of autocorrelation and partial autocorrelation functions. p is established as the "p"th posterior truncated tail under the partial autocorrelation of the AR model (Equation (1)), and q is established as the "q"th posterior truncated tail under the autocorrelation of the MA model (Equation (2)). The details are shown in Figure 3.

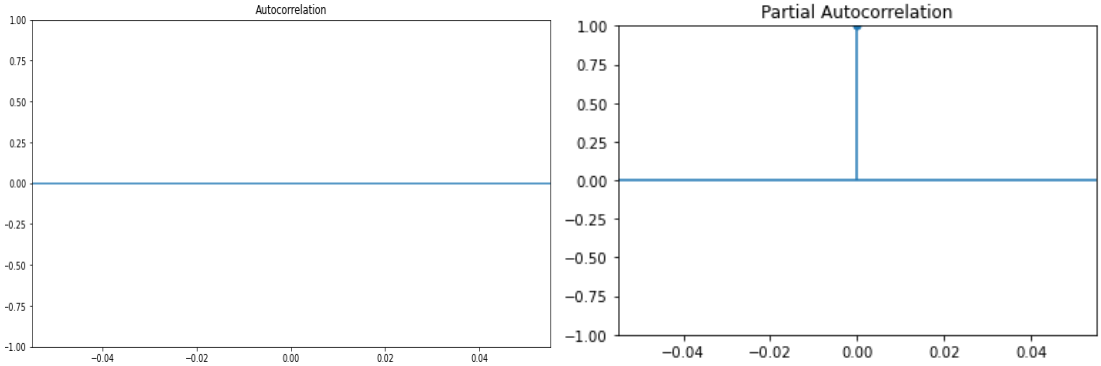

Figure 3. The results of different parameter

From the above Figure 2-3, we can see that d=2,q=0,p=1

$$y_t = \mu + \sum_{i=1}^{1} \gamma_i \, y_{t-i} + \epsilon_t + \sum_{i=1}^{0} \theta_i \epsilon_{t-i} \;\; (d = 2) \tag{4}$$

The prediction was then made (see the appendix for the specific code solution, due to space limitations the results are shown directly here): the expected range of total difficulty model scores on March 1, 2023 was obtained as（-7.251311e+05，7.406134e+05 ）

Note: Since our constructed model has the advantage of exact estimation, we give the point estimate results in order to improve the estimation accuracy. point estimate on March 1, 2023 reports the result as: 7741（See Appendix A for specific codes）

2.2.4 Word Attributes Exploration

In addition to the total daily score (reported results) given in the current data, the percentage data for each number of attempts before success is also presented, where we then explore whether the attributes of the words affect the percentage.

First of all, we divided the 359 words in the current data into 14 categories according to the two attributes of lexicality and the number of letters that make up the word (for example, EERIE consists of three letters E, R and I), and the naming rule for each category is lexicality + number of letters that make up the word (Note: the word given in EXCEL table D38 is "clen" which does not meet the 5 bytes, after looking up and checking the corresponding "clen" to "clean")are adj+5, adj+3+4, n+5, n+3+4, v+5, v+3+4, n. adj+5, n. adj+3+4, n. adj+4, n. adj+3+4, n. adj+3+4, n. adj+3+4. adj+3+4,

n.v+5, n.v+3+4, n.v.adj+5, n.v.adj+3+4, other+5, other+3+4. After that, the percentage of scores for the difficult mode was calculated.

We used the ANOVA model to calculate whether the data differed between the groups. The details are shown in Table 1.

Table 1. Differential analysis of data from different groups

ratio	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.012	13	.001	1.923	.027
Within Groups	.165	345	.000		
Total	.177	358			

As can be seen from Table 1, the significance of the test result is 0.27 at the 0.05 level of significance, and we choose to reject the original hypothesis and consider that there is a significant difference between the groups.

We explored the effect of each of the two attributes on the percentage of difficulty mode scores using scatter plots. The details are shown in Figure 4.

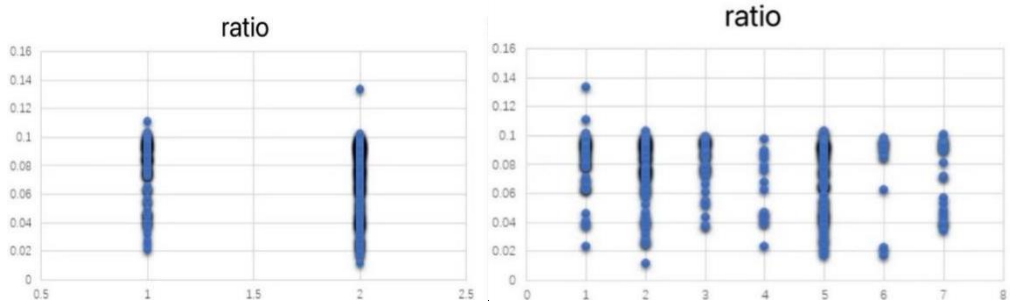

Figure 4. The effect of each of the two attributes on the percentage of difficulty mode scores.

Figure 4 (a):“1”represents words with repeating letters and“2”represents words without repeating letters. From the scatter plot, we can see that: the values of group 1 are mainly concentrated in 0.07-0.1, and the values of group 2 are mainly concentrated in 0.04-0.1. If the concentration range of group 2 is larger, the total mean value will be lower than that of group 1.

Figure 4 (b): Where 1~7 represent: adj, n, v, n.adj, n.v, n.v.adj, other lexical. We can get: the percentage of difficult mode scores of group 1 will have higher values generated; the concentration point of the percentage of difficult mode scores of group 4 is relatively down; the percentage of difficult mode scores of group 6 is more scattered, mainly concentrated at the two ends, around 0.02 and between 0.08-0.1; and the values of the percentage of difficult mode scores of groups 2, 3, 5 and 7 are relatively continuous, especially the data of groups 2 and 5 concentration range is greater.

In summary, we conclude that the attributes of the words have a significant effect on the percentage of scores reported that were played in Hard Mode.

3. Model establishment and solution

3.1. Exploration of the percentage distribution of "EERIE"

3.1.1 Data process

In this question we need to build a model to predict the percentage distribution of a specific word "ERRIE" in the future and to show how accurate the model is. First, we need to pre-process the data.

(1) the data given in the "Word" to extract, into a separate text format file

(2) this question to solve the "2023-03-01, eerie" to predict its "1 try" "2 tries" ... ... "7 or more tries(x)" results, the two belong to different languages, so we first pre-process the text data to build a 29-dimensional word vector, and then convert the time data into floating point numbers, and the word vector merged into a new data frame x

(3) Convert“x、y”into an array

3.1.2 Keras model building and solving

Since there are large differences in the dimensionality of the dependent variables of the independent variables, and they are all multidimensional data, we use a fully linked neural network model for machine training and model generation.[6]

We need to construct the various substructures of the model, first constructing the fully connected layer of the model, which is actually implemented in the forward propagation as a matrix multiplication operation：

$$\overrightarrow{a_i} = \overrightarrow{X_i}W \tag{5}$$

where (X_i) $\rightarrow$ is the input vector of our"i"th sample (assuming we have a total of n samples), W is the weight matrix, i.e., the weight of the connection between the neurons in the previous layer and the neurons in this layer, and (a_i) $\rightarrow$ is the output vector of this layer.

Next, in back propagation, we need to calculate the gradients of X and W according to the above equation.

$$\frac{\partial \overrightarrow{a_i}}{\partial \overrightarrow{X_i}} = W^T,\ \frac{\partial \overrightarrow{a_i}}{\partial W_{ab}} = \overrightarrow{X_i}^T \tag{6}$$

For a classification problem like word recognition, we want the model to classify different samples into pre-defined categories. The most common way to solve multi-category problems by neural networks is to set t output nodes, where t is the number of categories (in this problem, t = 7.) For each sample, the neural network can obtain a t-dimensional vector as the output, and each dimension in the vector is corresponding to a category for each output node. Ideally, if a sample belongs to category k, then the output value of the output node corresponding to this category should be 1, while the output of all other nodes is 0.

The effect of a neural network model and the goal of optimization are defined by the loss function, and for the same neural network, different loss functions can have an important impact on the model obtained by training. In classification problems, to describe how close an output vector is to an expected vector, we use a loss function called cross-entropy, which characterizes the distance between two probability distributions.

Given two probability distributions p and q, the cross-entropy of p, expressed through q, is

$$H(p,q) = -\sum_X p(x)\log q(x) \tag{7}$$

The above equation is rewritten to obtain the formula for the model cross-entropy loss function, where (Y_i) $\rightarrow$=(Y_i1, Y_i2...Y_in) represents the true value of the "i"th sample, and (P_i) $\rightarrow$=(P_i1, P_i2...P_in) represents the predicted value of the "i"th sample

$$Loss = -\overrightarrow{Y_i}\log\overrightarrow{P_i} = -\log P_{ij}\left(\overrightarrow{Y_i} = \left(0,0,\dots,Y_{ij},\dots 0\right), Y_{ij} = 1\right) \tag{8}$$

Then the gradient is found for it in back propagation as:

$$\frac{\partial Loss}{\partial \overrightarrow{P_i}} = \left(0,0,\dots,-\frac{1}{P_{ij}},\dots,0\right) = \left(-Y_{i1}\frac{1}{P_{i1}}, -Y_{i2}\frac{1}{P_{i2}},\dots,Y_{it}\frac{1}{P_{it}}\right) \tag{9}$$

Based on the above, using forward and backward propagation, we can combine the previously constructed modules to build our fully connected neural network model

We divided the training set and the test set among the processed data, and conducted machine training. Finally, we found that the accuracy of the model was 0.56.

Using this model, predict the data needed to solve the problem. And normalize the data and output into the desired form. The details are shown in Table 2.

Table 2. Normalize the output result table

Tiimes	Try 1	Tries 2	Tries 3	Tries 4	Tries 5	Tries 6	X
Original result	5.48365	12.042	19.9349	31.9569	24.4813	2.97655	0.013413
Normali zation	0.056597	0.124287	0.20575	0.329831	0.252674	0.030721	0.000138
Final result	6	12	21	33	25	3	0

3.2. Difficulty determination of "EERIE"

3.2.1 Data process

We set 3 factors that affect the difficulty of the words.

(1) INITIAL LETTER：We processed the data to get the following graph, the number of words starting with s is the largest at 15.44%, while the number of words starting with z is the smallest at 0.09%. Since the higher the overall frequency of the initial letter, the higher the guessing rate of the word with the initial letter, the lower the difficulty, and the lower the expected number of words, the smaller the value assigned to the letter with the higher frequency. The details are shown in Figure 5.

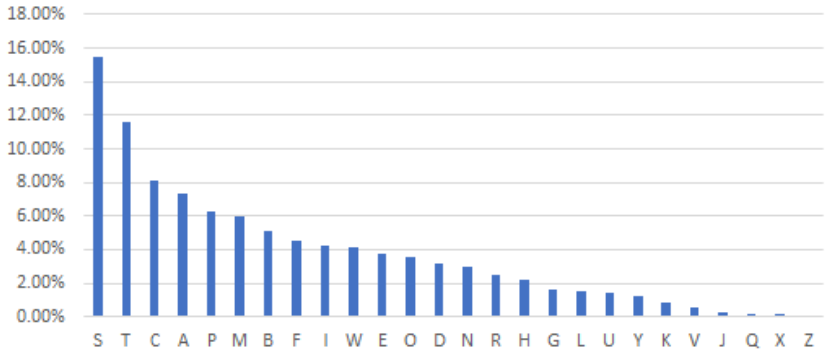

Figure 5. Initial processing result graph

(2) NUMBER OF LETTERS（For example, the number of letters in EERIE is 3）: Guessing words in Wordle will have color hints, if the letters are not in the right position will be realistic yellow, the more the number of letters appears yellow hints probability is greater, expect the number of cracks is smaller, can more easily help us locate the correct word. In the current 359 words in a word may have 3, 4, 5 lexical, respectively assigned to 3, 2, 1.

(3) NUMBER OF WORD ITEMS：For words with more lexical properties, we think they are more common and easier to guess with less difficulty, so 1, 2, 3 and 4 lexical properties are assigned as 4, 3, 2 and 1 respectively.

Then, the expected number of cracks for each word is calculated according to the formula, and the difficulty is divided into three grades: less than 3.5 is the lower difficulty, between 3.5 and 4.5 is the medium difficulty, and more than 4.5 is the higher difficulty. E(Tries)= $\sum p * Tries$

3.2.2 Decision Tree Model - Problem Solving

Based on the Gini coefficient, we build the CART decision tree [7,8] with the following equation:

$$gini(M) = 1 - \sum p_i^2 \tag{10}$$

Where pi is the frequency of category i in sample M, i.e., the ratio of samples with category i to the total number of samples.

Since we have 5 categorical variables, thus our Gini coefficient becomes

$$gini(M) = \sum_{i=1}^{5} \frac{s_i}{\sum_{i=1}^{5} s_i} gini(m_{i)} \tag{11}$$

Where $s_{i.}$ is the sample size of each of the five categories after division，$gini(M_i)$are the Gini coefficients for each of the five categories.

We then used the python program "sklearn.model" for machine learning[9] and prediction. We divided the data sets into 4 groups, groups 1-3 were the training group and group 4 was the test group,

and we performed classification practice to find the best nodes of the decision tree and the best branching method to reduce impurity and achieve importance test. Through python, we obtained the following results: the importance of the first letter is 0.376, the importance of the number of letter kinds is 0.235, and the importance of the number of lexical words is 0.389.

Finally, we took in the values of each feature of "EERIE", i.e., 30 points for those with three letters, 3 points for those with the initial letter e, and 4 points for those with only one type of word, and finally judged the difficulty of "EERIE" to be moderately difficult.

The accuracy rate of the model is about 80%, which indicates that the model is more reliable.

4. Conclusions

In order to do predictive analytics for "World", we first started collecting data in January to predict what will happen on March 1, 2023. The predicted results are (-7.251311e+05, 7.406134e+05) and 7741. Second, we built a date, word, and guess result model. Through this model, we finally find that March 1, 2023 is the date, take EERIE word vector as input, and then output the prediction result of the date. Our prediction is (1,2,3, 4, 5, 6,X) -- (6, 12, 21,33 years old,25 years old,3,0). Finally, we also study the factors that influence the difficulty of guessing words. We divide words into three categories: low difficulty, medium difficulty and high difficulty. Through modeling, we find that after introducing the characteristics of EERIE words into the model, it is classified as medium difficulty, so we predict that on March 1, 2023, the average number of guesses for EERIE words will be between 3.5 and 4.5.

References

[1] Yu Yang, Liu Chunyan, Cui Yanqun. Research on Handwritten Digit Recognition based on Convolutional Neural Network [J]. Information and Computer (Theoretical Edition),202,34(17):171-173.

[2] Jiang P. Deep Learn-based Sentiment Classification and its Application in public opinion Analysis [D]. Nanchang university, 2020. DOI: 10.27232 /, dc nki. Gnchu. 2020.003285.

[3] WU Xike. Research and Application of improved Neural network algorithm in prediction Method [J]. Computer and Digital Engineering,2022,50(10):2276-2279+2344.

[4] Yan Wang,Patricia Rodríguez de Gil,Yi-Hsin Chen,Jeffrey D. Kromrey,Eun Sook Kim,Thanh Pham,Diep Nguyen,Jeanine L. Romano. Comparing the Performance of Approaches for Testing the Homogeneity of Variance Assumption in One-Factor ANOVA Models [J]. Educational and Psychological Measurement,2017, 77(2).

[5] Liu Jinkun. Discussion on case teaching based on Fuzzy Neural network Algorithm [J]. University Education,2022,(12):93-96.

[6] WU Xike. Research and Application of Improved Neural Network Algorithm in Prediction Method [J]. Computer and Digital Engineering,202,50(10):2276-2279+2344.

[7] Suphirat C., Chomtee B.,Borkowski J.J.. Expected mean squares for model effects in the two-way anova model when sampling from finite populations [J]. Songklanakarin Journal of Science and Technology,2021,43(1).

[8] DING Wei. Sentiment Analysis Based on Combination of Dictionary and Machine Learning [D]. Xi 'an University of Posts and Telecommunications,2017.

[9] Jing Yuanyuan. Research on Decision Tree Algorithm Based on Decision Path [D]. Shandong University of Technology,2022.