Problem Chosen	2023	Team Control Number
C	MCM/ICM Summary Sheet	2314130

AI Improves Wordle Game Results: Leveraging ML Algorithms to Analyze Game Data

Summary

Artificial intelligence can be applied in Wordle games, and analyze the game performance data feedback from Twitter to objectively evaluate players' performance and improve the game results. For Question 1 and 2, random forest and RNN/Spearman/Linear Regression models are used for prediction respectively; for Question 3, seven SVR models are used for prediction; for Question 4, K-Means/SVM models are used; finally, outliers are mined and part-of-speech is counted.

For Model I, K-fold Cross-validation was used to partition the training set and test set in order to alleviate the small dataset issue. Random forest algorithm was used to fit the distribution relationship between Contest number and various features and many random forest models with different hyperparameters were established. Multiple decision tree models were independently and parallelly trained in each forest and were combined together in a Bagging way to greatly reduce the variance of the model. Finally, the report result quantity range of March 1,2023 was predicted to be [28000, 31000] and the MSE value was 0.001 in the evaluation result of the model.

For Model II, in order to analyze the features related to Hard Mode Percentage Scores from the data, the first thing to do is to look at the words themselves. Utilize the RNN model to transform the words into temporal and semantic information without loss. Because the training dataset is too small, transfer learning was used to take the pre-trained RNN model and to retrain with a smaller learning rate. By analyzing the information with Hard Mode Percentage Scores, it turned out that they are not related. Using the Spearman algorithm for correlation analysis, it was found that Hard mode percentage scores has a strong correlation with Contest number. Finally, the two were subjected to univariate linear regression to predict that the Hard mode percentage score on March 1, 2023 was 0.132.

For Model III, the data is firstly dimension-reduced and two features related to the frequency distribution of Hard mode percentage scores and Contest number are analyzed. Seven SVR algorithm models are used in series to predict the value of each part one by one, and the test samples of each model come from the prediction results of the previous model. Finally, the following results were obtained: 1 try was 0%, 2 try was 4.44%, 3 try was 22.11%, 4 try was 34.89%, 5 try was 24.51%, 6 try was 10%, and 7 or more tries (X) was 3.02%. The MSE evaluation results of models 1-7 were 0.5, 9, 22, 12, 4, 7, and 4 respectively, and due to the series connection of these models, they are extremely unstable, but I still have certain expectations for it.

For Model IV, after analyzing and dimensionality reducing the dataset, the difficulty of words is only related to Hard Mode percentage scores, 1-6 try and 7 or more tries (X). K-Means algorithm reveals that when K=3, the data distribution of each classification is good, thus the difficulty level of words is defined as “Easy”, “Medium” and “Difficult”. Then based on 3 categories of data, SVM model is trained and predicts that the possibility of EERIE being Easy is 99%. In the evaluation result of SVM model, all the indicators are greater than 0.9, which indicates that the classification effect of this model is good.

Based on the discrete points in the raw data, additional information is analyzed, such as the update of the server on November 30th, 2022. After that, the part-of-speech of all words in the data file was counted and analyzed, and most words in the Wordle game came from news articles or reports, because they usually take an objective description as the main.

Keywords: Random Forest; RNN; Spearman's Correlation Coefficient; SVR; K-Means; SVM

Contents

1 Introduction

1.1 Problem Background

Wordle is a popular and trendy spelling game that not only requires no expensive gaming equipment nor vast online social networks for players, but also can improve their language skills, providing them with an easy and interesting way to learn and enhance their English ability. However, Wordle also has a potential social meaning, it changes players' attitude towards language learning, increasing their interest in words and strengthening their cognition of English. Wordle helps players to establish good language learning habits, which will contribute to their future English learning and enhance their learning ability, master English skills, and broaden their career prospects.

Artificial Intelligence can be applied in Wordle games by giving the AI model the context and specific objective and letting it calculate the best guess of the game. It can gain the optimal results according to the guess times, the correct letters usage and the game rules. AI techniques can also be used for analyzing the game results data from the feedbacks on Twitter, to evaluate the player's performance in different modes and improve the game result in a more comprehensive and objective way.

1.2 Restatement of the Problem

We built multiple models based on the topic and each of them focused on solving the following problems:

- This study analyzed the correlation between certain features in the data and the number of reported results, and established a predictive estimation interval for it on March 1, 2023.
- This research aims to identify the features in the data files that are related to words, and to determine which of these features have an effect on the percentage of people who select the difficult mode. The research will then point out which features have what impacts.
- We analyze the previous data distribution of EERIE to predict the data distribution on March 1st, and then evaluate the model.
- This paper aims to classify words according to their difficulty, identify the features that form the basis of the classification, and determine the difficulty of EERIE.

1.3 Our work

Our approach to this article is illustrated in the following diagram:

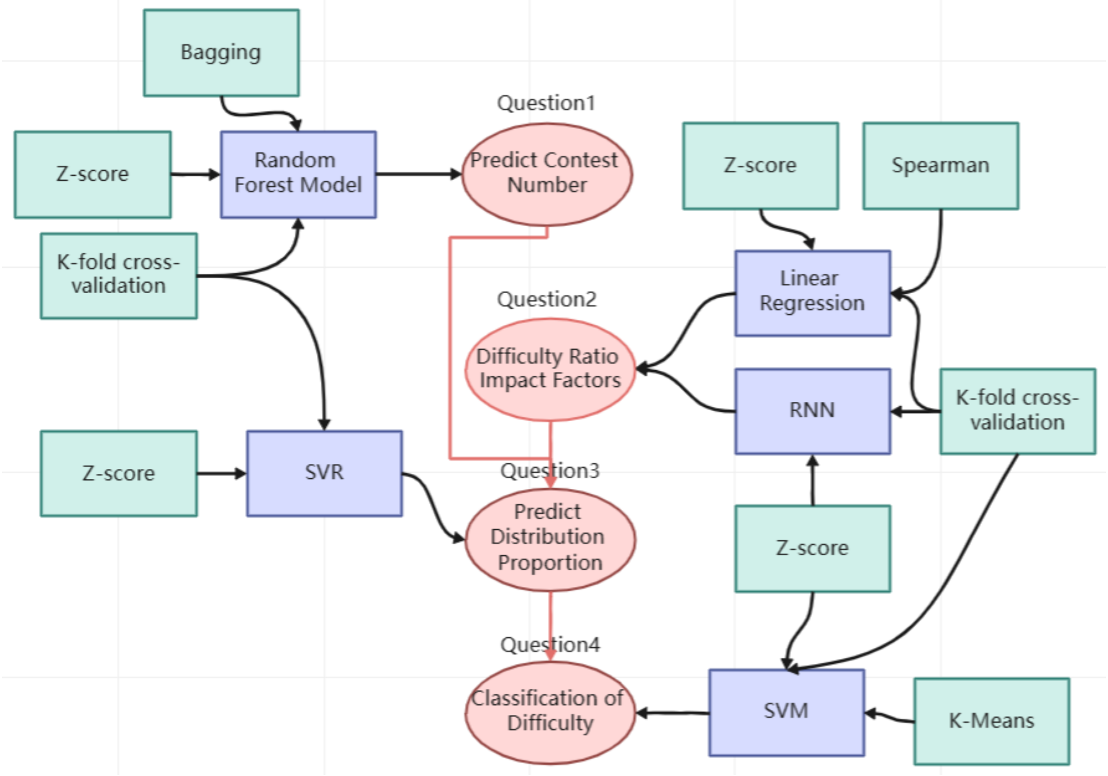


Figure 1: Model Overview

2 Assumptions and Justifications

In order to simplify the problem and make the model more general, we should assume the following assumptions to be valid by default.

➢ **Assumption 1:** There is no drastic change in the social environment.

Justification: The data records the attitudes of people living in the current social environment towards leisure activities. If a similar event such as a financial crisis occurs that can affect the entire social ecology, people's attitudes towards leisure or recreation may change drastically, resulting in a drastic decrease in the precision rate of the predictive algorithm model based on the previous data.

➢ **Assumption 2: No unexpected interactions between players and games.**

Justification: When playing Wordle, players can see the letters in each space and they have six attempts to guess the word. After each word submission, they get feedback with yellow tiles indicating that the letter in the tile is in the word but in the wrong position, green tiles indicating that the letter in the tile is in the word and in the correct position, and gray tiles indicating that the letter in the tile is not at all contained in the word. The above rules are followed normally.

➢ **Assumption 3: Assume players can solve the problem by guessing and eliminating**

Justification: Players have enough intelligence to effectively analyze the problem, classify the possible words, eliminate the possible words according to the feedback information, and think carefully after each attempt in order to finally solve the problem.

3 Notations

The mathematical symbols and their meaning used in this paper are shown and explained in the following table 1.

Table 1: Mathematical symbol used in this paper

Symbol	Description
σ_X	Standard deviation
X	Training sample
E(X)	Mathematical expectation
T	Aggregate data
s_t	The state information of a certain point in time
ρ	The correlation between two variables
$\overline{x}$	The average of a sample
$\overset{\wedge}{\alpha}$	The true label in supervised learning
$\overline{\mathrm{C}}$	The clustering center in K-Means algorithm
MSE	Expected squared difference between predicted and actual value.
RMSE	The expectation value of the difference
MAE	The average absolute error
MAPE	A transformation of MAE
K	The number of initial clustering centers selected in K-Means
ω	Normal vector in any hyperplane
b	Constant in formula
C	Penalty factor
ξ	Relaxation variable

4 Model Preparation

Data and features determine the upper limit of machine learning, and the data simulated by models and algorithms can only approach the upper limit infinitely. In order to ensure that the model training effect is close to the truth, it is necessary to clean the data and obtain the main features for model training. This data set is the same type of data and has excellent integrity, so there is no missing value. Therefore, we detect noise values in the data, and finally scientifically display the distribution of the data of this label.

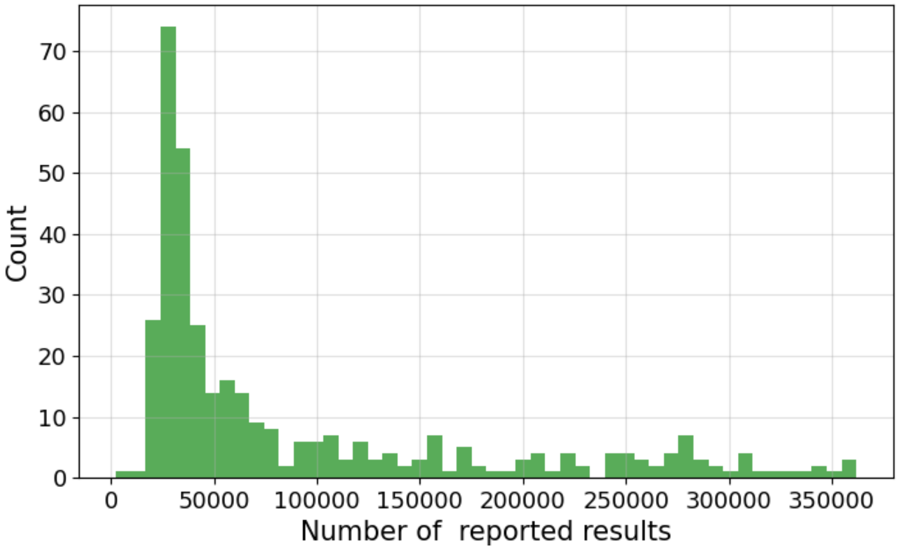


Figure 2: Data Distribution

Due to the lack of sufficient data and computing power, it is difficult to ensure that the model training reaches the desired intensity. Therefore, we intend to use transfer learning technology to transfer the model trained by authoritative enterprises in related tasks to our local tasks. Then we make certain fine-tuning on our own tasks so that the model can better adapt to the new task. In this process, in order to avoid the influence of different scales on the model in different tasks, and make the loss function smoother, and make the model convergence easier, we z-score standardize as

$$E(X) = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \sigma_X = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - E(X)\right)^2}{n}} \tag{1}$$

The histogram in the figure 3 below shows the normality test of the standardized number of reported results. If the normal graph shows a bell-shape (higher in the middle and lower on the sides), it can be accepted as a normal distribution, even though it is not strictly normal.

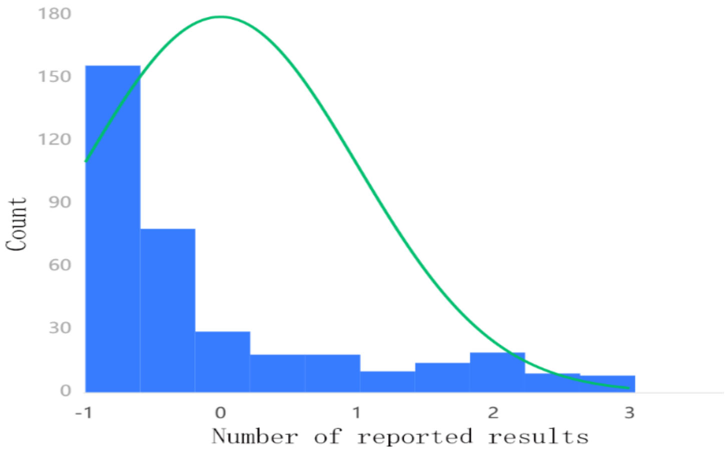


Figure 3: The Distribution of Standardized Data

Due to the requirement of training many neural network layers and large amount of data in deep learning, with a high cost involved. Thus, first of all we have to look for pre-trained models that we want and check on what kind of dataset it is trained on and whether it is applicable for our own task. Table 2 shows some sources from which we can obtain the pre-trained models that we desire.

Table 2: Pre-trained Model Sources

Model Providers	Model Websites Data
TensorflowHub	https://tfhub.dev/
ModelZoom	https://www.modelzoo.co/
ModelHub	https://github.com/modelhub-ai/
TIMM	https://github.com/rwightman/pytorch-image-models/

5 Model I: Reported Results Predictive Model

5.1 Description of Question

Fit a model based on the feature column Number of reported results data provided in the data file in order to reflect the most general rule of the data and use the predicted model to predict the possible result range of Number of reported results on a certain future day.

.

5.2 Random Forest Model

Random Forest uses independently trained multiple decision tree algorithms as its base learners and combines them by bagging, as shown in Figure 4. Random sampling is applied to both the sample observations and feature variables of the modeling dataset, resulting in a tree each time, with its own rules and classification results. The forest finally integrates all decision tree rules and classification results to achieve the classification of Random Forest.

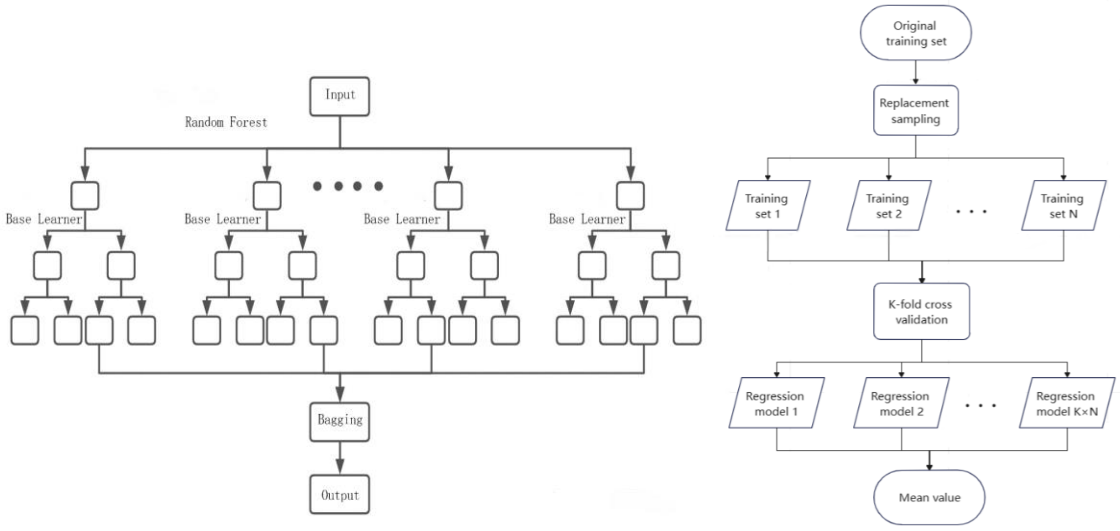


Figure 4: The Random Forest Structure and Process Flow Chart

5.2.1 K-Fold Cross Validation

The method for generating validation datasets: randomly divide the existing datasets into two categories, one is the training dataset and the other is the validation dataset. Train a model on the training dataset, calculate the error on the validation dataset, and use the validation error to approximate the generalization error.

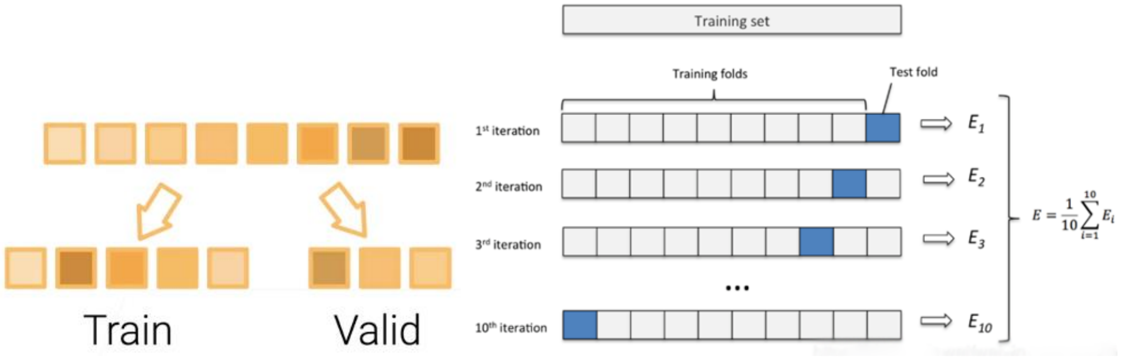


Figure 5: The Flowchart of K-Means Algorithm

Typically 10%~50% of the total data is chosen for validation samples, depending on whether the data is sufficient or not; when the data is sufficient, 50% can be chosen. When the data is very limited and cannot form a whole set for validation, this method can be used to divide the data set into k parts. For the i-th model, we can use the i-th part as the validation data, and the rest as the training data set. Then, the average of the k validation errors (i.e., the average errors of the k models) can be taken as my validation error.

5.2.2 Loss Function and Hyperparameters

Every algorithm model has its own loss function, and the gradient of each parameter of the loss function is calculated after each iteration of training. Then the optimizer updates each parameter in the negative gradient direction based on the current learning rate. Random forest is obtained by taking the sum of the results of multiple decision trees and then averaging them. For each decision tree, let S be the set of s data samples. Assume that the class label attribute has m different values, and define m different classes Ci (i = 1,...,m). Let si be the number of samples in class Ci. The expected information required for a given sample classification is given by the following formula, More information can be found in [1].

$$I(s_1, s_2, \cdots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad (2)$$

Where pi=si/s is the probability of any sample belonging to C_i. Note that the logarithm has a base of 2, which originates from information being encoded in binary. With other hyperparameters held constant, change an important one in the model to observe how the result changes. The hyperparameters of Random Forest are as shown in Table 4.

Table 3: Parameter estimation results

Parameter	Value1	Value2	Value3	Value4
Decision Tree Quantity	2	10	100	1000

Data Splitting	1	1	1	1
Data Shuffling	True	True	True	True
Cross Validation	7	7	7	7
Splitting Evaluation Criteria	MSE	MSE	MSE	MSE
Maximum Tree Depth	10	10	10	10
Maximum Number of Leaf	10	10	10	10
With Replacement Sampling	Ture	Ture	Ture	Ture

5.3 Results

The result of fitting the data with a Random Forest is shown on the left of the figure below. With other parameters fixed, the number of decision trees was changed and the model was observed to converge after certain point. The forecast of the reported results for March 1, 2023 is shown on Figure 6 below after fitting and predicting the situation.

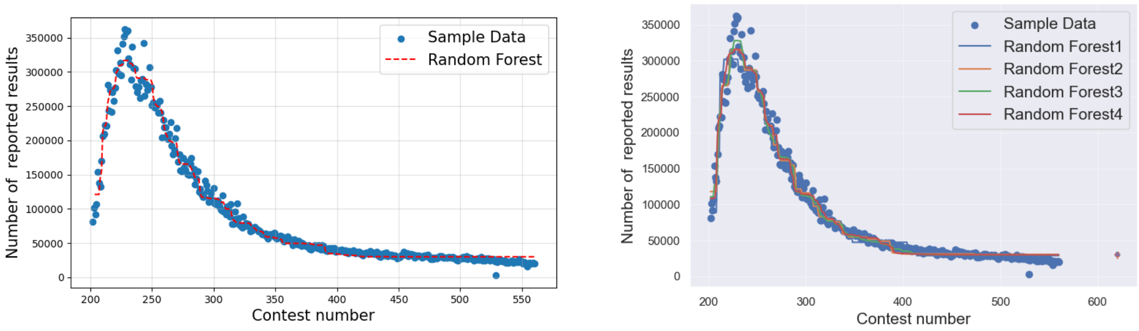


Figure 6: Comparison of Performance on Different Hyperparameters

On March 1, 2023, the corresponding Contest number was 620, and the Number of reported results value was different under different hyperparameters. When the number of decision trees is 2, the value of 3.1 is 28886; when the number of decision trees is 10, the value of 3.1 is 30272; when the number of decision trees is 100, the value of 3.1 is 29823; when the number of decision trees is 1000, the value of 3.1 is 29935. Therefore, it is inferred that the prediction estimation interval for 3.1 is [28000, 31000].

5.4 Model Evaluation

Using the formula to analyze the variance issue of random forest after bagging, assuming the variance of the variables in the subsets is σ^2 , and the correlation between any two variables is ρ, then the variance of the model is as shown below, for more details, see [2].

$$\begin{gathered}\operatorname{Var}\left(\frac{\sum X_t}{K}\right)=\frac{1}{K^2}\operatorname{Var}\left(\sum X_t\right)\\ \Rightarrow\frac{1}{K^2}\left(K\operatorname{Var}\left(X_1\right)+2\sum_{t=1}^{K}\ \sum_{j=1}^{K}\ \operatorname{cov}\left(X_t,X_j\right)_{t\neq j}\right.\end{gathered}\tag{3}$$

$$\rho = \frac{\mathrm{cov}(X_t, X_j)}{\sqrt{X_i}\sqrt{X_j}} \quad \Rightarrow \frac{1}{K^2}(K \cdot \mathrm{Var}\ (X_1) + 2\sum_{t=1}^{K}\sum_{j=1}^{K} \mathrm{cov}\ (X_i, X_j)_{i=j})$$

$$\Rightarrow \frac{1}{K^2}(K \cdot \mathrm{Var}\ (X_1) + 2\sum_{i=1}^{K}\sum_{j=1}^{K} \rho \cdot \sigma^2)$$

$$\Rightarrow \mathrm{Var}\ (\frac{\sum X_t}{K}) = \rho \cdot \sigma^2 + (1-\rho)\sigma^2/K \tag{4}$$

It can be obtained from Equation (4) that the variance of bagging is reduced. Therefore, after bagging, the bias of the model is close to that of the submodel, but the variance is greatly reduced, which reduces the complexity of the model and reduces the overfitting phenomenon. Here is the model evaluation result table 4:

Table 4: Parameter estimation results

	MSE	RMSE	MAE	MAPE
Training Set	0	0.013	0.008	3.773
Test Set	0.001	0.034	0.02	7.984

- MSE stands for the expectation of the difference between the predicted value and the actual value. The smaller the value, the higher the accuracy of the model.
- RMSE is the square root of MSE. The smaller the value, the higher the accuracy of the model.
- MAE is the average of absolute errors, which can reflect the actual situation of the prediction error. The smaller the value, the higher the accuracy of the model.
- MAPE is a variant of MAE and is a percentage. The smaller the value, the higher the accuracy of the model.

6 Model II: Difficult Mode Correlation Analysis

6.1 Question Analysis

Because the words are categorical variables, and the percentage of the number of people in the difficult selection mode is a quantitative type variable, in order to explore the relationship between some attributes of words and the percentage of the number of people in the difficult selection mode, we first need to convert these two types of data into the same type. After we encode each word with one-hot, then combine all the one-hot words to get the one-hot encoding of the word, but this operation will lose the temporal information between each letter of the words, so we use RNN model to retain the temporal information of the previous stage to the next stage. Finally, the word data with temporal information and the percentage of people in the difficult selection mode are analyzed for correlation.

6.2 Recurrent Neural Network

Recurrent Neural Networks (RNN) are an artificial neural network models used for processing sequence data. It processes specific data inputs in a recursive manner, producing a series of output results. The strength of RNNs lies in their ability to handle variable-length sequences, capture relationships between sentences within a sentence and identify the implicit structure of language, as well as discover relationships between words. The algorithm flow is illustrated in the following figure:

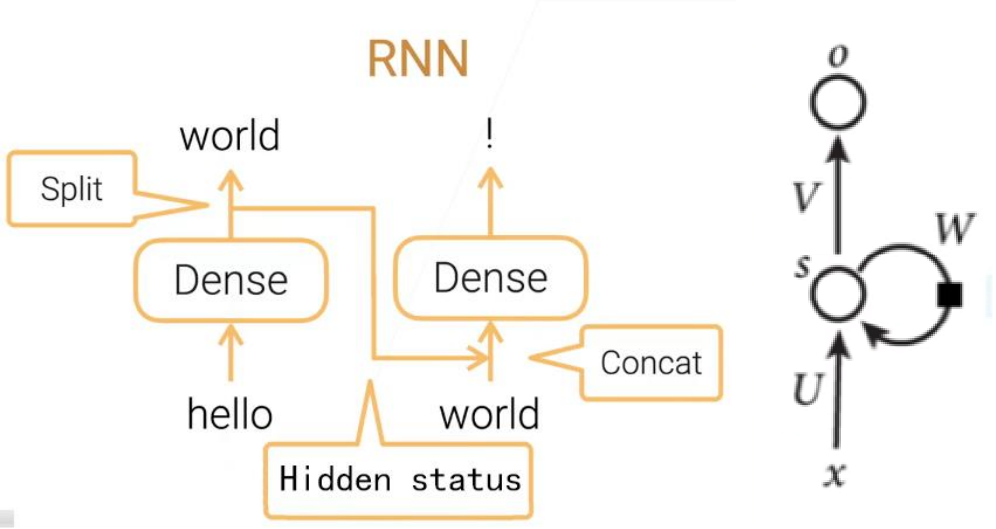


Figure 7: The process of RNN algorithm

The vector x_t is 57-dimensional, U has 128 neurons and the input is 57 values; x_t goes through U outputted in a 128-dimensional vector, Ux_t, s_{t-1} is the state at time t-1 goes through a W layer to output a 128-dimensional vector, Ws_{t-1}.

$$s_t = f(Ux_t + Ws_{t-1}) \qquad (5)$$

At time t+1, st also provides the state at time t, V has 57 neurons, the input is a 128-dimensional vector, the output is a 57-dimensional vector, for Vs_t.

$$o_t = \text{softmax}\ (Vs_t) \qquad (6)$$

In this paper, a recurrent neural network (RNN) was used to perform one-hot encoding of words in a way that preserves the meaning of the words, in order to investigate the correlation between the intrinsic properties of the words and their hard-mode percentage scores. o_t is the predicted value for current moment.

We obtained feature states saving the time-series information of each word through RNN, and it is easy to judge the correlation between two data objects by drawing a scatter plot. Different correlation data objects present the scatter distribution as shown in the following graph, more details can be seen in [3].

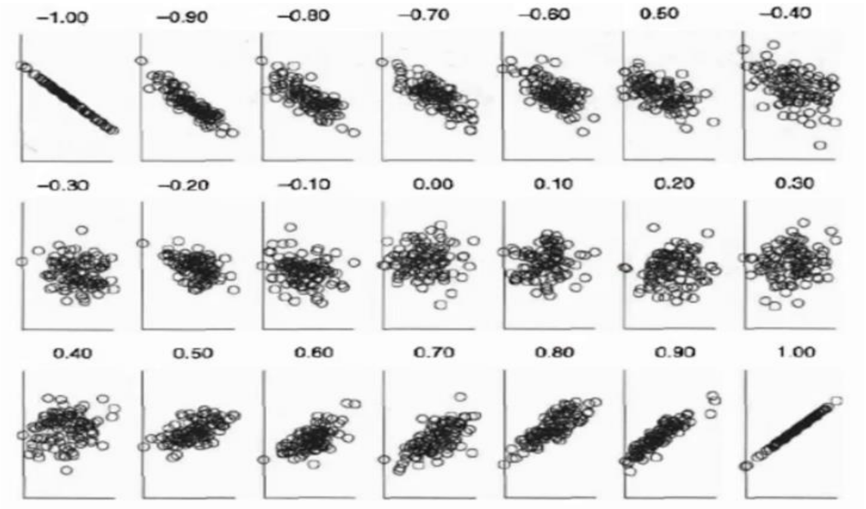


Figure 8: Explain scatter plot of correlation from -1 to 1

Thus, we used scatter plot to observe the distribution form after standardizing the data of the proportion of people who took the selection difficulty mode and the feature states that contain each word's temporal information, and found that the correlation was close to 0, so it can be judged that the proportion of people taking the selection difficulty mode is completely unaffected by the word factor.

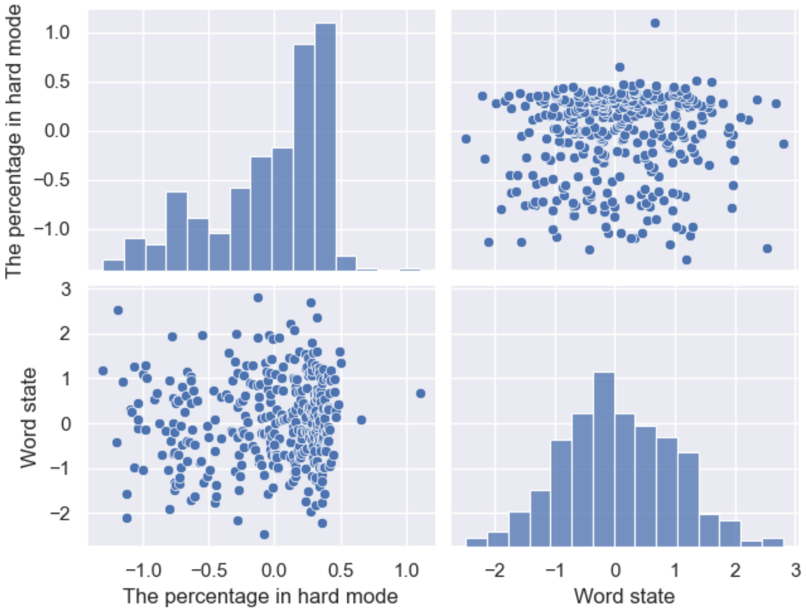


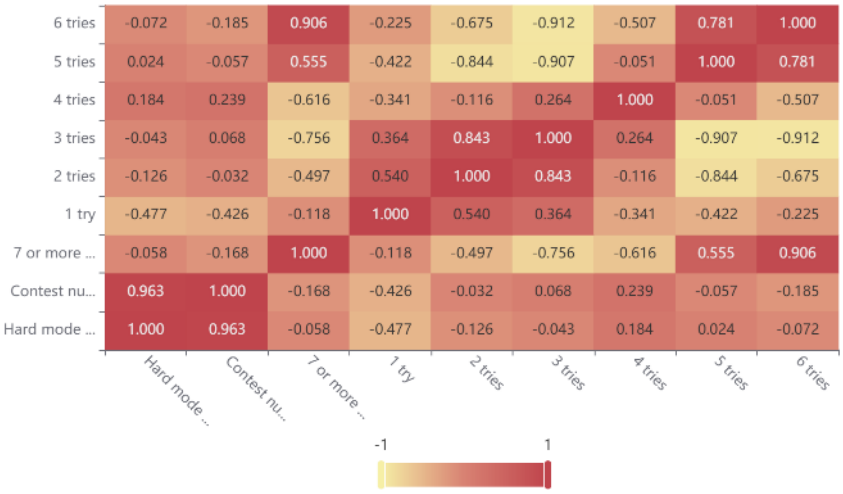
Figure 9: Scatter Plot of Feature Data

6.3 Spearman's Correlation Coefficient

It has been proved that the Date and Contest number are irrelevant to the meaning of the words and difficulty mode ratio, so we can exclude them. We will then explore whether the difficulty mode ratio is related to other factors of the word's properties.The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the ranks of two level variables. For a sample with size n, the n original data are transformed into level data, and the correlation coefficient ρ is

$$\rho = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}} \tag{7}$$

Calculate the Spearman correlation coefficient respectively by putting the Hard mode percentage scores together with other features in the data, and from the correlation coefficient heat map, it is obvious to see that Hard mode percentage scores have a strong correlation with Contest number.



	Hard mode ...	Contest nu...	7 or more ...	1 try	2 tries	3 tries	4 tries	5 tries	6 tries
6 tries	-0.072	-0.185	0.906	-0.225	-0.675	-0.912	-0.507	0.781	1.000
5 tries	0.024	-0.057	0.555	-0.422	-0.844	-0.907	-0.051	1.000	0.781
4 tries	0.184	0.239	-0.616	-0.341	-0.116	0.264	1.000	-0.051	-0.507
3 tries	-0.043	0.068	-0.756	0.364	0.843	1.000	0.264	-0.907	-0.912
2 tries	-0.126	-0.032	-0.497	0.540	1.000	0.843	-0.116	-0.844	-0.675
1 try	-0.477	-0.426	-0.118	1.000	0.540	0.364	-0.341	-0.422	-0.225
7 or more ...	-0.058	-0.168	1.000	-0.118	-0.497	-0.756	-0.616	0.555	0.906
Contest nu...	0.963	1.000	-0.168	-0.426	-0.032	0.068	0.239	-0.057	-0.185
Hard mode ...	1.000	0.963	-0.058	-0.477	-0.126	-0.043	0.184	0.024	-0.072

Figure 10: Heatmap of Correlation Coefficients

6.4 Linear Regression

Regression analysis that includes only one independent variable and one dependent variable, and the relationship between the two can be approximately represented by a straight line, is called univariate linear regression analysis. If regression analysis includes two or more independent variables, and the relationship between the dependent variable and the independent variables is linear, it is called multiple linear regression analysis. Generally speaking, linear regression can be solved by the least squares method or the gradient descent method to calculate its equation and calculate the line for y=bx+a.

We conducted a linear regression with a one-variable independent variable, Contest number, and a dependent variable, Hard mode percentage scores, to establish a model with the following results (see the figure 11 below).

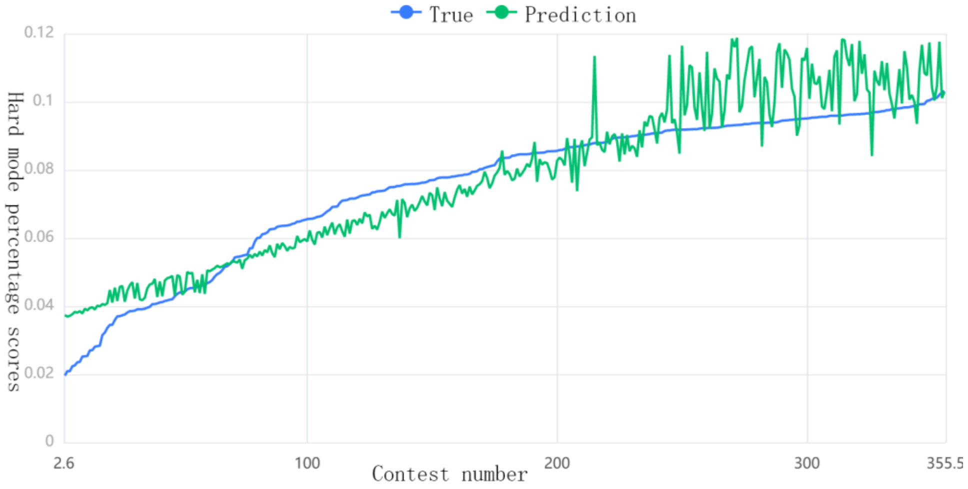


Figure 11: Linear Regression with One Variable

Finally, we can use the model to calculate that the Hard mode percentage scores for March 1st 2023 will be 0.132.

7 Model III: Probability Distribution Predictive Model

7.1 Data Analysis

The word is a guess made by players on related dates and competition numbers, so the word is related to related dates and competition numbers. And the dates and competition numbers are increasing series, they also have a strong correlation. For three feature data, we only select the feature data of competition numbers and scale it using min-max standardization. From 6.4, we know that there is a strong correlation between Number of reported results and Number in hard mode, so we take Hard mode percentage scores as the percentage of Number in hard mode in Number of reported results.

Here we use 7 SVR algorithms to predict the result. The training samples x for the first model are: competition numbers, Number of reported results, and the training samples y hat for the first model are: 1 try; the training samples x for the second model are: competition numbers, Number of reported results, 1 try, and the training samples y hat for the second model are: 2 tries; the training samples x for the third model are: competition numbers, Number of reported results, 1 try, 2 tries, and the training samples y hat for the third model are: 3 tries; the training samples x for the fourth model are: competition numbers, Number of reported results, 1 try, 2 tries, 3 tries, and the training samples y hat for the fourth model are: 4 tries; the training samples x for the fifth model are: competition numbers, Number of reported results, 1 try, 2 tries, 3 tries, 4 tries, and the training samples y hat for the fifth model are: 5 tries; the training samples x for the sixth model are: competition numbers, Number of reported results, 1 try, 2 tries, 3 tries, 4 tries, 5 tries and the training samples y hat for the sixth model are: 6 tries; and the training samples x for the seventh model are: competition numbers, Number of reported results, 1 try, 2 tries, 3 tries, 4 tries, 5 tries, 6 tries and the training samples y hat for the seventh model are: 7 or more tries (X).

The result predicted by the first model is used for the testing data of the second model, and the result of the second model and the prediction result of the first model are used for the testing data of the third model, and so on. There are 7 models in total, and in the test data of the last model, there are 6 results from the first 6 models (1 try, 2 tries, 3 tries, 4 tries, 5 tries, 6 tries), and the remaining two are from the data set analysis results (Hard Mode percentage scores and min-max standardized competition numbers).

7.2 Support Vector Regression

For a general regression problem, given the training sample D = {(x1, y1), (x2, y2), ..., (xn, yn)}, we hope to learn a f(x) that is as close as possible to y, and w, b are parameters to be determined. In this model, only when f(x) is exactly the same as y, the loss is zero. The support vector regression assumes that we can tolerate at most ε difference between f(x) and y. Only when the absolute difference between f(x) and y is greater than ε, the loss is calculated. At this time, it is equivalent to constructing an interval belt with a width of 2ε centered on f(x). If the training sample falls into this interval belt, it is considered to be predicted correctly (the relaxation degree on both sides of the interval belt can be different).

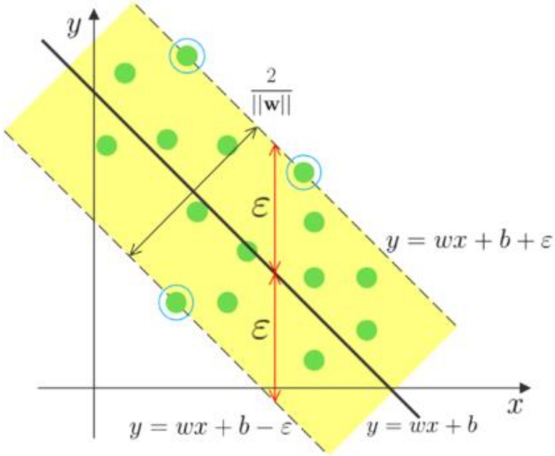


Figure 12: Interval Diagramse

Thus, the SVR problem can be transformed into (the left part of the below equation is the regularization term) [4].

$$\min_{\boldsymbol{w},b}\frac{1}{2}\parallel \boldsymbol{w}\parallel^2+C\sum_{i=1}^{m}\ \ell_c(f(\boldsymbol{x}_i)-y_i) \qquad (8)$$

The left part of the above equation is the regularization term and the right part is the loss function.

$$\ell_\epsilon(z)=\{\begin{matrix}0, & \text{if}|z|\leqslant\epsilon\\ |z|-\epsilon, & \text{otherwise}\end{matrix} \qquad (9)$$

So a relaxation factor was introduced, and the SVR problem became

$$\begin{gathered}\min_{\boldsymbol{w},b,\xi_i,\hat{\xi}_i}\frac{1}{2}\parallel w\parallel^2+C\sum_{i=1}^{m}(\xi_i+\hat{\xi}_i)\\ \text{s.t.} f(x_i)-y_i\leqslant\epsilon+\xi_i\\ y_i-f(x_i)\leqslant\epsilon+\hat{\xi}_i\\ \xi_i\geqslant 0,\hat{\xi}_i\geqslant 0, i=1,2,\dots,m\end{gathered} \qquad (10)$$

The Lagrange function can be obtained by introducing the Lagrange multiplier finally.

$$\begin{gathered}L(\boldsymbol{w},b,\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}},\boldsymbol{\xi},\hat{\boldsymbol{\xi}},\boldsymbol{\mu},\hat{\boldsymbol{\mu}})\\ =\frac{1}{2}\parallel \boldsymbol{w}\parallel^2+C\sum_{i=1}^{m}(\xi_i+\hat{\xi}_i)-\sum_{i=1}^{m}\ \mu_i\xi_i-\sum_{i=1}^{m}\ \hat{\mu}_i\hat{\xi}_i \qquad (11)\\ +\sum_{i=1}^{m}\ \alpha_i(f(\boldsymbol{x}_i)-y_i-\epsilon-\xi_i)+\sum_{i=1}^{m}\ \hat{\alpha}_i(y_i-f(\boldsymbol{x}_i)-\epsilon-\hat{\xi}_i)\end{gathered}$$

Differentiate the four traversals, set the partial derivatives to zero, and obtain.

$$\begin{aligned} \boldsymbol{w} &= \sum_{i=1}^{m} (\hat{\alpha}_i - \alpha_i)\, \boldsymbol{x}_i \\ 0 &= \sum_{i=1}^{m} (\hat{\alpha}_i - \alpha_i) \\ C &= \alpha_i + \mu_i \\ C &= \hat{\alpha}_i + \hat{\mu}_i \end{aligned} \tag{12}$$

Inserting the above equation into it, the dual problem of SVR can be obtained.

$$\begin{gathered} \max_{\alpha,\hat{\alpha}} \sum_{i=1}^{m} y_i(\hat{\alpha}_i - \alpha_i) - \epsilon(\hat{\alpha}_i + \alpha_i) \\ -\frac{1}{2}\sum_{i=1}^{m} \sum_{j=1}^{m} (\hat{\alpha}_i - \alpha_i)\,(\hat{\alpha}_j - \alpha_j)\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j \\ \text{s.t.} \sum_{i=1}^{m} (\hat{\alpha}_i - \alpha_i) = 0 \\ 0 \leqslant \alpha_i, \hat{\alpha}_i \leqslant C \end{gathered} \tag{13}$$

The above process needs to satisfy the KKT conditions.

$$\begin{cases} \alpha_i(f(\boldsymbol{x}_i) - y_i - \epsilon - \xi_i) = 0 \\ \hat{\alpha}_i(y_i - f(\boldsymbol{x}_i) - \epsilon - \hat{\xi}_i) = 0 \\ \alpha_i\hat{\alpha}_i = 0, \xi_i\hat{\xi}_i = 0 \\ (C - \alpha_i)\xi_i = 0, (C - \hat{\alpha}_i)\hat{\xi}_i = 0 \end{cases} \tag{14}$$

Finally, the solution of SVR can be obtained.

$$\begin{gathered} f(\boldsymbol{x}) = \sum_{i=1}^{m} (\hat{\alpha}_i - \alpha_i)\, \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x} + b \\ b = y_i + \epsilon - \sum_{i=1}^{m} (\hat{\alpha}_i - \alpha_i)\, \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x} \end{gathered} \tag{15}$$

The hyperparameters of the seven models are shown in the table below. For Model 1, it fits the 2-dimensional feature data and can be displayed on a 3-dimensional plane, while for the others, due to the high dimensional feature data, it is unable to show the full view of the models. The model 1 fitting and modeling graph is shown as follows:

Table 5: Model Hyperparameters

Parameter Name	Parameter value
Training Time	0.024S
Data Splitting	0.7
Data Shuffling	Ture
Cross Validation	7
Penalty Coefficient	1
Kernel Function	Linear
Kernel Coefficient	Scale
Kernel Constant	0
Highest Power	3
Error Convergence	0.01
Maximum Iterations	1000

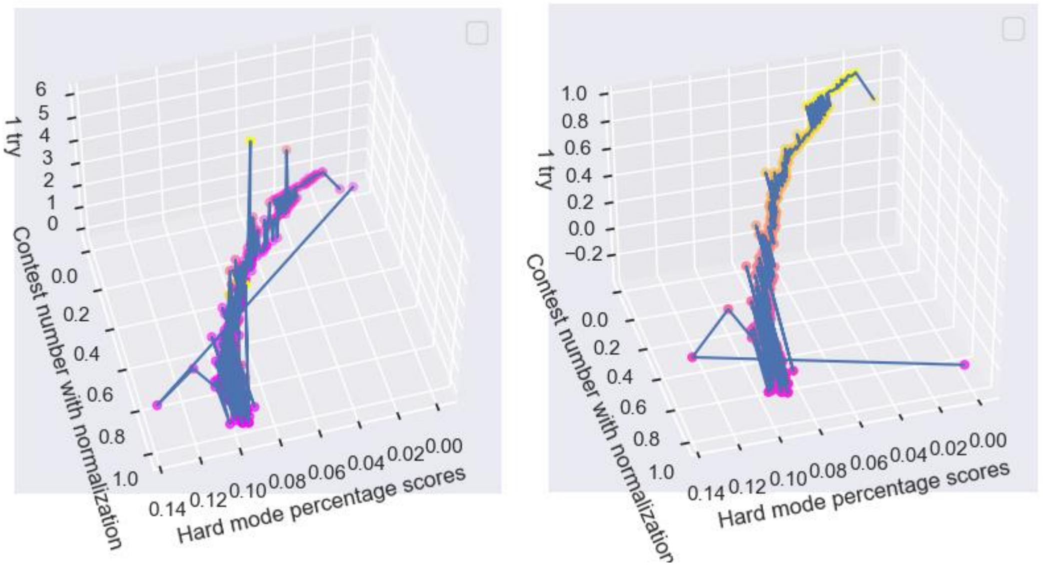


Figure 13: Real Model (Left) vs. Fitted Model (Right)

For March 1st, 2023, the word EERIE corresponds to Contest number 620, and the Hard mode percentage score for March 1st, 2023 is 0.132. The first model therefore yields a result of -0.22 (approximately 0). All the previous predictions are then used as the test data in the next model, and the results of the second model are shown in the table below.

Table 6: Parameter estimation results

Date	Contest number	Hard mode percentage scores	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	6+ tries	reported results
2023-3-1	620	0.132	-0.22(0)	4.44	22.11	34.89	24.51	10	3.02	[28000，31000]

7.3 Model Evaluation

7.3.1 Uncertainty Factor

The prediction results of the nth model in these 7 models must be based on the prediction results of the n-1 model (n>1), and the prediction results of the 1st model must be based on the one-dimensional linear regression model in Chapter 6.5. These models are linked together, which may cause a large error in one place, resulting in a huge change, and the final result may be very different from the real value.

7.3.2 Evaluation Results

The following are the evaluation results of the seven models implemented in this chapter:

Table 7: Result 1

Model1	MSE	RMSE	MAE
Test	0.454	0.674	0.384

Table 8: Result 2

Model2	MSE	RMSE	MAE
Test	8.96	2.993	2.232

Table 9: Result 3

Model3	MSE	RMSE	MAE
Test	22.46	4.739	3.865

Table 10: Result 4

Model4	MSE	RMSE	MAE
Test	12.171	3.489	2.678

Table 11: Result 5

Model5	MSE	RMSE	MAE
Test	3.819	1.954	1.244

Table 12: Result 6

Model6	MSE	RMSE	MAE
Test	7.021	2.65	1.043

Table 13: Result 7

Model7	MSE	RMSE	MAE
Test	3.872	1.968	1.102

Through the evaluation results of the model, I feel that the error degree of the model is not very large. But the data in the dataset is too small, which may make it difficult for the model to learn some general information in the data. I still have some confidence in my model.

8 Model Ⅳ: Vocabulary Difficulty Classification Model

8.1 Question Analysis

The difficulty of a word does not have any relation to the word itself, but to the Contest number, Date and Word. The difficulty of a word is reflected in the result displayed by the people when dealing with the word in the same activity. Since 6.2 states the correlation between Number of reported results and Number in hard mode, the Hard Mode percentage scores are introduced as the percentage of Number in hard mode in Number of reported results to represent the former two. Therefore, we explore the relationship between the difficulty of a word and the Hard Mode percentage scores, 1 try, 2 try, 3 try, 4 try, 5 try, 6 try, 7 or more tries (X).

8.2 K-Means Clustering Algorithm

K-Means algorithm is an unsupervised learning algorithm, which is also a partition-based clustering algorithm. Generally, the Euclidean distance is used as the index to measure the similarity between data objects, and the similarity is inversely proportional to the distance between data objects. The larger the similarity is, the shorter the distance is. The algorithm needs to specify the initial clustering number k and the k initial clustering centers in advance. According to the similarity between data objects and clustering centers, the position of clustering centers is constantly updated and the sum of squared errors (SSE) of clusters is constantly lowered. When SSE does not change or target function converges, the clustering ends and the final result is obtained.

Its core idea is: first, randomly select k initial clustering centers C_i (1⩽i⩽k) from the dataset and calculate the Euclidean distance between the other data objects and the clustering center C_i. Find the clustering center C_i closest to the target data object and assign the data object to the cluster corresponding to the clustering center C_i. Then calculate the average value of the data objects in each cluster as the new clustering center and iterate until the clustering center does not change or the maximum iteration times are reached. The Euclidean distance calculation formula between data objects and clustering centers in space is as follows: see [5].

$$d(x, C_i) = \sqrt{\sum_{j=1}^{m} (x_j - C_{ij})^2} \tag{16}$$

Where x is the data object, C_i is the i-th clustering center, m is the dimension of the data object, x_j, C_{ij} is the j-th attribute value of x and C_i. The formula of the sum of squares of errors SSE of the entire dataset is

$$SSE = \sum_{i=1}^{k} \sum_{x\in C_i} |d(x, C_i)|^2 \tag{17}$$

When the number of clustering k takes different values, the results are shown:

Table 14: K=2 Results

Classification	Count	Percentage
Kind_1	185	51.532
Kind _2	174	48.468
Sum	359	100

Table 15: K=3 Results

Classification	Count	Percentage
Kind _1	185	51.532
Kind _2	97	27.019
Kind _3	77	21.448
Sum	359	100

Table 16: K=5 Results

Classification	Count	Percentage
Kind _1	107	29.805
Kind _2	25	6.964
Kind _3	39	10.864
Kind _4	64	17.827
Kind _5	124	34.54
Sum	359	100

Table 17: K=4 Results

Classification	Count	Percentage
Kind _1	132	36.769
Kind _2	128	35.655
Kind _3	62	17.27
Kind _4	37	10.306
Sum	359	100

When K > 3, the distribution of the data for each category is extremely uneven, while when K = 2 or K = 3, the distribution of the data for each category is good. Here, we select K = 3 as the level to differentiate the difficulty, which are easy, medium, and difficult.

8.3 Support Vector Machine

SVM seeks to find the farthest distance between different sample points and the hyperplane, which is to find the largest margin hyperplane. Any hyperplane can be described by the following linear equation: see [6].

$$w^T x + b = 0 \tag{18}$$

The following is the formulation of the optimization problem.

$$\begin{gathered} \min_{\omega,b}(\frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}} \cdot \boldsymbol{\omega} + C\sum_{i=1}^{N} \xi_i) \\ \text{s.t.} y_i(\boldsymbol{\omega}^{\mathrm{T}} \cdot x_i + b) \geqslant 1 - \xi_i \\ \xi_i \geqslant 0, i = 1,2 \cdots, N \end{gathered} \tag{19}$$

In the equation, ω is the normal vector, b is the constant term, C is the penalty factor, and ξ_i is the slack variable. To find the optimal normal vector ω and constant term b, the optimal classification surface can be obtained. In order to transform the equation into a quadratic programming problem, the corresponding Lagrange function is introduced, so that the classification problem is transformed into

$$\begin{gathered} L(\boldsymbol{\omega}, b, \lambda) = \frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}} \cdot \boldsymbol{\omega} + C\sum_{i=1}^{N} \xi \cdot \\ \sum_{i=1}^{N} \alpha_i[y_i(\omega^{\mathrm{T}} x_i + b) - 1 + \xi_i] - \sum_{i=1}^{1} \beta_i \xi_i \end{gathered} \tag{20}$$

In the equation, α_i, β_i are Lagrange multipliers, α_i,⩾0, β_i⩾0.According to the dual principle, the above equation can be transformed into

$$\begin{gathered} \max_{\alpha} L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{N} \alpha_i \alpha_j y_i y_j (\Phi_{xi} \times \Phi_{xj}) \\ \text{s.t.} \sum_{i=1}^{N} y_i \alpha_i = 0, 0 \leqslant \alpha_i \leqslant C, \end{gathered} \tag{21}$$

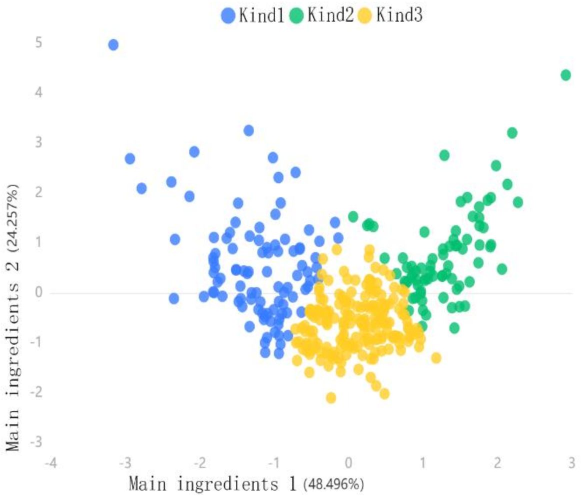
The kernel function $\Phi_{xi} \times \Phi_{xj} = K(x_i, y_j)$ is used for non-linearly separable problems, and support vector machines need to introduce kernel functions to project data from low-dimensional space to high-dimensional space, in order to achieve the purpose of linear separability. Commonly used kernel functions include linear kernel function, polynomial kernel function, radial basis kernel function and perceptron kernel function, and different kernel functions will lead to different forms of support vector machines.

According to the difficulty level divided in section 8.2, all training samples are labeled as simple, medium, and difficult. Here, the data label classification of simple difficulty level is defined as 1, the data label classification of medium difficulty level is defined as 2, and the data label classification of difficult difficulty level is defined as 3. These samples are used as the training data of the SVM model, and the classification label of the difficulty level is used as the category of the training data to train the model. Here, the data category distribution chart and the configuration of the parameters of the SVM model are as follows:

Table 18: Parameters of SVM

Parameters	Count
Training Time	0.054s
Data Splitting	0.7
Data Shuffling	True
Cross Validation	7
Kernel Function	Linear
Coefficients	Scale
Error	0.001
Iterations	1000

Figure 14: Clustering Distribution



The results obtained by introducing the results of Chapter 7.2 into the SVM model are shown in the following table.

Table 19: Prediction Result

Category	Kind_1.0	Kind _2.0	Kind _3.0
1.0	0.995	0.003	0.001

8.4 Accuracy Analysis

The following table is the evaluation result of the model：

Table 20: Model Accuracy

	Accuracy	Recall	Precision	F1
Training Set	0.996	0.996	0.996	0.996
Cross Validation Set	0.984	0.984	0.986	0.984
Test Set	0.963	0.963	0.964	0.963

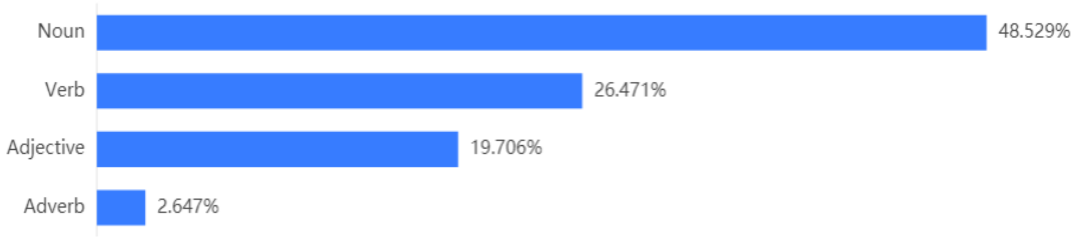
The analysis showed that the classification evaluation indicators of the training set, cross-validation set, and test set were all greater than 0.9, indicating that the model had good classification effect on the test set, and the model had practicality.

9 Interesting Features

On November 30th, 2022, there were only 2569 submissions of answers to the Wordle game, with 2405 of them coming from the Pro Mode, accounting for 93% of the total submissions. This figure is obviously abnormal. If the data reflects the real situation, that day should have been Thanksgiving, a national holiday when people gather together and have plenty of leisure time. Normally, the number of players should have increased instead of decreased on that day. So it is likely that on November 30th, 2022, due to server updates the game was still in the testing phase and not yet officially released, so there were not enough players to participate in the game.

Analysis of the part of speech for all words in the table data reveals that nouns dominate, verbs and adjectives occupy the middle ground, and other parts of speech such as adverbs are almost nonexistent. This can be indicative that the words used in the Wordle game originated mostly from news articles and reports because they typically rely on objective descriptions and utilize a large number of nouns compared to, say, prose or fiction that are traditionally characterized by a large amount of verbs and adjectives. The word statistical chart is shown below:

Figure 15: Word Part-of-Speech Statistics



10 A Letter to Puzzle Editor

Dear Sir/Madam:

I am writing to introduce a unique technology-aided approach to helping you develop puzzles for The New York Times. My research has explored four different machine learning models that can be used to better understand the data behind the puzzles, as well as to predict potential outcomes.

Model I uses the K-fold Cross-validation to split the data and then employs the Random Forest algorithm to fit the relationship between the Contest number and the features of the data. Through Bagging, the variance of the model is reduced, and the model can predict that the number of reports on March 1, 2023 will be between 28000 and 31000.

Model II focuses on the relationship between the Hard Mode Percentage Scores and the words used in the puzzles. After using the RNN model to turn the words into meaningful and chronological information, the Spearman correlation is used to analyze the correlation between the Hard Mode Percentage Scores and Contest number. The result is a univariate linear regression to predict the Hard Mode Percentage Scores on March 1, 2023 to be 0.132.

Model III is used to reduce the data and analyze the relationship between the frequency distribution and the two features Hard Mode Percentage Scores and Contest number. Seven SVR models are used to predict each individual number, and each model's test data comes from the prediction results of the previous model. The prediction results show that 1 try will be 0%, 2 try will be 4.44%, 3 try will be 22.11%,4 try will be 34.89%, 5 try will be 24.51%, 6 try will be 10%, and 7 or more tries (X) will be 3.02%.

Model Ⅳ uses the K-Means algorithm to clustering the data and define the word difficulty level as easy, medium, and hard. An SVM model is then trained on the three classes of data and is used to predict that the difficulty of EERIE is 99% easy. The evaluation results for the SVM model show that all the indicators are greater than 0.9, indicating that the model has good classification effect.

Finally, based on the original discrete points of the data, some additional information can be derived. For example, on November 30, 2022, a possible server update may occur. In addition, the statistics of the part of speech of all the words in the data indicates that most of the words in Wordle game originate from news articles or reports, as it is usually descriptive.

I hope this letter gives you some new ideas and inspiration in your process of creating puzzles for The New York Times.

Sincerely,
Team #2314130

References

[1] Yang, X.B., & Zhang, J. (2007). Decision Tree Algorithm and Its Core Technology. Computer Technology and Development, 01, 43-45.

[2] Zhou, Z.-H. (2016). Machine Learning. Beijing: Tsinghua University Press.

[3] Pang-Ning Tan. Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann, 2011.

[4] Vapnik, V.．Statistical learning theory. 1998 (Vol. 3).．New York, NY：Wiley，1998：Chapter 10-11, pp.401-492.

[5] Saroj，Kavita.Review：study on simple k mean and modified K mean clustering technique[J].International Journal of Computer Science Engineering and Technology，2016，6（7）：279-281.

[6] Li, H. (2012). Statistical Learning Methods. Beijing: Tsinghua University Press, Chapter 7, pp. 95-135.