A Random Forest-Based Word Difficulty Prediction Model

Chenye Xi *, Gong Chen

Department of Southampton Ocean Engineering Joint Institute, Harbin Engineering University, Heilongjiang China, 150001

* Corresponding Author Email: 15103453333@163.com

Abstract. The aim of this paper is to analyse the relationship between the word of the day and the corresponding distribution of the number of attempts in the Wordle game and to give a prediction method for the proportional distribution of word attempts. Firstly, the paper preprocesses the data provided by Question C of the 2023 American Collegiate Mathematical Modelling Competition. By constructing a model, this paper quantifies the word information entropy and people's preference for choosing common letters. Considering the above features and the influence of previous attempts on the follow-up, this paper constructs a regression model to verify the correlation between the word composition features and the distribution of the number of attempts. Meanwhile, considering the subjectivity of feature selection, this paper constructs a random forest model for further analysis. Comparing the results of the model analysis, the random forest model fits better, and the proportion of word EERIE attempts from 1 to 7 is 0%, 1.97%, 15.99%, 36.31%, 29.83%, 13.34%, and 2.24% respectively . This paper provides a theoretical basis for predicting the number of attempts of the corresponding words, which helps Wordle to optimally adjust the lexicon.

Keywords: Wordle, Information entropy, Random Forest, Difficulty prediction.

1. Introduction

Wordle is a very popular online word-guessing game in which the player needs to select letters in turn, and the colour of the letter prompts the player after each attempt. The player needs to guess a 5-letter word in a maximum of 6 attempts. This simple and fun game has gained immense popularity in early 2022, and there are now multiple aspects of research that have been conducted on Wordle that are important for the development and promotion of the game. For example, Siddhant's team used a reinforcement learning-based approach to help people guess words correctly in the least number of times [1]; Yitai's team used a Prophet model to explain the reasons for the change in the number of reports and to predict the interval between the changes in the number of reports on a particular day in the future [2]; and the change in the number of players over time can be predicted using an LSTM time series as well [3]. However, considering the actual situation, these methods have certain problems in prediction, such as the use of time series prediction does not take into account the impact of the word itself on the number of results of the day's participation in the report; reinforcement learning methods have a high demand for training, and the sampling efficiency is too low, and so on. So far, no systematic research has been conducted on the relationship between the distribution of the number of attempts for word guessing and the word itself. Therefore, the aim of this paper is to analyse the relationship between the word and the corresponding attempt count distribution, and to provide a method for predicting the count distribution.

2. Materials and methods

2.1. Data acquisition and processing

In this paper, a total of 355 data contained in the appendix of Question C (https://www.comap.com/resources/free-materials) of the 2023 Mathematical Modeling Competition for American College Students are statistically analyzed. In order to reduce the influence of wrong data on the results, this paper first cleans the data. By checking the Twitter posts of @WordleStats, this paper corrects the wrong words or data in the given data. For the data that can't guess the correct

word within six times, this paper classifies it as seven times or more. For the data that the percentage of attempts is not 1, this paper divides each percentage by the data of the percentage sum to obtain the proportion.

2.2. Introduction to the methodology

For the purpose of the study, the following hypothesis is made during the guessing process, people will tend to choose more familiar words or letters, and all people in the given sample data will have the same preference for the letters of commonly used words. Also, the information given by previous wrong choices is taken into account in subsequent attempts to guess the same word.

For the analysis of the distribution of the number of attempts, this paper considers that the word composition and characteristics are important factors affecting the distribution of the number of attempts. Therefore, this paper chooses the information entropy and frequency of occurrence of each letter of a word as important indicators for prediction analysis. Considering people's preference for commonly used words and the hint of letter colour after each attempt, this paper believes that previous attempts have a greater impact on the subsequent guessing results. Therefore, this paper chooses the information entropy of word letters, the preference of commonly used letters and the proportion of previous attempts as the indicators for analysing the relationship between words and the distribution of corresponding attempts.

In order to study the relationship between the above indicators and the distribution of the number of attempts, this paper first uses linear regression to verify the strength of the relationship between word features and the distribution of the number of attempts. Considering that the subjectivity of choosing complex features of words may affect the prediction results, in order to reduce the influence of this factor on the prediction results, this paper constructs a random forest model to analyse the relationship between word features and the distribution of the number of attempts and cross-checks the results of the analysis to verify its reliability. By comparing the prediction results of the two models, the prediction results are obtained [4].

2.2.1 Information entropy

Information entropy draws on the concept of thermodynamics and refers to the average amount of information after eliminating redundancy in a message. Shannon's information theory research points out that for uncertain source symbols, can be measured according to the probability of its occurrence of its information, the higher the probability of occurrence of the event, the lower the amount of information it carries; vice versa it carries a high amount of information. In order to study the relationship between the distribution of the number of attempts and the word itself, this paper argues that the amount of information in the word will have a significant impact on how easy or difficult it is for people to guess the word, and therefore, information entropy is used to quantitatively represent the complexity of the word [5].

In a source, it is not the uncertainty of the occurrence of a particular individual symbol that is considered, but the average uncertainty of all possible occurrences of this source. If the source symbols have n values: U_1 ...U_i ...U_n, the corresponding probabilities are P_1 ...P_i ...P_n , and the occurrences of the various symbols are independent of each other. At this point, the average uncertainty of the source should be the statistical average (E) of the individual symbol uncertainties -logP_i, which can be called the information entropy, i.e.:

Where p (x_i) denotes the probability of occurrence of the source symbol.

$$H(x) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{1}$$

2.2.2 Preference quantification (word position entropy)

Considering that people's preference for the selection of commonly used letters when guessing words affects the selection results, this feature is quantified in this paper. At the same time, due to the influence of previous attempts on the follow-up in the process of guessing, the range of letter selection

will be narrowed with the increase in the number of attempts, which can be regarded as a gradual decrease in the influence of the commonly used letters, therefore, this paper assigns the preference indicators based on the sequence of the number of attempts, and uses the results after the assignment for the prediction analysis [5-6].

Entropy weight method is an objective assignment method, in the specific use of the process, according to the degree of dispersion of the data of each indicator, using the information entropy to calculate the entropy weight of each indicator, and then according to the indicators of the entropy weight of certain corrections, so as to obtain a more objective weight of the indicators. This method can to a certain extent reduce the impact of the subjectivity of hierarchical analysis on the assignment results.

First, the evaluation object is determined, the evaluation index system is established, the level matrix R is constructed and normalised. Calculate the entropy value of the evaluation indicators using the formula:

Among them:

$$H_j = -k\sum_{i=1}^{m} f_{ij} \ln f_{ij} \tag{2}$$

$$f_{ij} = \frac{r_{ij}}{\sum_{i=1}^{m} r_{ij}}, k = \frac{1}{\ln m} \tag{3}$$

The entropy weights W_i can be obtained by bringing in equation (2):

$$W_i = \frac{1-H_i}{m-\sum_{i=1}^{m} H_i} \tag{4}$$

2.2.3 Linear regression

Regression analysis is a predictive modelling technique that examines the relationship between dependent and independent variables. This technique is commonly used in predictive analyses, time series modelling and to discover causal relationships between variables, assuming that the relationship between the dependent variable y and the independent variable x satisfies a linear function $y = \omega x + b$. A curve/line is usually used to fit the data points with the goal of minimising the difference in distance from the curve to the data points [7], i. e:

$$L(\omega, b) = \frac{1}{n}\sum_{i=1}^{n} (\omega x_i + b - y_i)^2 \tag{5}$$

The relationship between the independent and dependent variables can be analysed by solving for the values of w and b using the least squares method:

$$\omega = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2} \tag{6}$$

$$b = \frac{1}{n}\sum_{i=1}^{n} (y_i - \omega x_i) \tag{7}$$

2.2.4 Random Forests

Random Forest is a supervised learning algorithm for classification and regression problems. It is an integrated model consisting of many decision trees. The core idea is that when training data is fed

into the model, instead of building a large decision tree with the entire training dataset, Random Forest uses different subsets and feature attributes to build multiple smaller decision trees, which are then combined into a more powerful model. By combining the results of multiple decision trees, random forests can enhance the model. Another important feature of random forests is that each subset is built from randomly selected samples and randomly selected feature attributes. This randomisation reduces the sensitivity of the decision tree to the training data, thus preventing overfitting [8].

A visualisation of the structure of a random forest is shown in Figure 1:

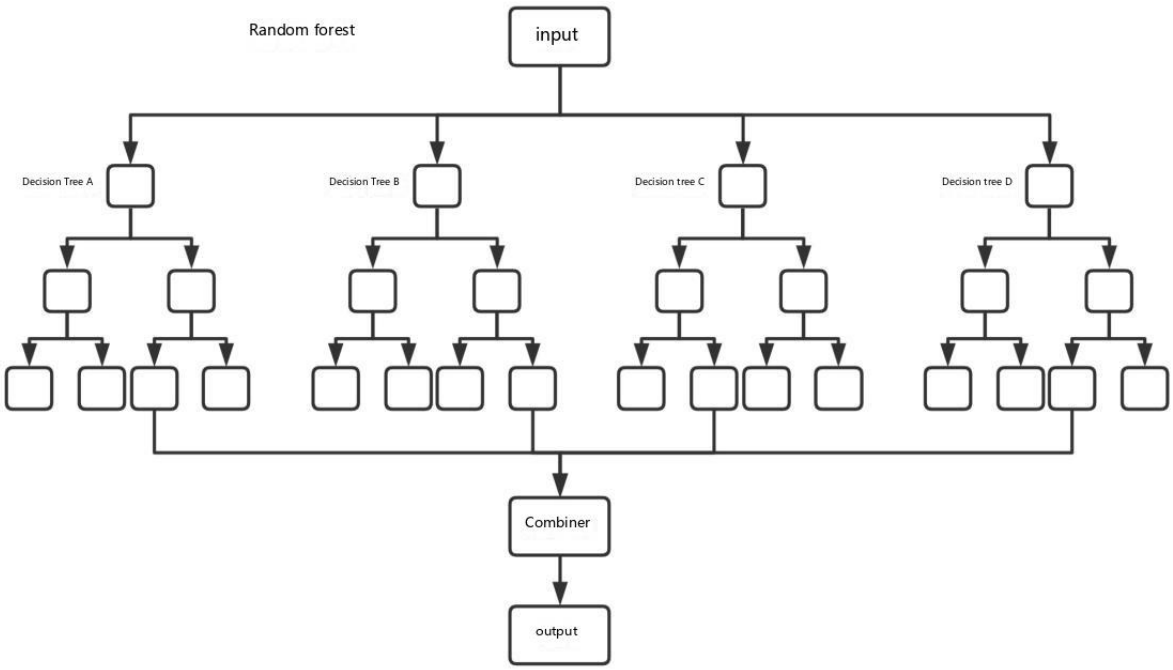


Figure 1. Random forest structure

Random forest is an integrated learning method consisting of decision trees, and its basic principle can be summarised in three steps: random sampling, random feature selection, and majority voting.

The first step in Random Forest is random sampling, i.e., sampling the training data set with putback. Assuming that the size of the training dataset is N, n data are selected from N data to form a new training set. This new training set is randomly selected from the original data with putback, i.e., the same data can appear multiple times in the new training set or may not be selected from the original data set. The purpose of random sampling is to reduce the risk of overfitting by averaging multiple random subsets, which can improve the model's ability to generalise to new data [8].

In each decision tree of a random forest, the feature selection for each node is random. Specifically, instead of all features being used for model training, a subset of k features is randomly selected from the full set of feature attributes, and each decision tree selects an optimal feature attribute from this subset of k features as the division node, so each node only considers a random subset of features, not the full set of feature attributes. Therefore, compared to the method of applying decision trees alone, the random selection of features can make the rules of decision trees more randomised, thus reducing the variance of the algorithm and improving the stability and accuracy of the algorithm [8].

The final prediction of the Random Forest is derived by combining the results of all the decision trees. For classification problems, each decision tree outputs a classification label, and the Random Forest performs a majority vote on the classification labels output by all decision trees to arrive at the final classification result. For regression problems, the regression results output by each tree are subjected to an averaging operation to produce the output of the random forest. The purpose of majority voting is to combine the results from multiple trees to enhance the accuracy and generalisation of the model.

3. Modelling and solving

3.1. Information Entropy and Extraction of Preference Indicators

Using equation (1), the information entropy of all letters is calculated. Where, where p (x_i) denotes the probability of occurrence of each letter in In this paper , the probability is replaced by the frequency of occurrence of the letter in this paper. The greater the information entropy of a word, the more information it contains. When guessing words, players tend to choose words with more information, i.e., words with higher information entropy. However, in the first few choices, players avoid words with repeated letters to gain more useful information; similarly, the later the number of guesses, the more likely people are to use repeated letters, at which point the entropy of repeated letters cannot be ignored. This difference may affect the prediction of the distribution of reported results. Therefore, the information entropy is optimised in this paper [9]. For words with repeated letters, this paper multiplies the information of the second occurrence of the letter by 0.5 and multiplies the information of the third or more occurrences of the letter by zero.

In addition, people's preference for letters of commonly used words when guessing words is considered. This paper quantifies this preference metric for use in the analysis. In this paper, we statistically analyse the frequency of occurrence of each letter in different positions and express the sum of the letter frequencies of each word to quantify people's preference for choosing commonly used words, i.e...

$$\delta = \sum_{j=1}^{n} \gamma_{\alpha j} \tag{8}$$

Where$\gamma_{\alpha j}$ denotes the probability of the letter α appearing in column j. In this question, this paper uses the frequency of the letter appearing instead of the probability. In order to reduce the subjective influence of the model on quantifying people's word choice preferences as and to verify the reliability of the model, this paper obtains the word frequency of the corresponding words by consulting the frequency of commonly used words in the Corpus of Contemporary American English, and takes it as a quantitative standard for the preference of commonly used word choices, and the quantitative data of the preference value is basically reliable[10].

In this paper, we consider the influence of previous choices on subsequent choices, with the increase of the number of attempts, the cue of letter colour affects the player's choice result, and the influence of common word choice preference decreases with the increase of the number of attempts. Therefore, in this paper, based on the sequence of the number of attempts, the probability of the appearance of letters in different attempts was assigned using hierarchical analysis-entropy weighting method. The magnitude of the entropy weights for each attempt was calculated using equations (2) (3) (4), and the weights were obtained by combining the results of the hierarchical analysis method. The sum of the weighted probabilities of the letters for each word was calculated separately, which is the final quantitative value of the preference.

The quantified values of information entropy, optimised information entropy and preference for the sample words are shown in Table 1.

Table 1. Quantitative values of word feature

word	Information entropy	Optimised information entropy	Preference quantization value
gorge	1.186	1.04	0.496
dodge	1.065	0.913	0.459
egret	1.373	1.036	0.341
slump	0.958	0.958	0.397

3.2. Linear regression models

In the process of guessing the same word, the results of previous guesses will have an effect on subsequent guesses. Considering this factor, this paper uses a recursive approach to predict the

following data. For example, in predicting the percentage of third attempts, this paper considers the impact of the first attempt, the results of the second attempt, information entropy and preference indicators.

For word guessing attempts, the variables show significance and hence this paper uses least squares for prediction. In this problem, this paper uses information entropy, preference data and percentage of previously predicted attempts as independent variables x_1, x_2 .... .x_k, , and study the relationship between the above independent variables and the k+1st attempt. Make an independent observations of y and x_1, x_2. Obtain x and n sets of observations (x_{t1}, xt_2 ..., x_{tk}) which satisfy the following equation.

$$y = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \varepsilon_t \tag{9}$$

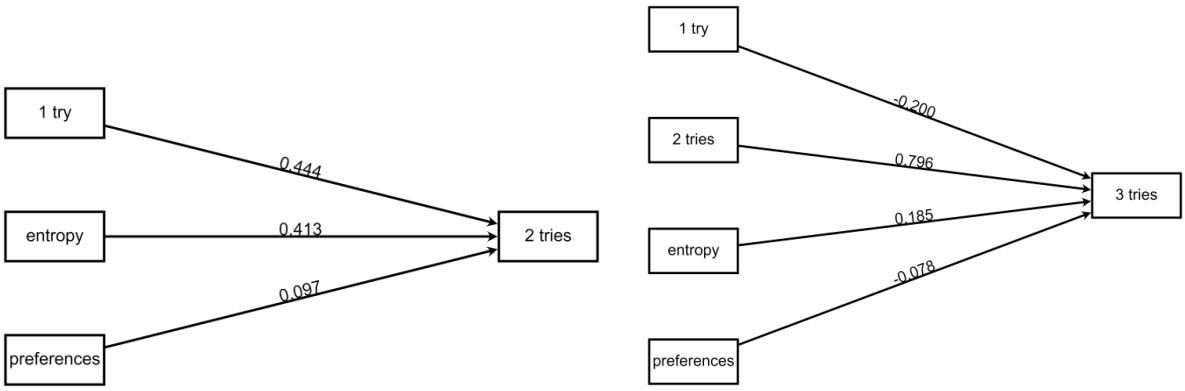
Construct the matrix and use the least squares method to solve for the correlation coefficients.

$$\beta = (X^T X)^{-1} X^T Y \tag{10}$$

$$u(b) = \frac{\sigma}{\sqrt{\sum (x - \bar{x})^2}} \tag{11}$$

$$u(a) = u(b)\sqrt{\frac{\sum x_i^2}{n}} \tag{12}$$

The effect of each variable on the number of attempts was analysed sequentially to obtain the correlation coefficients for the predictive analysis of two or three attempts as shown in Figure 2.



(a) Three attempts to predict results (b) Two attempts to predict results

Figure 2. Results of the impact of word characteristics

The model model fit R^2 is close to 1, which is a good fit, indicating that the word features are related to the distribution of the number of guesses. A graph comparing the prediction results for the third attempt with the actual data is shown in Figure 3:

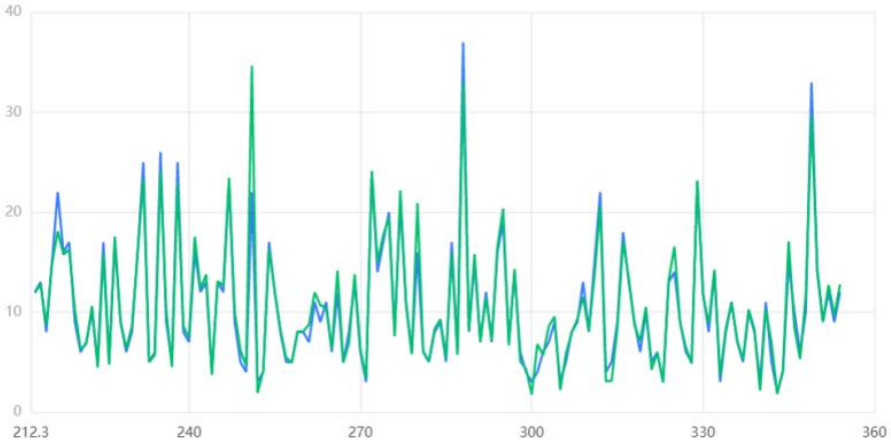


Figure 3. Fitting effect

3.3. Random Forest Modelling

The essence of attempt distribution analysis based on Random Forest algorithm is to make regression prediction for each attempt. The prediction results are obtained by analysing the data of different features of the word. The extraction of feature quantity is the key factor for the prediction analysis based on Random Forest Algorithm, so it is necessary to extract the appropriate feature quantity from the factors affecting the complexity of the word [6]. The collated data is used as the input to the Random Forest algorithm, and the subsequent prediction percentage is used as the output of the Random Forest algorithm. In the random forest algorithm, a large number of feature quantity data need to be trained several times, and finally get a more stable model [7-8].

From the above, it can be seen that word information entropy, quantitative value of choice preference and percentage of previous attempts will have a greater impact on predicting the percentage of subsequent attempts, so the above metrics are chosen as feature quantities.

During training, P samples are randomly selected from the given word data to construct P decision trees. Each decision tree can grow without constraints and pruning while remaining independent without correlation. For any sample X in the given word data, the random forest model trains P sub-models and will generate P predictions. Assuming that the kth submodel's prediction is Y_k , then the output value of the total model Y_E is the mean of the predicted values of all submodels, i.e:

$$Y_E = \frac{1}{P}\sum_{k=1}^{P} Y_k \tag{13}$$

The distribution of the sixth attempt of the given word was predicted and the comparison between the predicted and actual results is shown in Fig. 4:

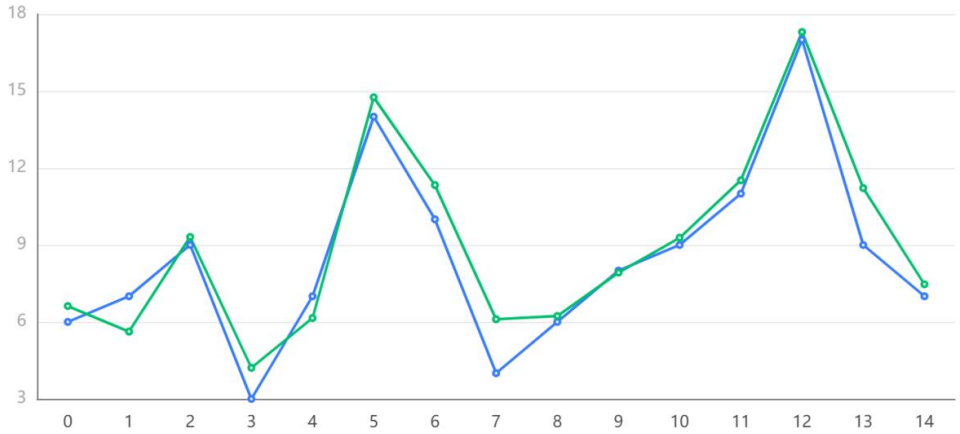


Figure 4. Comparison of random forest fitting results

Meanwhile, the contribution of each feature involved in the decision tree operation is examined in this paper to verify the validity of the selected feature quantity [5]. The importance analysis of the features is mainly based on out-of-bag (OOB) data, which is a dataset consisting of sample points that are not selected each time the model performs random sampling and replacement of the training set. The importance of a variable is measured by the percentage increase in the mean square error (IncMSE%) of the OOB data.IncMSE% indicates that removal of the variable reduces the accuracy of the target prediction, so more important variables have a higher IncMSE%. For the decision tree, the corresponding variables from the OOB data are placed into the decision tree before and after scrambling, and then their IncMSE% is calculated. Assuming that there are N trees in the forest, the IncMSE% for K trees is.

$$IncMSE\%(i) = \sum_{K=1}^{N} \frac{(OOB_{k2} - OOB_{k1})}{OOB_{k1}} \times 100\% \tag{14}$$

In the prediction of the sixth attempt**,** the model OOB score was 0.914, with the third and fourth attempts accounting for 51.8% of the data having a greater impact on the prediction.

In this paper, we have evaluated the model performance using k-flod cross-test by dividing the word data into k folds, using k-1 folds of them as the training set each time and the remaining one as the validation set, which was used to evaluate the model performance. This was repeated k times, each time using a different validation set. Ultimately, the mean of the k validation results was used as the performance metric of the model. By repeatedly performing cross-validation, the generalisability of this model is better represented and the possibility of overfitting phenomenon occurring is reduced.

The relevant indicators for model evaluation are shown in Table 2:

Table 2. Model evaluation parameters

	MSE	RMSE	MAE	MAPE	R^2	obb_score
training set	0.586	0.766	0.486	4.521	0.986	0.914
test set	1.625921312754241	1.275116195785404	0.9885462838599038	12.879278000131864	0.9143775672244915	-

The model was tested to have a high goodness of fit for the distribution of the seven attempts. Therefore, it is concluded that the model has a good fit to the data distribution of the reported results. The goodness of fit of the model for seven attempts is shown in Table 3:

Table 3. Goodness of fit to the attempted distribution.

variable	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 tries
R^2	0.71	0.75	0.79	0.91	0.97	0.98	0.97

3.4. Comparison of models

By comparing the goodness of fit R^2 of the two models, the random forest model has better prediction, and the results of the comparison of the goodness of fit of the random forest model and the linear regression model are shown in Figure 4:

Table 4. Comparison of model fit goodness of fit

variable	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 tries
random forest	0.71	0.75	0.79	0.91	0.97	0.98	0.97
linear regression (math.)	0.52	0.56	0.62	0.53	0.88	0.94	0.92

The random forest model was applied to analyse the distribution of the number of guesses for the example word ERRIE and the results are shown in Figure 5:

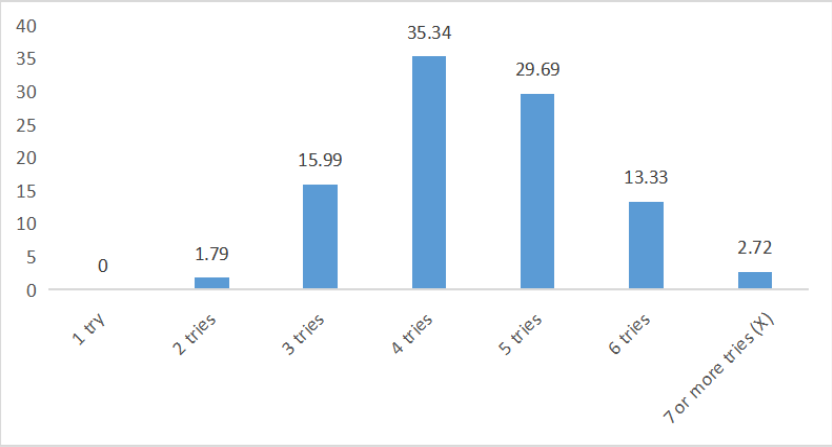


Figure 5: Attempted distribution projections

The random forest model has a strong anti-interference ability, and the accuracy of the prediction results is higher. At the same time, the model is more resistant to overfitting and is more effective in regression analysis.

4. Summary

In this paper, we constructed a regression model and a random forest model, analysed the relationship between daily words and the corresponding attempts distribution in Wordle game and predicted the distribution of attempts share. For word complexity, this paper chooses information entropy to quantify this index. Considering the influence of previous attempts on the follow-up when guessing words, this paper integrates the quantitative index of common letter preference, information entropy and the proportion of previous attempts for prediction analysis. By comparing the model results, the random forest fits well and predicts the distribution of the percentage of attempts well. This paper provides a theoretical basis for predicting the distribution of word attempts, and it is hoped that it will help to optimise the Wordle game thesaurus, help it to expand more ways to play in order to enhance the fun, provide players with a better gaming experience, and attract more players.

References

[1] Siddhant Bhambri, Amrita Bhattacharjee, and Dimitri Bertsekas. Reinforcement Learning Methods for Wordle: A POMDP/Adaptive Control Approach, 2022.

[2] Huang Yitai, Zhong Zeheng, Fang Zhaoyang. Prediction and Classification Model Based on Wordle's Date [J]. Advances in Computer, Signals and Systems, 2023, 7(5).

[3] Xuyi Shi, Jiachen Guang, Liangsu Shao. Wordle data analysis based on time series analysis model [J]. Academic Journal of Mathematical Sciences, 2023, 4(2).

[4] Weicun Zhang. (2022). Compare Linear regression, Decision Tree Regressor, and Random Forest Regressor based on python, a restaurant company on Kaggle as a case... (eds.) Proceedings of 2022 International Conference on Company Management, Accounting and Marketing, CMAM 2022: 323-330.

[5] Dan T, Mingchao L, Yang S, et al. Intelligent mining of safety hazard information from construction documents using semantic similarity and information entropy [J]. Engineering Applications of Artificial Intelligence, 2023,119.

[6] Yang Q, Fang Y, Zheng Y. Word Data Research and Prediction Based on Wordle Game [J]. Academic Journal of Computing & Information Science, 2023, 6(4).

[7] Guoji X, Huan W, Jinsheng W, et al. A Local Weighted Linear Regression (LWLR) Ensemble of Surrogate Models Based on Stacking Strategy: Application to Hydrodynamic Response Prediction for Submerged Floating Tunnel (SFT) [J]. Applied Ocean Research, 2022, 125.

[8] Linrong X, Jiyong D, Liping Y, et al. Random forest algorithm-based accurate prediction of rat acute oral toxicity [J]. Molecular Physics, 2022, 120(24).

[9] Siddhant D, S. A K. Development of function-specific indices for assessing water quality based on the proposed modifications of the expected conflicts on existing information entropy weights [J]. Environmental Monitoring and Assessment, 2022, 194(12).

[10] Yitai H, Zeheng Z, Zhaoyang F. Prediction and Classification Model Based on Wordle's Date [J]. Advances in Computer, Signals and Systems, 2023, 7(5).