

## Data prediction model for Wordle software based on regression analysis and SVM classifier

Yanlu Guo<sup>#,\*</sup>, Jiaqi Wang<sup>#</sup>, Peilin Liang<sup>#</sup>

Houston International Institute, Dalian Maritime University, Liaoning Dalian, China, 116026

\*Corresponding author: m18539851185@163.com

<sup>#</sup>These authors contributed equally.

**Abstract.** The field of gaming is an emerging field that receives a lot of attention nowadays, so its good development is important for the economy and society. In order to solve the potential problems of game development, this thesis collects the total number of users playing Wordle game, the difficulty of words and the number of attempts for the past 12 months, firstly pre-processing to eliminate the abnormal data, then using SPSS software to analyze the data from regression model and SVM model, using its dichotomous method to accurately fit the data, relying on the regression model The prediction model of the total number of users playing the game was established based on the principle of regression model, while the concept of SVM classification model was used to define the difficulty of the words of the Wordle game to establish the prediction model of the number of attempts per day for the users, and then the trend of the number of users and the number of attempts per day could be grasped through the two prediction models. Thus, Wordle game operators can take corresponding measures to maintain the game's popularity and ensure the number of users per day, which can help the game operators to generate more economic benefits.

**Keywords:** Wordle, Regression Analysis, Prediction Model, SVM, SPSS.

### 1. Introduction

In today's era of Internet of everything, various kinds of software are emerging, among which game software is especially prominent among Internet products, loved by more and more users, providing users with leisure and entertainment and having great potential for development. With the booming development of the game industry, the competitive pressure of similar games, the short life cycle of online games and the declining interest of users can also pose a threat to game developers, and how to retain users and achieve healthy and sustainable development of enterprises is the biggest problem faced by relevant enterprises at present. This paper takes Wordle game as an example and analyzes the development trend from its initial development to the explosion of the game and then the loss of the number of users, which eventually tends to stabilize, in order to predict and analyze the overall playing data of the game user group, and realize the prediction of the number of users playing on a specific date and the number of attempts to solve the target word on that day, so that Wordle game operators can better maintain the existing user base and maintain the novelty of the game's unique word solving model.

### 2. Materials and Methods

#### 2.1. Data acquisition and pre-processing

This thesis selects Wordle game as the research object, for this purpose, the recent 12 months of user data were downloaded from its official website, including the specific date, the number of users playing online every day, the words of the day and the percentage of attempts to answer words by playing users, based on this summary of user data and pre-processing work, mainly screening out the individual abnormal data, such as the number of The 359 valid data are all continuous in time and have obvious longitudinal reference value, so they are useful for the establishment of prediction models.

## 2.2. Method Introduction

### 2.2.1 Regression prediction model

The essence of the regression prediction model is to use regression analysis to statistically analyze the expression of the function obtained by curve fitting, so as to verify the credibility of the regression model function and judge whether each independent variable has a significant effect on the dependent variable, then as long as it is proved that the function obtained by curve fitting meets the credibility test of the regression model, this function can be used to calculate the prediction of the number of users playing online on a specific date[1]. The specific validation method is to do interval estimation or hypothesis testing for the obtained function to determine whether the confidence interval range is reliable and whether there are no anomalies[2].

For the solution of the curve fitting function, this study will use MATLAB to implement, and then use SPSS software to achieve statistical analysis of the sample data set, so as to verify the reliability of the function, as long as the significance between the variables is significant enough to use the coefficients of the function to predict the required user data.

### 2.2.2 SVM prediction model

SVM is a generalized linear classifier that performs binary classification of the target data set in a supervised learning manner, and its decision boundary is the hyperplane of the maximum margin solved for the sample data set. The support vector classifier has the characteristics of both sparsity and stability. The sparsity is reflected in the fact that the hyperplane of the binary classification is accurate and infinite, and any sample points are spaced from each other and distant from the hyperplane of the classification, while the stability is reflected in the fact that any sample adapted to this classifier can be used for this classification criterion and will not be easily broken. These two characteristics are very suitable for the Wordle sample dataset in this study, which can accurately classify the existing sample data and maintain a uniform standard for classifying new input data.

The SVM classifier is able to classify all words in the sample dataset into one difficulty level, and then the variance chi-square test is performed for each word of different difficulty levels to obtain the normalized distribution pattern of the number of user attempts, and then the distribution pattern is used in combination with the segmentation sub-test of SPSS software to predict the percentage of user attempts for the target words.

## 2.3. Model Evaluation Metrics

In this study, the influence of four factors on the difficulty level of words was explored, and the importance of the four factors can be assigned with the information from the references and official annual reports as shown in Table 1 below[3], so that the percentage distribution of the total number of users and the number of attempts to predict special words[4].

**Table 1.** Symbolic representation of each of the four influencing factors and weight assignment[5].

| Symbol | Description                | Weight |
|--------|----------------------------|--------|
| m      | Repeated number            | 15%    |
| n      | Initial letter's frequency | 25%    |
| p      | Monogram's frequency       | 30%    |
| q      | Common letter's frequency  | 30%    |

## 3. Results and analysis process

### 3.1. Model 1: Regression prediction model

Firstly, the pre-processed data will be corresponded one by one in Excel to form a data table, and then the data table will be imported into SPSS to create a scatter plot as shown in Figure 1, thus making the data visualized[6].

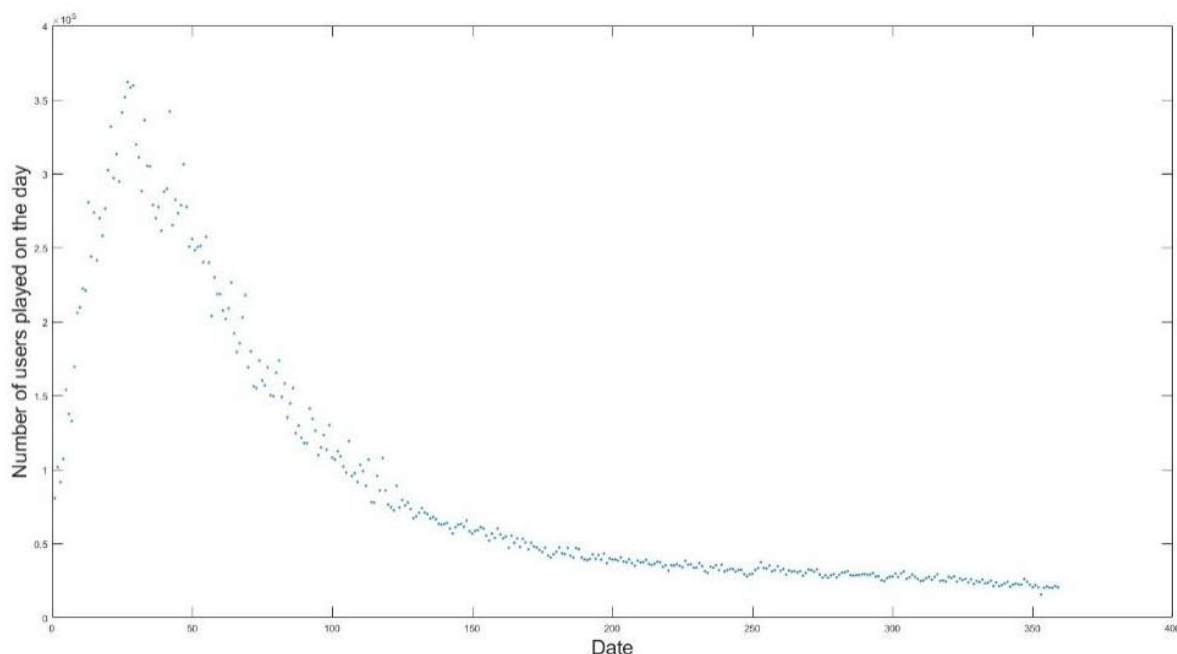


**Figure 1.** Data scatter plot.

As can be seen from the figure, the overall number of users playing daily surges at the beginning, starts to decline significantly after reaching a certain number, and then slowly declines and stabilizes. Subsequently, by analyzing the results presented in the scatter plot, we found that the data can be divided into two parts and thus MATLAB fitting analysis, the first part is the rapid growth part at the beginning as shown in Figure 2, which is selected as a cubic polynomial function model; the second part is the decreasing part at the later stage as shown in Figure 3, which is selected as an exponential and cubic polynomial function model[7].

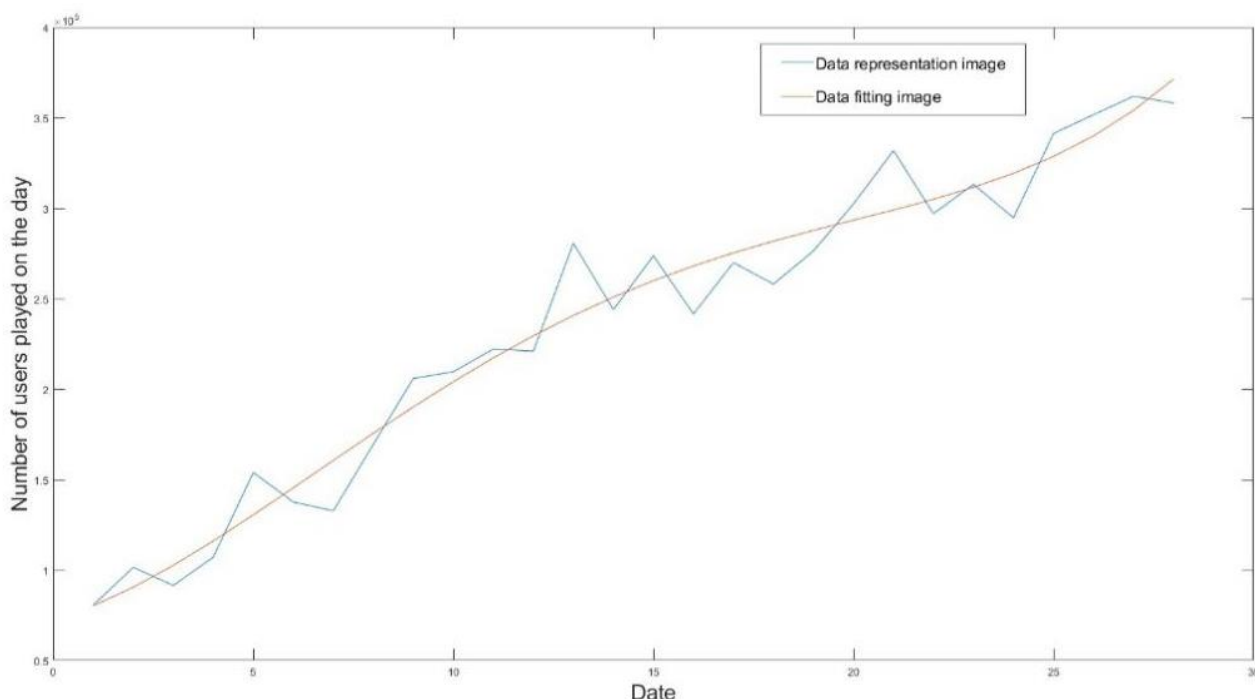


**Figure 2.** A fitting graph of data growth trends.

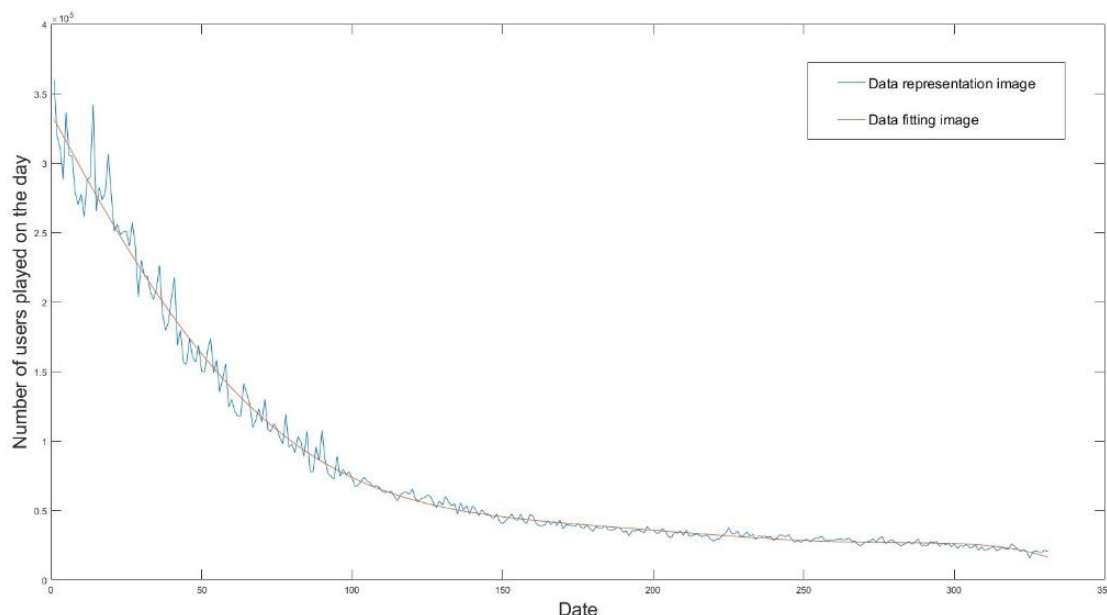


**Figure 3.** A fitting graph of data decline trends.

From the scatter plot, we can get the dividing point of the two trends is the number of users corresponding to 2022.02.04. With this as the dividing point, we can get the required function model for the two fits, and then carry out spss regression analysis on the data of the number of users respectively, and we can get the prediction results of the three-time function and the exponential function as shown in Table 2 below[8, 9].

**Table 2.** A table of data that fits the parameters of the function.

| Dependent variable: V4 |           |                  |                     |                     |              |                      |         |        |       |
|------------------------|-----------|------------------|---------------------|---------------------|--------------|----------------------|---------|--------|-------|
| Model Summary          |           |                  |                     |                     |              | Parameter Estimation |         |        |       |
| Equation               | R Squared | F                | Degree of freedom 1 | Degree of freedom 2 | significance | constant             | b1      | b2     | b3    |
| Cubic                  | 0.773     | 16<br>3.2<br>1.7 | 3                   | 144                 | 0.000        | 19209.288            | 182.042 | -0.992 | 0.004 |
| Exponential            | 0.442     | 11<br>5.5<br>59  | 1                   | 146                 | 0.000        | 20520.410            | -0.004  |        |       |

The significance of both functions is 0, so the significance is good that the two function models have a good fit for the existing data set, but the R-squared of the cubic polynomial is larger than that of the exponential, closer to 1, which proves that the fit is better for the existing data cubic polynomial, but because of the need to predict the data after a long date, analysis and comparison of the cubic polynomial function and the exponential function can be found that the cubic This will lead to a mismatch between the model and the actual problem. Therefore, for predicting the number of users after a specific date, the exponential function is chosen in order to meet the real situation and ensure that the prediction results are closer to reality and more accurate.

When analyzing the image trend, the research also found that the factors affecting the total number of users also include the users' own subjective factors and objective factors such as word difficulty, which can also lead to errors between the actual and predicted values.

### 3.2. Explanation of model evaluation indicators

Through relevant linguistic literature and the data survey of Wordle game official website, this thesis concludes that four factors which influence the difficulty level of words relatively significantly

are the number of letter repetition index, the frequency of initial letter occurrence index, the frequency of common letter occurrence index and the frequency of common letter combination index, and the score range of each factor is defined as 0-3, then the total score after weighting is in the range of 0-12. Then, the total score was divided into four levels of difficulty according to the total score, with the following criteria: easy 9-12 points, medium 6-9 points, difficult 3-6 points, and extremely difficult 0-3 points, and then the four factors were analyzed one by one.

The first factor is the number of letter repetition index, in this thesis, the number of letter repetition in a word is inversely related to the difficulty of the word, the more the number of letter repetition, the more difficult the word is, so if three identical letters appear in a word, the value is 1 point, two identical letters are assigned 2 points, one identical letter is assigned 3 points, and no identical letters appear. The first influence factor can be quantified after this process.

The second factor is the frequency of the initial letter. By checking the official annual usage report of the word, we know that the frequency of different initial letters appearing in the first position of the word is different, so we can assign the 26 letters from 3 to 0 points according to the frequency of appearing in the initial letter, as shown in Table 3.

**Table 3.** Assignment table for frequency of occurrence of initials.

| Initial | Weight | Initial | Weight |
|---------|--------|---------|--------|
| x       | 0      | n       | 1.3    |
| y       | 0.1    | e       | 1.4    |
| z       | 0.2    | h       | 1.5    |
| q       | 0.3    | r       | 1.6    |
| j       | 0.4    | d       | 1.7    |
| k       | 0.5    | u       | 1.8    |
| v       | 0.6    | b       | 1.9    |
| w       | 0.7    | t       | 2      |
| l       | 0.8    | m       | 2.1    |
| g       | 0.9    | a       | 2.2    |
| o       | 1      | c       | 2.3    |
| i       | 1.1    | p       | 2.4    |
| f       | 1.2    | s       | 2.5    |

The third factor is the frequency of occurrence of commonly used letters, and the second factor is the same as the frequency of occurrence of initial letters, through the official annual report data can be seen that the frequency of use of different letters in the composition of the word is different, which can also be assigned to the 26 letters of the alphabet, as shown in Table 4.

**Table 4.** Assignment table for the frequency of common letters.

| Letter | Weight | Letter | Weight |
|--------|--------|--------|--------|
| e      | 2.6    | a      | 2.5    |
| r      | 2.4    | o      | 2.3    |
| t      | 2.2    | l      | 2.1    |
| i      | 2      | n      | 1.9    |
| s      | 1.8    | c      | 1.7    |
| h      | 1.6    | p      | 1.5    |
| u      | 1.4    | y      | 1.3    |
| m      | 1.2    | d      | 1.1    |
| g      | 1      | k      | 0.9    |
| f      | 0.8    | b      | 0.7    |
| w      | 0.6    | v      | 0.5    |
| x      | 0.4    | q      | 0.3    |
| z      | 0.2    | j      | 0.1    |

The fourth factor is the frequency of common letter combinations, which is selected from the 24 monigram attributes in Table 5, excluding the letter combinations with no obvious pattern and those that are relatively rare, and then assigning a value to each of them. If a word contains these common letter combinations, it is assigned a corresponding score, and the highest priority principle is followed in the assignment process, for example, if there are both ai and ee, it is assigned a score of 3.

**Table 5.** Assignment table of the frequency of common letter combinations.

| Letters | Weight | Letters | Weight |
|---------|--------|---------|--------|
| ai      | 3      | oa      | 3      |
| ee      | 2      | ea      | 3      |
| ie      | 2      | ei      | 3      |
| oo      | 1      | oi      | 3      |
| ou      | 3      | au      | 1      |
| ar      | 2      | ch      | 1      |
| er      | 2      | gh      | 1      |
| ir      | 2      | ph      | 2      |
| or      | 1      | sh      | 1      |
| ur      | 1      | th      | 3      |
| aw      | 0      | ay      | 1      |
| ew      | 0      | ey      | 1      |

### 3.3. Model 2: SVM prediction model

The SVM support vector classifier is a mathematical prediction model with four influencing factors and word sample dataset as inputs and word difficulty classification levels as outputs, followed by defining the required inputs as well as outputs for the model in turn[10].

Based on the given training set, the parameter T is defined to represent,

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\} \in (X \times Y)^l, \quad (1)$$

For this formula  $x_i \in X = R^n$ ,  $X$  is called the input space, and each point in the input space  $x_i$  consists of  $n$  attribute features,  $y_i \in Y = \{-1, 1\}$ ,  $i = 1, \dots, l$ ,  $Y$  is called the output space, while in this set of data training set  $X=359$ ,  $n=4$ , thus establishing the definition of the concept of training set. Then perform the scoping of the training set  $T$ , because of  $\exists \omega \in R^n, b \in R$  and positive numbers  $\varepsilon$ , such that for all subscripts  $i$  that make  $y_i=1$  have  $(\omega \cdot x_i) + b \geq \varepsilon$ , and for  $y_i=-1$  the subscript  $i$  has  $(\omega \cdot x_i) + b \leq -\varepsilon$ , then this training set  $T$  is linearly divisible.

Denote the two types of sample sets as

$$M^+ = \{x_i | y_i = 1, x_i \in T\}, M^- = \{x_i | y_i = -1, x_i \in T\}.$$

Define the convex package of  $M^+$  is  $\text{conv}(M^+)$

$$\text{conv}(M^+) = \{x = \sum_{j=1}^{N_+} \lambda_j x_j | \sum_{j=1}^{N_+} \lambda_j = 1, \lambda_j \geq 0, j = 1, \dots, N^+; x_j \in M^+\}, \quad (2)$$

The convex package of  $M^-$  is  $\text{conv}(M^-)$

$$\text{conv}(M^-) = \{x = \sum_{j=1}^{N_-} \lambda_j x_j | \sum_{j=1}^{N_-} \lambda_j = 1, \lambda_j \geq 0, j = 1, \dots, N^-; x_j \in M^-\}. \quad (3)$$

Where  $N^+$  denotes the number of +1 class sample points,  $N^-$  denotes the number of sample points of class -1, from the sufficient conditions for the linear differentiability of the training set  $T$ , we know that the convex packages of  $M^+$  and  $M^-$  are separated from each other, and there must

exist a unique canonical hyperplane  $H = \{x \in R^n : (\omega \cdot x) + b = 0\}$  and  $\varepsilon > 0$  that always holds, therefore this hyperplane can be used as our criterion to classify the sample set. In order to ensure the accuracy and uniqueness of the classification criteria, the following verifies that the convex package of  $M^+$  and  $M^-$  are all separated and do not overlap convex packets, also normative hyperplane exists and is unique.

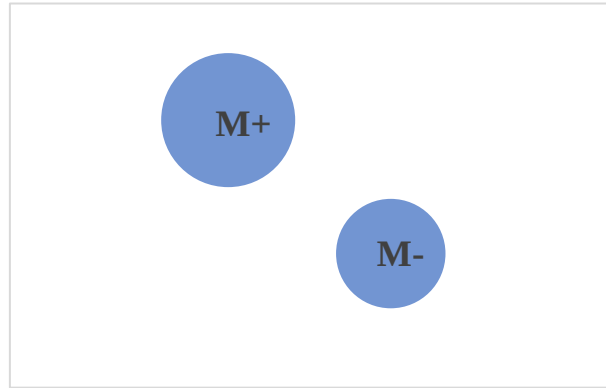


**Figure 4.** Graphical representation of the convex packet of two types of sample hits when the training set T is linearly divisible.

When the canonical hyperplane exists and is unique, it satisfies the following verification equation

$$\begin{cases} (\omega \cdot x_i) + b \geq 1 & y_i = 1; \\ (\omega \cdot x_i) + b \leq -1 & y_i = -1. \end{cases} \quad (4)$$

Verify that  $x_i$  satisfying  $(\omega \cdot x_i) + b = \pm 1$  in the formula becomes an ordinary support vector, for the SVM linearly differentiable case, only they play a role in building the classification hyperplane, while the ordinary support vectors represent only a small fraction of the sample set, therefore, it also shows that SVM has sparsity, Thus it can be defined that for sample points of class  $y_i = 1$ , their interval from the canonical hyperplane is

$$\min_{y_i=1} \frac{|(\omega \cdot x_i) + b|}{\|\omega\|} = \frac{1}{\|\omega\|}, \quad (5)$$

For sample points of class  $y_i = -1$ , their interval from the canonical hyperplane is

$$\min_{y_i=-1} \frac{|(\omega \cdot x_i) + b|}{\|\omega\|} = \frac{1}{\|\omega\|}, \quad (6)$$

Then the interval of the ordinary support vector is  $\frac{2}{\|\omega\|}$ , optimal hyperplane means maximizing  $\frac{2}{\|\omega\|}$ . As shown in Figure 5

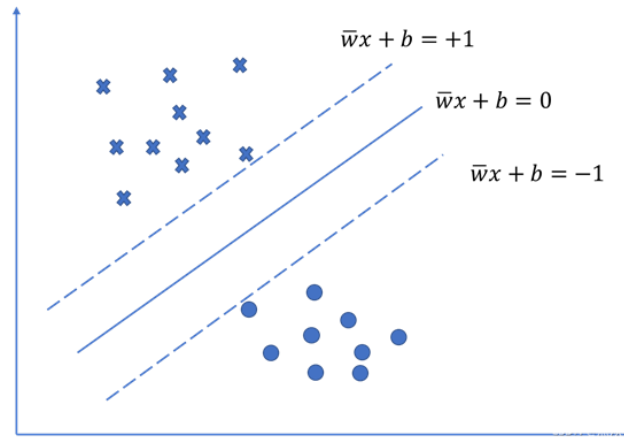


**Figure 5.** Graphical model of linear separable support vector classifier.

$(\omega \cdot x) + b = \pm 1$  is called the classification boundary, and it is the classification boundary that this model needs to obtain, so the problem of finding the optimal hyperplane can be transformed into the following quadratic programming problem.

$$\begin{aligned} \min \quad & \frac{1}{2} \|\omega\|^2 \\ \text{s.t.} \quad & y_i((\omega \cdot x_i) + b) \geq 1 \quad i = 1, \dots, l \end{aligned} \quad (7)$$

This objective function  $\frac{1}{2} \|\omega\|^2$  is a convex function of  $\omega$  and the constraints are linear, for which the Lagrange function is introduced for simplification

$$L(\omega, b, \alpha) = \frac{1}{2} \|\omega\|^2 + \sum_{i=1}^l \alpha_i (1 - y_i((\omega \cdot x_i) + b)) \quad (8)$$

Where  $\alpha = (\alpha_1, \dots, \alpha_l)^T \in R^{l+}$  is the Lagrange multiplier, and according to the definition of the wolfe dual, by setting the partial derivatives of each variable in the original problem to zero we get

$$\begin{aligned} \frac{\partial L}{\partial \omega} = 0 \quad \Rightarrow \quad \omega &= \sum_{i=1}^l \alpha_i y_i x_i \\ \frac{\partial L}{\partial b} = 0 \quad \Rightarrow \quad \sum_{i=1}^l \alpha_i y_i &= 0 \end{aligned} \quad (9)$$

Lagrange pairing problem with Lagrange function into the original problem

$$\begin{aligned} \max_a \quad & -\frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j a_i a_j (x_i \cdot x_j) + \sum_{i=1}^l a_i \\ \text{s.t.} \quad & \sum_{i=1}^l y_i a_i = 0 \\ & \alpha_i \geq 0, i = 1, \dots, l \end{aligned} \quad (10)$$

Solve the above optimization problem and obtain the optimal solution  $\alpha^* = (\alpha_1^*, \dots, \alpha_l^*)^T$ , calculating

$$\omega^* = \sum_{i=1}^l \alpha_i^* y_i x_i \quad (11)$$

From the KKT complementarity condition, get that

$$\alpha_i^*(1 - y_i((\omega^* \cdot x_i) + b^*)) = 0 \quad (12)$$

The corresponding  $\alpha_i^*$  is positive only when  $x_i$  is a support vector, otherwise it is zero. Choose a positive component  $\alpha_j^*$  of  $\alpha^*$  and use it to calculate

$$b^* = y_j - \sum_{i=1}^l y_i \alpha_i^* (x_i \cdot x_j), \quad (13)$$

The classification hyperplane  $(\omega^* \cdot x) + b^* = 0$  is then constructed, and the decision function is thus derived

$$g(x) = \sum_{i=1}^l \alpha_i^* y_i (x_i \cdot x) + b^*, \quad (14)$$

Following that, the classification function is obtained

$$f(x) = \text{sgn}\left(\sum_{i=1}^l \alpha_i^* y_i (x_i \cdot x) + b^*\right) \quad (15)$$

The overall word data set can be divided into two categories by the calculation of the primary classification function, and then the sub-sample set after the primary classification can be more carefully bifurcated by using the SVM linear vector classifier again[11], thus we can get the four categories of difficulty classification we need: easy, medium, hard, and extremely hard, and our classification results are shown graphically in Figure 6 which is the form of a tree diagram below:

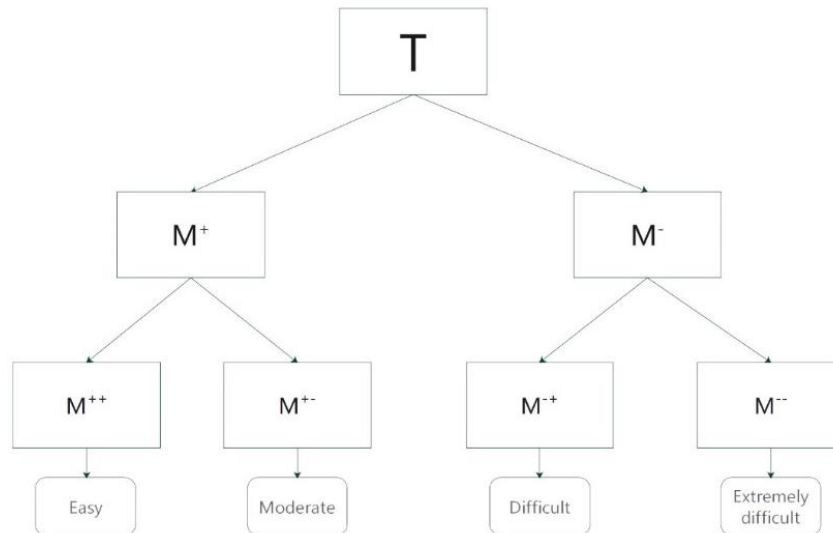


**Figure 6.** Flow chart of SVM model for classifying word difficulty.

The classification model can thus classify the words to be predicted in terms of difficulty, and after categorizing them into four difficulty levels, the percentage distribution of the number of attempts for a particular word can be predicted based on the normalized distribution law corresponding to the respective difficulty level obtained from the SPSS analysis. Here the EERIE word is selected as an example for prediction analysis which shown in Figure 7, and its classification result obtained according to the SVM vector classifier model is M+-, that is, the result of the first classification function is +1, the result of the second classification function is -1, and the final difficulty is defined as medium.

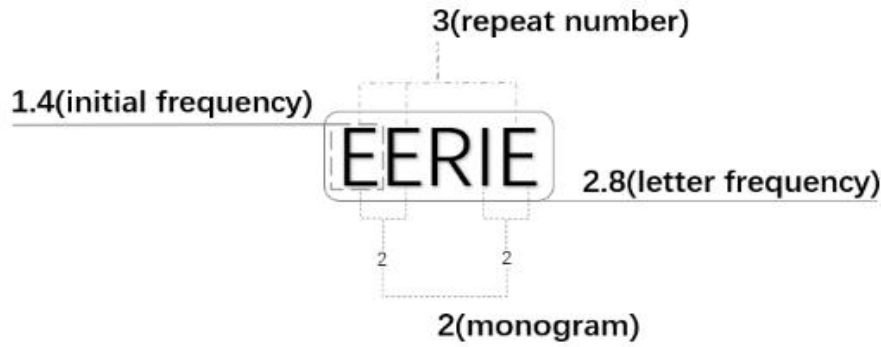


**Figure 7.** EERIE word specific score chart of influencing factors.

According to the SPSS summary rule for the normalized distribution of the number of attempts of moderately difficult words, the following formula for calculating the respective number of attempts can be obtained

$$a_1 = -0.69 * m + 0.098 * n + 0.196 * p + 1.095 * q - 2.166 \quad (16)$$

$$a_2 = -4.2 * m + 0.848 * n + 1.281 * p + 9.012 * q - 15.644 \quad (17)$$

$$a_3 = -6.743 * m + 2.122 * n + 1.989 * p + 10.364 * q - 4.172 \quad (18)$$

$$a_4 = 3.045 * m - 0.222 * n + 0.503 * p - 4.479 * q + 48.116 \quad (19)$$

$$a_5 = 5.170 * m - 1.788 * n - 1.396 * p - 9.643 * n + 48.004 \quad (20)$$

$$a_6 = 2.243 * m - 1.06 * n - 1.574 * p - 5.198 * q + 25.903 \quad (21)$$

$$a_7 = 1.273 * m + 0.072 * n - 0.944 * p + 0.148 * q + 3.571 \quad (22)$$

After that, based on the regression results of this model, the distribution of the number of occurrences of the word EERIE is predicted, and the following results can be obtained as shown in Table 6

**Table 6.** Distribution prediction table of the number of EERIE word attempts.

| 1 try | 2 tries | 3 tries | 4 tries | 5 tries | 6 tries | 7 or more tries |
|-------|---------|---------|---------|---------|---------|-----------------|
| 0%    | 1%      | 11%     | 40%     | 31%     | 12%     | 6%              |

## 4. Conclusion

This study uses regression prediction method and SVM support vector machine classification prediction method to predict the data of Wordle, a specific game software, using the fitting prediction characteristics of the regression model to find the fitting trend function of the data, so as to obtain the total number of users of the game on a specific date, as well as using the characteristics of SVM accurate classification to define the difficulty level of words and thus achieve the percentage of attempts of a specific word prediction. These two prediction data can be continuously optimized to ensure the novelty and fun of Wordle software, which is a good reference value for the healthy and sustainable development of the game.

## References

- [1] Vo Nhi N.Y., Liu Shaowu, Li Xitong et al. Leveraging unstructured call log data for customer churn prediction [J] Knowledge-Based Systems, 2021, 212.
- [2] Baghla Seema and Gupta Gaurav. Performance Evaluation of Various Classification Techniques for Customer Churn Prediction in E-commerce [J]. Microprocessors and Microsystems, 2022, 94.
- [3] Simone A. du Toit, Saif A. Kade, Craig T. Danielson et al. The effect of emotional working memory training on emotional and cognitive outcomes in individuals with elevated social anxiety[J] Journal of Affective Disorders, 2020, 261(C).
- [4] Herlambang Mega B. Taatgen Niels A. Cnossen Fokie Modeling motivation using goal competition in mental fatigue studies[J] Journal of Mathematical Psychology, 2021, 102.
- [5] Marcello A. Gómez-Maureira, Isabelle Kniestedt Exploring video games that invoke curiosity [J] Entertainment Computing, 2019, 32(C).
- [6] Zhao Huiqin,Zhu Jianping. How to calculate the results of factor analysis applications with SPSS software[J]. Statistics and Decision Making,2019,(20):72-77.
- [7] Zhang Hanxia. Scenario analysis applied to linear regression and logistic regression [J]. Automation and Instrumentation,2022,(10):1-4+8.
- [8] Hou S. L., Wang J. G., Wang J. L.. Support vector machine based multi-target classification and identification[J]. Firepower and Command Control, 2016, 41(09): 189-192.
- [9] Guo Hongfei.Exploration of the application of F-test and T-test in the process of method validation[J]. Shanxi Metallurgy,2019,(04):114-116.
- [10] Yi, Chaoqun. Support vector classification machine based on feature analysis [D]. Supervisor: Li Jianping. University of Defense Science and Technology, 2010.
- [11] Zhang Fugui,Ye Lei,Li Delun,Wu Xuemei. A tobacco grade classification model based on RF-PSO-SVM[J]. Computer and Digital Engineering,2022,(08):1833-1838.