Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2318226

"The Big Bang Theory" of Wordle

Summary

Wordle is a popular puzzle currently offered daily by the New York Times, where players try to guess a five-letter word in six chances, with the number of players changing daily since the game was released.

Problem 1:In order to accurately predict the number of players,we improved the basic Logistic Growth Model and built the **Improved Logistic Growth Model** to simulate this process which the number of players decreases after reaching the peak and finally plateaus. Finally, we made predictions and obtained the confidence **interval for March 1, 2023, when α = 0.05, 28245$\pm$2.8745.** In order to investigate which word attributes affect the percentage of player scores reported in the difficult mode, we first reasonably evaluated the percentage of player scores by using **TOPSIS** to get the composite score, then defined and extracted the word attributes by **Hypothesis Testing** to determine whether the attributes affect the composite score and establish a word attribute extraction model, **found a significant relationship between word sentiment and whether the word contains repeated letters and the distribution of difficulty.**

Problem 2: In order to predict the distribution of player scores for a given word, the word attribute extraction model is used to determine the variable input to the model, and after preprocessing the data, a **Neural Network Model** for predicting the distribution of scores is built to output the distribution of player scores. The scores of the model on the sample set and the training set were 60.64% and 68.94%, respectively, and **the final score distribution prediction results for EERIE were (0,6,23,36,24,9,2).**

Problem 3: We built an improved clustering decision tree model (**Cluster-DT Model**), using the attached data to cluster the words first, and output the classification rules through the decision tree model, while taking into account the assurance of rationality and stability of the number of difficulty categories, classify the words into three difficulty levels: simple, average, and hard. The combined use of the word feature extraction model The word score distribution is obtained for any word by using the neural network score prediction model and inputting the cluster-DT model to obtain the difficulty level of the word. In order to study the relationship between the difficulty classification of the Cluster-DT model and the word attributes, we used **PCA** to downscale the attributes and calculate the scores of each attribute, concluded that **(a) the more frequently a word is used in life, the lower the difficulty of the word; (b) the initial letter of a word directly affects the difficulty of the word. (c) The presence of repeated letters in a word increases the difficulty. (d) Words with positive meanings tend to be easier.**

Problem 4: Through the exploratory data analysis in the process of modeling described above, **we point out some interesting points about the data:** (i) the number of players in the difficult mode is gradually in growth and in a flat proportion. (ii) The number of players tends to increase more on weekdays (except Thursdays) and decrease on weekends. (iii) Wordle's solution words have more nouns and most of them are neutral words, and the proportion of letters appearing in words is roughly similar to that in Wikipedia.

Keywords: Logistic Growth Model, Neural Network Prediction, Cluster-DT Classification Model, PCA, EDA

Contents

1 Introduction

1.1 Problem Background

Wordle, a popular daily puzzle currently offered by The New York Times, is a type of crossword puzzle that has a long history and wide audience in English-speaking countries, with a wide variety of variations. This game is updated daily, and the player's goal is to guess a five-letter word within six attempts, using the information returned after each attempt.

By analyzing the question, we know that for a game, it is most important to predict the number of players accurately and reasonably, and to analyze the reasons for the fluctuation of the number of players. Then back to the game mechanism of wordle itself, different words make the distribution of players' scores different on that day, which attributes of the words affect the players' scores, and further whether we can build a model to predict the players' scores for a given word on a certain day in the future, and finally to classify the difficulty level of the words.

Four major problems are discussed in this paper, which are:

- Build a model to accurately predict the number of players
- Build a model to extract the attributes that affect the player's score
- Build a model to predict the distribution of the score using the attributes of the word.
- Build a model to classify solution words by difficulty.

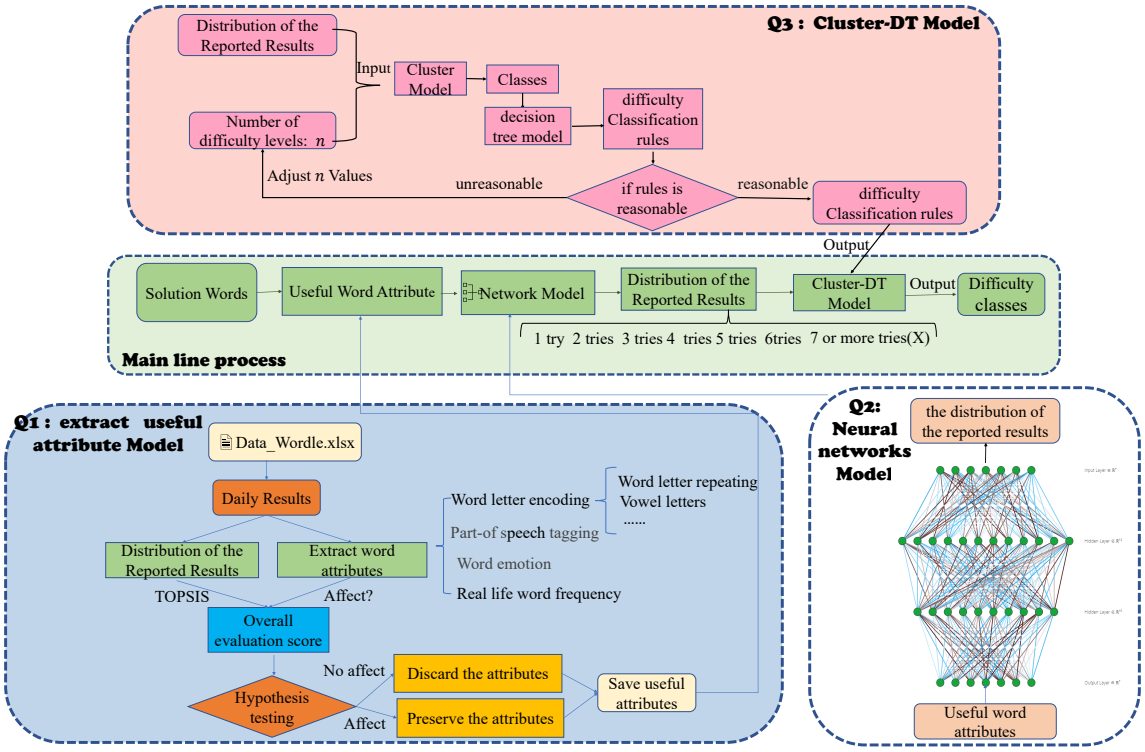


Figure 1: Flowchart of our work

1.2 Our work

Some data problems were found during the modeling process, which have been corrected by querying Twitter. In order to solve the first question of the first problem, We built an Improved Logistic Growth Model to characterize the trend changes in the number of people and to forecast the future number of people.

The rest of our models are showed in the following flow chart 1, Among them, for the sake of completeness and readability of the article, we put all the contents of the prediction and difficulty rating of the word EERIE in the fourth part to apply to the model.

2 Preparation of the Models

2.1 Assumptions

To simplify the model, we make the following assumptions:

- It is assumed that players will not know the answer in advance neither will they share the answer.

  ⇒Otherwise it will affect the accuracy of the Distribution of the Reported Results.

- It is assumed that the Wordle has a certain number of followers Before attachment data collection.

  ⇒Otherwise the game can not reach his social attributes and slow development.

- It is assumed that the player guesses correctly on the first try is likely to be a wild guess.

  ⇒It is almost impossible to guess right without receiving any feedback, a small weight should be given in the comprehensive evaluation.

Table 1: Notations

Symbol	Definition
$N(t)$	number of players per day
T or t	time
r	per rate of increase
r_{max}	maximum per rate of increase
K	carrying capacity
ρ	growth rat
N_0	starting number
N_m	maximum number of people
$z(t)$	external factor
D	Distribution of scores
A	Word attribute matrix

2.2 Notations

The primary notations used in this paper are listed in Table 1.

3 The Models

3.1 Task 1:Predict the Number of Reported Results

For a new game, the number of players changes all the time. When the game is first launched, it is hot and the number of players keeps growing, but as time changes, when the number reaches its maximum, the game passes its hotness and the number of players decreases until it finally plateaus. To explain this change and predict the number of players at a future time, we consider its overall trend and use a modified Logistic Growth Model with a coefficient of variation to portray the number of players.The trend of the number of games is shown below.

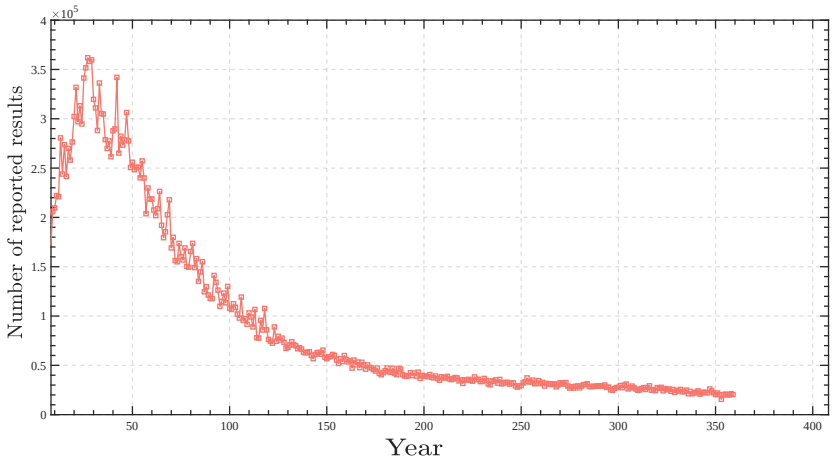


Figure 2: The trend of the number of games

3.1.1 The Logistic Growth Model

Logistic function or logistic curve is a common S-shaped function named by Pierre-François Vélule in 1844 or 1845 when he studied its relationship with population growth. The generalized logistic curve can mimic the S-shaped curve of population growth (P) in some cases. The initial phase is roughly exponential; then the increase slows down as it starts to become saturated; finally, the increase stops when it reaches maturity.

The detail can be described by equation :

$$\frac{dN}{dT} = r \cdot N \tag{1}$$

$$\frac{dN}{dT} = r_{max} \cdot \frac{K-N}{K} \cdot N \tag{2}$$

The equation (1) is a general equation for the population growth rate (change in number of individuals in a population over time).In this equation,$\frac{dN}{dT}$ is the growth rate of the population in a given instant, N is population size, T is time, and r is the per capita rate of increase that is, how quickly the population grows per individual already in the population.

We can make more specific forms of it to describe two different kinds of growth models: exponential and logistic.When the per capita rate of increase (r) takes the same positive value regardless of the population size, then we get exponential growth. When the per capita rate of increase (r) decreases as the population increases towards a maximum limit, then we get logistic growth.r_{max} is the maximum per capita rate of increase for a particular species under ideal conditions, and it varies from species to species.The maximum population growth rate for a species, sometimes called its biotic potential, is expressed in the equation (1) ,which is called Exponential Growth.

Exponential Growth may happen for a while, if there are few individuals and many resources. But when the number of individuals gets large enough, resources start to get used up, slowing the growth rate.We can mathematically model logistic growth by modifying our equation for exponential growth, using an r(per capita growth rate) that depends on population size N and how close it is to carrying capacity K. Assuming that the population has a base growth rate of r_{max}, when it is very small, we can write the equation (2).

3.1.2 Improved Logistic Groewh Model

Accprding to The Logistic Growth Model,we develop a Improved Logistic Growth Model to describe the variation of the number of reported results.

First, we define the expansion degree, i.e., the growth rate ρ, which represents the ratio of the growth rate of the number of people to time,

$$\rho = \frac{dN}{dt} \cdot \frac{1}{N} \tag{3}$$

where N represents the number of users, which can be seen as a differentiable function of time t due to the large number, and $N(t)$ represents the number of people playing the game at moment t. Since the number of people playing the game will not keep growing continuously, it will be limited by the effect of blocking such as freshness on the daily number of people N, making the expansion degree ρ decrease with the increase of the number of people. The p will be expressed as a function of the daily number of people N,$\rho(N)$, so there is

$$\frac{dN}{dt} = \rho(N) \cdot N, N(0) = N_0 \tag{4}$$

by assuming that:

$$\rho(N) = \rho - S \cdot N, (\rho > 0, S > 0) \tag{5}$$

Set the maximum number of changes in the number of people N_m, when $N = N_m$, the number of players no longer grow, at this time the growth rate $\rho(N_m) = 0$, brought into the above formula

$$\rho(N) = \rho \cdot (1 - \frac{N}{N_m}) \tag{6}$$

brought into a formula to get

$$\frac{dN}{dt} = \rho \cdot N \cdot (1 - \frac{N}{N_m}) \tag{7}$$

and finally solved

$$N(t) = \frac{N_m}{1 + (\frac{N_m}{N_0} - 1) \cdot e^{\rho t}} \tag{8}$$

The fitted equation is as follows:

$$N(t) = \frac{361908}{1 + (\frac{361908}{80630} - 1) \cdot e^{-0.004t}} \tag{9}$$

The model fitting effect is analyzed as the folowing figure 3 showed, when only the data before N_m is fitted or the data after N_m is fitted, the effect is better, but as far as the overall trend is concerned, because the number of players keeps decreasing after reaching the maximum, so the fitting curve receives a tendency of compression more than growth.

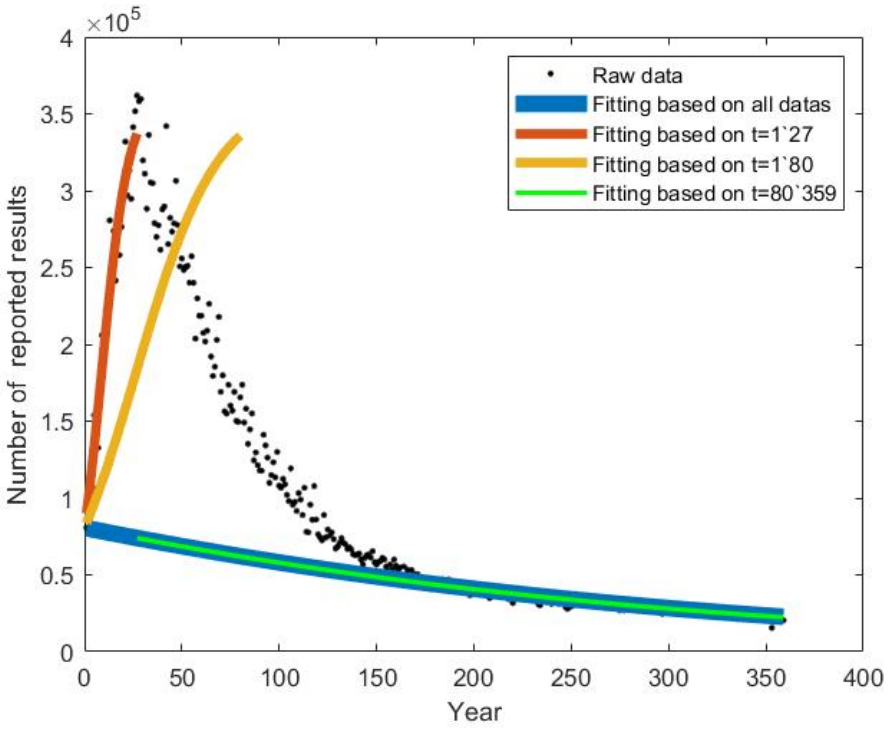


Figure 3: Fitting results

So the model is considered to be improved by adding a change factor to alleviate the larger change in the original data.The multiplicative factor of the change in the number of players due to external factors is denoted by $z(t)$,

$$z(t) = \frac{N(t)}{N_{t-1}} \tag{10}$$

Fitting the data for the rate of change yields the equation:

$$z(t) = 0.0003599t^2 - 0.071781t + 5.6725 \tag{11}$$

After multiplying the coefficient of variation with the fitted equation, we found that the shape of the resulting graph is already incomparably close to the original graph. Considering that at the beginning of the game release, there existed a certain fan base for the operation account, solidly adding a constant number to the existing model, representing the old players who started to spread the function. Finally, the results of the model fitting are shown below figure 4, where the confidence interval is obtained from the hypothesis test $\alpha = 0.05$ when fitting the data.And finally, the number of reported results predicted by this model for March 1, 2023, was 28245 ± 2.8745.

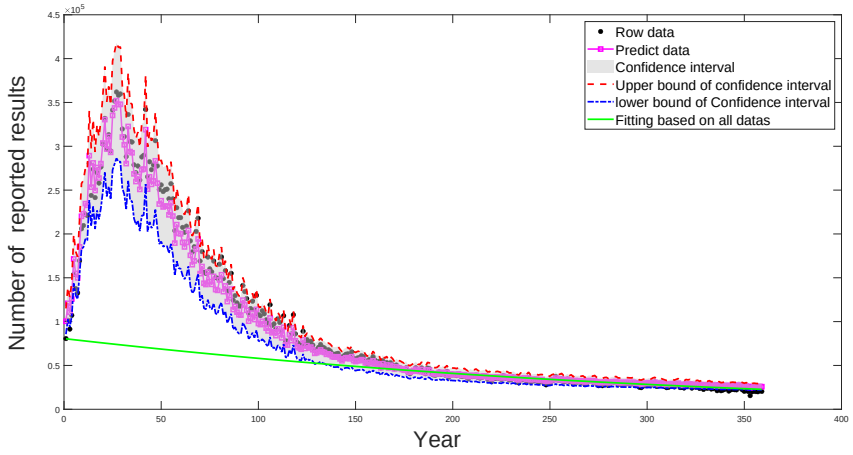


Figure 4: Final predict

3.1.3 Conclusion

Through the analysis, it can be seen that the Wordle game at the very beginning, due to the wide spread of the Internet and certain social attributes, the number of its players grew sharply and the growth rate accelerated, and then reached the maximum number of players, because it is only a web-based game, after the novelty, the number of players gradually decreased slowly and gradually tended to a stable state, the influence factor in the model to a certain extent.The influence factor in the model reflects the change of people's enthusiasm for this game to a certain extent, and finally acts on the number of players.

3.2 Task 2:Useful Attribute Extraction & Neural Network Prediction

Since the attributes of words are inextricably related to the percentage of scores in the difficulty mode, i.e., the percentage distribution of scores can roughly reflect the difficulty level of a word, and if a word is guessed correctly in the first few times, it indicates that it may have better attributes, and if it is guessed more times, it indicates that words with such attributes are not easy to be guessed. In order to investigate how these attributes affect the distribution of the final score, and to facilitate the subsequent prediction of this distribution, we summarize some common attributes of words based

on existing studies, and first use TOPSIS to make a comprehensive evaluation of the difficulty of words using the distribution of scores as an indicator.Once the final score is obtained, the words can be ranked in terms of difficulty based on this score, and then the attributes that affect the score of the words can be analyzed based on their characteristic attributes, and the influential attributes can be finally identified through ANOVA to complete the feature extraction.

After the useful feature attributes are available, a neural network can be trained as the input layer of the neural network, and the corresponding difficulty percentage can be used as the output to achieve the prediction effect.

3.2.1 Useful Attribute Extraction Model

Details about TOPSIS in Score of Comprehensive Evaluation

The Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) is a multi-criteria decision analysis method, TOPSIS is based on the concept that the chosen alternative should have the shortest geometric distance from the Positive Ideal Solution (PIS) and the longest geometric distance from the Negative Ideal Solution (NIS).It is a method of compensatory aggregation that compares a set of alternatives, normalising scores for each criterion and calculating the geometric distance between each alternative and the ideal alternative, which is the best score in each criterion.

We use the percentage distribution of the number of guesses of existing words as an indicator, where the less the number of guesses used represents the word is easier, it is defined as having better properties, and it is treated as a positive indicator, if the number of guesses used is more, it means that the word is difficult to be guessed, and its properties are defined as a less good class, and it is treated as a negative indicator, for existing indicators, it is defined as having different weights, and since it is assumed that the first guess is likely to be a random guess, its weight is reduced appropriately, and the specific details of the indicator are as follows:

Distribute	Indicators Classes	Weight
1 try	Positive indicators	0.15
2 tries	Positive indicators	0.2
3 tries	Positive indicators	0.1
4 tries	Intermediate indicators	0.1
5 tries	negative indicators	0.1
6 tries	negative indicators	0.15
7 or more tries (X)	negative indicators	0.2

The final scores were ranked and the results are as follows figure 5:

To investigate the effect of this model, we took the top four words and the bottom four words to observe their score distribution as shown in the figure 6:

We can see that the words with higher scores use fewer correct guesses and the score graph is left skewed, while the words with lower scores use more correct guesses and the score graph is right skewed; indicating that the model works well. The following is an analysis of each attribute feature separately.

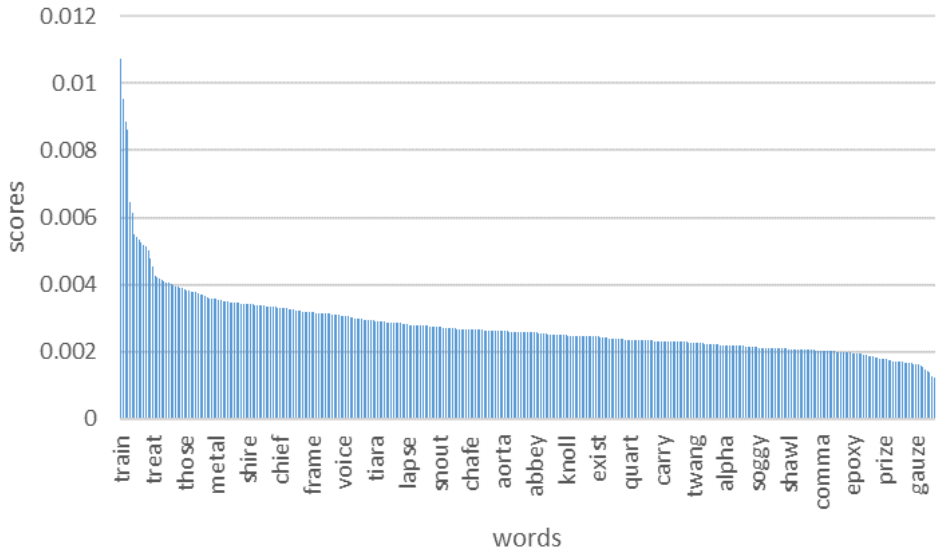


Figure 5: TOPSIS scores

	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more
train	6	26	32	22	10	3	0
slate	6	14	33	27	13	5	1
dream	5	14	31	29	15	4	1
feast	5	13	25	27	19	10	2
gawky	1	5	22	33	28	10	10
judge	2	8	16	26	33	14	14
coyly	0	4	17	28	35	15	15
mummy	1	4	14	27	37	18	18

Figure 6: TOPSIS renderings

Details about Hypothesis Testing

In this section, in order to further explore how the attributes of words affect the difficulty score, we select the parts of speech, the meaning of the word and the frequency of the word appearing in life, as well as the number of repeated letters in each word as its attributes. Among them, the parts of speech are divided into adj, adv, n, v, and other; The word meaning is divided into neg(negative),neu(neuter) and pos(postive), which are coded respectively and analyzed by ANOVA.

Finally, we found that the different word meanings and the presence or absence of repeated letters in the words have a significant effect on the difficulty, while the wordness can also be considered significant at $\alpha = 0.5$.The results of the ANOVA are shown in the following table :

Table 2: **Results of ANOVA**

Variables	SSA	SST	F	P
parts of speech	3.3611×10^{-6}	0.00039	0.77	0.5465
meaning of word	3.872×10^{-4}	0.00039	1.74	0.1773
number of repeated letters	0.00002	23	24.42	1.2×10^{-6}

Further, we also found that the frequency of words occurring in life also has a relationship with the final score, and the variation of words with score is shown in the following figure 7:

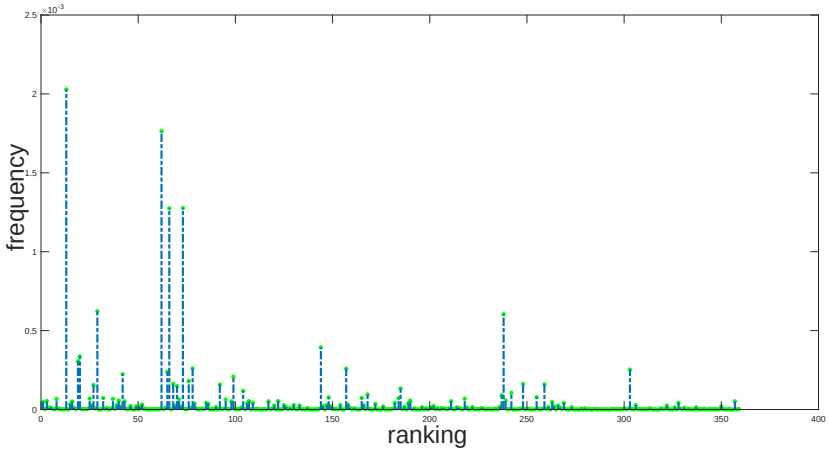


Figure 7: Change of frequency

3.2.2 Neural Network Model to Predict the Distribute of Score

Predicting the distribution of players' scores for a given word is actually a multiple-input multiple-output (MIMO) modeling problem, which is difficult to solve by statistical algorithms and machine learning in general. Again, unlike general NLP problems, the NLP problem involved in this question is relatively simple.So the choice was made to use a neural network to predict the distribution of scores, and to avoid overfitting,so we avoid using an over-complexity neural network,only the Multi-Layer Perceptron is used.

Preparing training data: word useful attribute encoding

A neural network model is largely dependent on the preparation of training data. Using attribute extraction model to get the useful attribute, such as:lexical category, word frequency of use in life, the emotion of the word, whether there are duplicate letters $\cdots$.

Use the word slump from 2022-Jan-7 for example Table 8.

Table 3: the feature attributes of slump

Attribute	Values	Attribute	Values
w0	18	adv	0
w1	11	n	1
w2	20	v	0
w3	12	neg	0
w4	15	pos	0
cf	0	compound	0
adj	0	frequency	1.37×10^{-6}

we can get the attribute data of the slump, by the same logic, you can get the attribute data A of

any word, where A is 14 dimensions. By normalizing frequency attributes:

$$A[frequency] = \frac{A[frequency] - \mu}{\sigma} \tag{12}$$

Here μ, σ means the mean value and standard deviation of the frequency of the attribute.

By this time the word attribute data has been prepared for training.

Neural network training

Using the attribute data A in the data preparation phase, the distribution of player scores D is used as the value to be predicted.

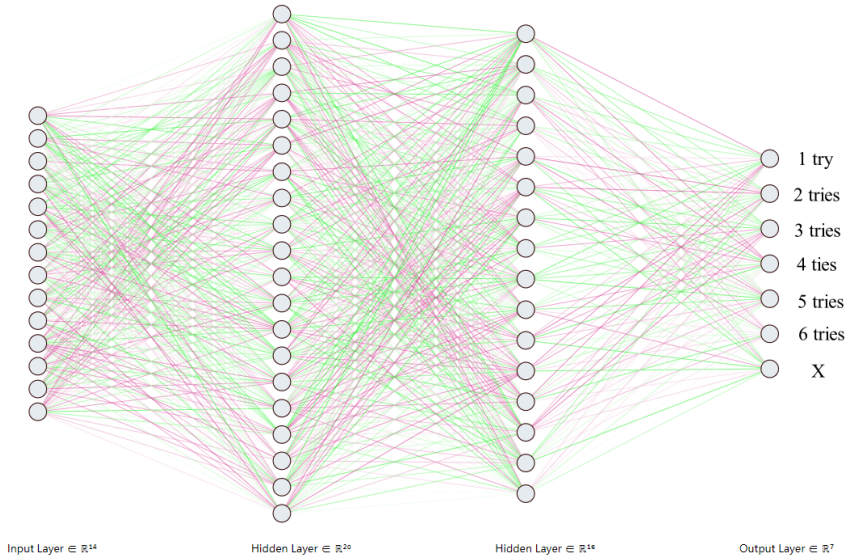


Figure 8: Neural Network Model

Basic Network Architecture Figure 8, the following is main parameters of the network:

- environment: python 3.9.6, Pytroch 1.13.1.
- hidden layer sizes: $(29, 48)$.
- activation function: *relu* function.
- Optimizer: Adaptive Moment Estimation.
- alpha: adaptive learning rate.

Evaluation of neural network models

By cross-validation, the correlation coefficient of the neural network reached 69% on the training set and 61% with 359 samples in the training set at the end of training, the more information as can see at Table 4.

The predicted and true values of the word slump's distribution are given below, Table 5,can see that the results are nice.

Table 4: Metrics for Neural Network Models

	MSE	MAE	Score
Sample set	10.521	2.192	60.64%
Train set	6.476	1.785	68.94%

Table 5: True distribution and neural network predicted values

	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (X)
Real Distribute	1	3	23	39	24	9	1
network model	0	4	25	38	21	8	1

Conclusion

For a given future word,to predict the distribution of the reported scores,Neural Network Model can help to predict the distribution of the reported. By using some useful attributes, after encoding these attributes, the neural network is trained with them as the input layer, and the final distribution of difficulty is used as the output, eventually reaching the point where given a word in the future, it can be used according to its attributes to predict the associated percentages of $(1,2,3,4,5,6,X)$. This model will also be used in the next difficulty classification model.

3.3 Task 3:Classify Solution Words by Difficulty.

3.3.1 Model Overview

Based on the model of problem 1 and problem 2, the extract model can be used to extract the feature attributes of a word such as "slump", the results are as follows,Table 8, and then the network model can be used to obtain the predicted score distribution of the word,Table 5.

From the above case, we can see that, based on the previous work, for any word that matches Wordle, the distribution of its score can be finally derived by the model, while the distribution of the score is exactly the reflection of the difficulty.

After that, using the distribution data of the scores, a combination of supervised and unsupervised learning is used to derive difficulty categories using a clustering algorithm, and further a decision tree model to derive rules for difficulty classification, so that for any word that satisfies the Wordle pattern a difficulty level can be obtained,algorithm 1.

3.3.2 Cluster-DT Difficulty Classification Model

1. Clustering process

As is well known, the clustering process is unsupervised learning, i.e., it is not necessary to know the category to which the words belong, and the distribution of scores is used to classify the words into several categories.

The main process in the model is to iteratively find n clusters by inputting the number of difficulty categories n and the distribution of word scores, which eventually minimizes the loss,The

Algorithm 1: Cluster-DT Difficulty Classification Model

Data: given word,Report Result Words,extract model,netWork model,Number of difficulty classes n

useful attribute ← extract model(Report Result Words);
distribution results← network model(useful attribute);
r=1;
while $r \neq 0$ **do**
    Difficulty Classification Model, Labels = K-Means(distribution results, n-cluster = n);
    Difficulty Classification Rules = Decision Tree(distribution results, Labels);
    if *Difficulty Classification Rules is reasonable* **then**
        rules= Rules;
        r=0
    else
        Adjust n values
    end
end
given word useful attribute← extract model(given word);
given word distribution results← network model(given word useful attribute);
given word diiculty class ← Rules(given word distribution results)

Result: given word diiculty class and the rules

loss function can be defined as $J(c,\mu)$

$$J(c,\mu) = \sum_{i=1}^{M} ||D(i) - \mu_{c_i}||^2 \tag{13}$$

The iterative process is the core step of clustering,Assign each word to the nearest cluster,then update the cluster.

$$c_i^t \leftarrow \underset{\mu}{argmin} ||D(i) - \mu_t^k||^2 \tag{14}$$

$$\mu_k^t \leftarrow \frac{1}{|C_k|} \sum_{D(i) \in C_k} D(i) \tag{15}$$

Finally, we get the class of difficulty,But no knowledge of classification rules.

2. Decision tree discovery the rules

To address the shortcomings of the K-Means model, the decision tree model is used to improve.

Decision trees are classical classification algorithms, which are supervised learning.

$$H(C) = \sum_{i=1}^{n} P(C_i) \cdot \log_2 P(c) \tag{16}$$

The core of the decision tree model is the selection of test attributes and the pruning technique of tree branches,Information entropy $H(C)$ is used.Using the categories classified by the clustering model, the decision tree model is used to explore the rules of difficulty classification.

You can derive the difficulty classification rules and the categories of all words by using this model, Go further using the previous model in combination, you can classify any word with a compound wordle rule and output the difficulty level.

In order to guarantee the homogeneity of the sample data volume and the significance of the differences in scores under different difficulty categories, it was finally decided to divide into three difficulty levels: **Hard,Average,Simple**.

So in now, the improved classification algorithm is derived,and named as: **Cluster-DT Difficulty Classification Model**.Table 6 is the three difficulty levels under the TOPSIS composite evaluation score.

Model classification rules,The following pseudo-code states the model of difficulty classification,algorithm 2:

Algorithm 2: Model classification rules

Data: Score distribution of one word: $D(i)$=percentages of$(1,2,3,4,5,6,X)$
Data: The difficulty category to which the word belongs: $C(i)$

```
if 3tries ≤ 22.5 then
    if 6tries ≤ 17.5 then
        c = Average
    else
        c = Hard
    end
else
    if 5tried ≤ 23.5 then
        c = simple
    else
        c = Average
    end
end
```

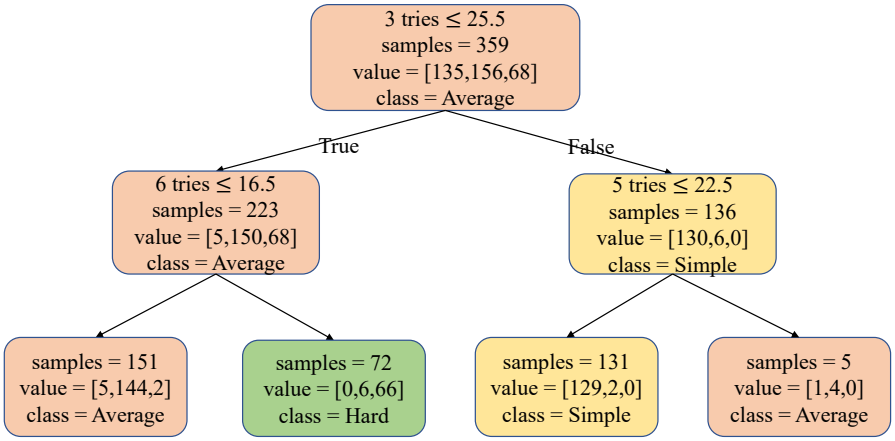


Figure 9: Difficulty classification rules

The result of the decision tree(Figure 9) matches the result of the clustering,It has an accuracy

rate of 0.95%,Performance can be considered good.

Table 6: Combined score of different difficulty levels

Classes	Sum	Mean	Standard deviation	Number
Hard	0.1389	0.0020	0.00044	68
Average	0.3842	0.0025	0.00038	156
Simple	0.1769	0.0035	0.00013	135

Take "slump" as an example were obtained using the previous model. Cluster-DT Difficulty Classification Model Using distribution data(Table 5) to get the difficulty category,Difficulty be classified as Average.

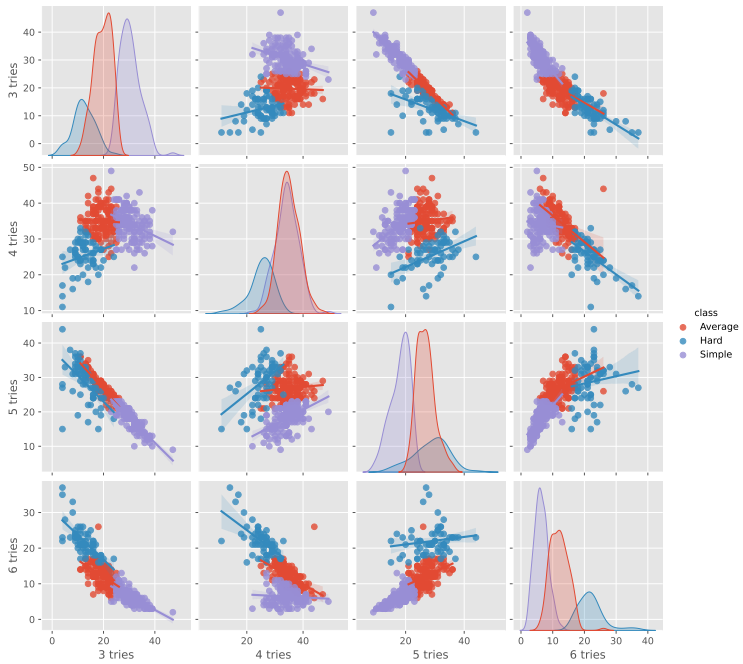


Figure 10: The relationship between the distribution of scores and categories

The model can give more cases of word difficulty classification,Table 7.

Table 7: Cluster-DT Model-based classification cases

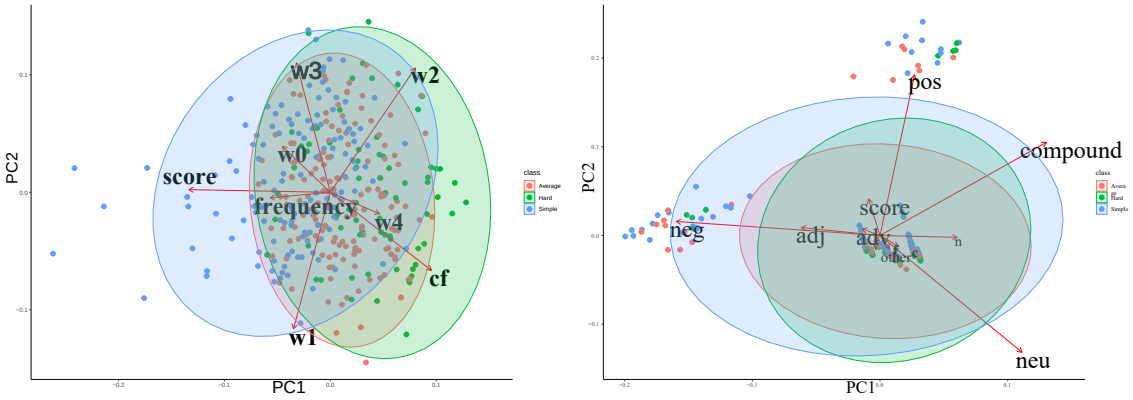
Word	**1 try**	**2 tries**	**3 tries**	**4 tries**	**5 tries**	**6 tries**	**7 or more**	**Topsis score**	**Class**
drink	1	9	35	34	16	5	1	0.003593523	Simple
panic	1	9	35	34	16	5	1	0.003593523	Simple
crank	1	5	23	31	24	14	2	0.002992659	Average
query	1	4	16	30	30	17	2	0.002760859	Average
dodge	1	3	15	29	27	19	7	0.002636123	Hard
trove	1	5	16	24	25	22	8	0.002654647	Hard

3.4 Correlation of Attributes Between Each Classification

Words can be classified into difficulty categories using the Cluster-DT model.In order to study more closely what attributes affect the difficulty division of words in the Wordle game,So a further study was done:

Using the Cluster-DT model to classify daily solution words by difficulty.Then analyze the correlation between the extracted word attributes and the difficulty classes.

The word attribute data A were dimensioned down by using principal component analysis(PCA), and the overall attributes were replaced by a few integrated word attributes, while the scores of each attribute were calculated, the scores were also extracted in a double-labeled plot(biplot).



(a) Alphabetic and frequency attributes of words (b) Lexical and emotional attributes of words

Figure 11: biplot of word attributes and difficulty categories

Conclusion

- As you can see from the Figure 11a :
  - Can know that words that appear **frequently in life are more likely to achieve better results** in Wordle, The more we use it, the more familiar we get with it.
  - **The difficulty of the first letter of a word directly determines the difficulty of the word**, which is logical. It is always known that a good start determines the success or failure of things, and the same is true in Wordle.
  - On the other hand, **the repetition of letters in words tends to make the words more difficult**. This can be explained from the point of view of the game mechanics itself and information theory, because when trying, players tend to enter more reasonable words made up of different letters as a way to get more information, and the feedback of one attempt with or without that letter, players should in most cases not choose to repeat the input, thus leading to worse performance for words with duplicated letters.
- As you can see from the Figure 11b :
  - From the word lexical and emotional point of view, **words with positive meanings tend to tend to achieve slightly better results**, because the first time the player may think of is something good.

– Vocabulary with verbs and some other types of words, players will tend to score a little worse, probably because they are rarely used in life.

4 Apply The Model to Analyze EERIE

All the models have been built,In this section the model will be used to complete the problem for Problem 2: prediction for the word EERIE on March 1, 2023,and then in Problem 2 complete the difficulty classification of EERIE.

1. Firstly using the Extract Model to extract the useful attributes.

2. Secondly using the Neural Network Model to get the score distribution.

3. Thirdly using the Cluster-DT Model to Classify words into difficulty levels.

Table 8: the feature attributes of EERIE

Attribute	Values	Attribute	Values
w0	4	adv	0
w1	4	n	0
w2	17	v	0
w3	8	neg	1
w4	4	pos	0
cf	1	compound	0
adj	1	frequency	2.527×10^{-6}

The useful attributes of EERIE analyzed by the model are Tabel 8:

To facilitate subsequent use, we note the attriburites of EERIE as a_{eerie}.

$$\begin{aligned} a = &\quad (w_0, w_1, w_2, w_3, w_4, cf, adj, adv, n, v, neg, pos, compound, frequency) \\ a_{eerie} = &\quad (4,4,17,8,4,1,1,0,0,0,1,0,0,2.527\times10^{-6}) \end{aligned} \tag{17}$$

By pre-processing a_{eerie}, the attributes are then fed into the the neural network model to Predicted score distribution d_{eerie}:

$$d_{eerie} = (0.05589584, 5.9055951, 22.61680411, 35.7167871, 23.75443074, 9.73908243, 2.07853881)$$

By rounding the raw output, percentages of $(1,2,3,4,5,6,X)$ can be obtained,Table 9.

Table 9: neural network Predicted Distribution of EERIE

	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (X)
network model	0	6	23	36	24	9	2

So at the end the Cluster-DT model is used to classify the difficulty level using the predicted distribution results d_{eerie} of the neural network.Finally, the difficulty classification level of the words can be classified as:**Average**, Figure 12.

Using the models together, the distribution of player scores and word difficulty levels can be derived for any word that satisfies the Wordle rule word.

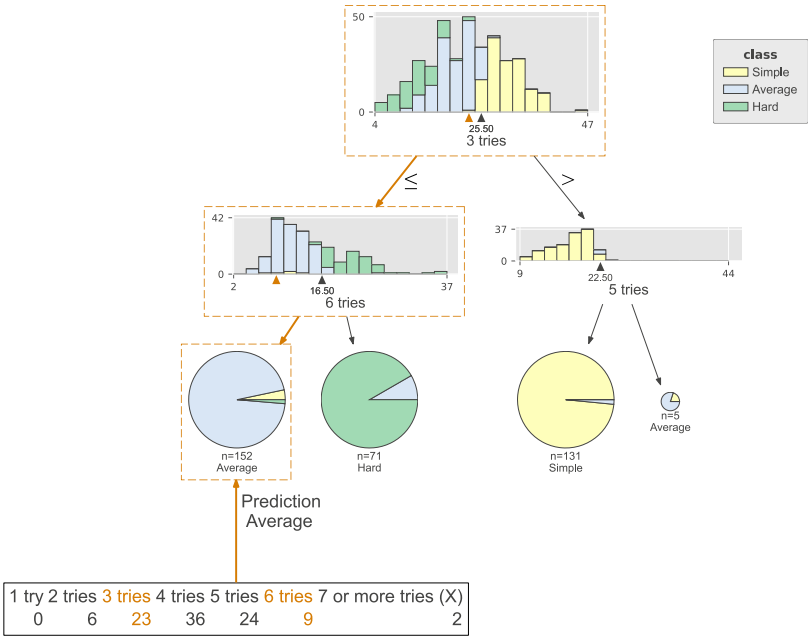


Figure 12: Difficulty classification of "ERRIE" by Cluster-DT model

5 Interesting Data Set Features

Further analysis of the characteristics of the original data set reveals some interesting features:

1. The number of normal players, the number of difficult players, the number of overall players, and the number of difficult players as a percentage of the overall players all show the same trend, with a rapid increase, a maximum value, and then a slow decrease, which is related to the fact that people's enthusiasm for new things will always be slowly consumed.

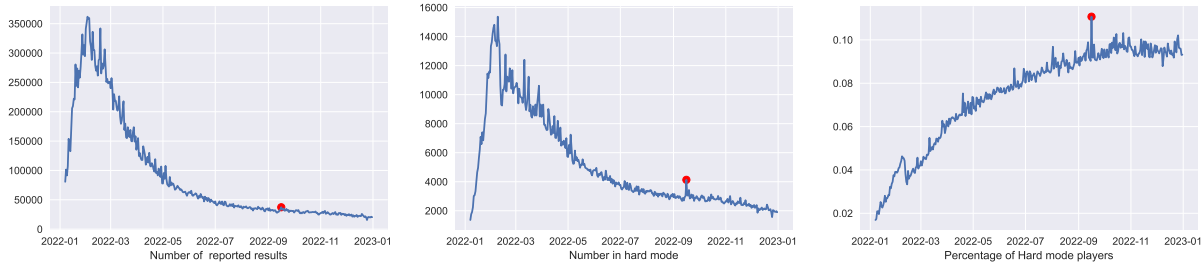


Figure 13: Changes in the number of players

In particular, on 2022-09-16, the proportion of players in difficult mode suddenly increased.

2. Further, we're more interested in knowing if the player growth is related to the workday? So we analyzed the change in the number of players between that day and yesterday.Find the proportional relationship between player growth and week. It can be found that players increase generally on weekdays, especially on Mondays, and the percentage of increase on Thursdays has weakened. The number of players decreases on weekends, and by checking the relevant

Wordle's communication forum, it is confirmed that players are more willing to play and communicate with each other at the same time.

Obviously the relationship between decrease and week is exactly the opposite of the above result, The relationship between growth and week can be seen in Table10.

Table 10: Percentage of weeks when players increased

Week	Proportional
Sunday	0.132184
Monday	0.183908
Tuesday	0.160920
Wednesday	0.166667
Thursday	0.097701
Friday	0.172414
Saturday	0.086207

3. Finally, we also analyzed the attribute characteristics of the words and found that among the lexical properties of words, nouns were the most numerous, followed by adjectives, while in terms of word meaning, the most neutral words were given, accounting for 84%, and the negative words and positive words accounted for very few words, and the proportions of both were roughly equal, indicating that the lexical database in the game was more evenly distributed and in line with the reality In the case of the game, there are more nouns and no leading meaningful colors included, and the word meanings are mainly neutral words.

Finally, we also counted the frequency of letter occurrences in words in the whole datasets and made the following word cloud map, Figure 14.

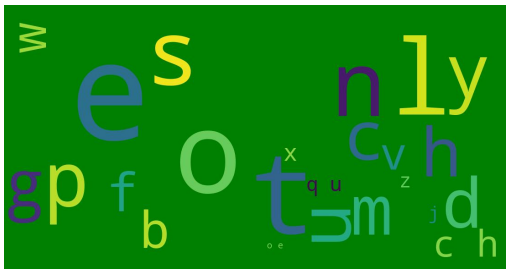


Figure 14: Alphabetic word cloud map

It is also possible to compare the relationship between the frequency of Wordle solution word letters and the frequency of real words, Table 11.

6 Strengths and Weaknesses

6.1 Strengths

- **Better portrayal of number variation:** Our improved logistic regression hysteresis growth model inherits the variation pattern of the original model and introduces a coefficient of variation to further portray the variation in the number of players.

Table 11: Letter frequency in Wordle game and real words

word	Wordle's Letter frequency	Letter frequency
e	10.250696	12.702
a	8.857939	8.167
r	7.465181	5.987
o	7.409471	7.507
t	7.242340	9.056
l	6.239554	4.025
i	5.682451	6.966
s	4.902507	6.327
...	...	...

- **Comprehensive evaluation of the difficulty of words:** We quantified whether the words are easy to guess based on the existing difficulty distribution using the comprehensive evaluation method.
- **Better characterization of word properties:** We studied the properties of the words themselves through the available data, and more comprehensively characterized the properties of a word.
- **Improved model design:** We use neural networks to establish the mapping between word attributes and their difficulty distribution, and use clustering models to classify the attributes of words with different difficulty levels, and finally output the classification through decision trees.

6.2 Weaknesses

- **Poor results with extreme changes in the number of people:** the model portraying the number of people has certain regular characteristics based on changes, and if there are extreme changes in the number of people it will lead to inaccurate model portrayal.
- **Possible overfitting:** Since there are a large number of parameters in models such as God's General Network and Decision Trees, and the word attributes as input features are filtered by our existing data, the model may be overfitted.
- **Model generalization ability is unknown:** Semantic analysis is impaired by the variability of the English language and the ability to use numerous word combinations to convey the same idea. Some keywords may be noticed and without more data for validation, the practical application capability of the model is unknown.

Letter

To: Puzzle Editor of the New York Times
From: Team 2318226
Date: February 20st, 2023
Subject: Modeling and suggestions for Wordle games

All the problems, we have built a corresponding model to make them well solved, the following will introduce the characteristics of our model one by one while sharing some of the problems found in the analysis of the problem, and finally give some suggestions about the development of Wordle.

Model 1: Logistic Growth Model

The most important thing for Wordle, is to have an accurate picture of the number of players in order to better determine the appropriate direction of development.As you can see from the image, Wordle went into flames as soon as it was introduced, But it is always impossible to grow indefinitely, the growth process that lasted for about 32 daysthen the popularity has weakened.

To better simulate this process, we avoid using time series models and some machine learning algorithms. We choose to use the differential equation model: Logistic Growth Model.Taking realistic resistance to growth into account to achieve a more realistic model, The final projection of the number of players for March 1, 2023 is completed and a reasonable range of intervals is given. I think this model in this problem is successful.

Model 2: Attribute Extraction Model

Consider from the perspective of game mechanics, a given word, then which of its attributes affect the distribution of the player's score in the final hard mode.This problem is interesting and equally it is crucial, because the useful feature extraction model built for this problem is the foundation for the subsequent models.

It is not necessary to define overly complex d attributes in this model.By reviewing the references, we defined some basic word properties such as word sentiment, word lexicality, whether the word has repeated letters, frequency of the word in real life etc. But the distribution of scores is an array of 7 dimensions, Therefore it is more difficult to analyze the effect of attributes. We then decided to evaluate the distribution of the scores together to derive a composite value by using TOPSIS. Then we use the hypothesis testing method to check whether the various kinds of words have an effect on the score or not.

Using the model we derived some attributes that affect the score distribution such as Frequency and presence of repeated letters.

Model 3: Neural network predictive model

The requirements of the subject matter are further enhanced. By using the attributes of the words as input to the model and then outputting the distribution of the scores of the words, unlike general NLP problems, accomplishing this goal is not simply, but it's not hard to beat us.

The first thought of statistical theory and general machine learning algorithms are not enough to solve this problem. This is a multiple-input, multiple-output modeling problem, Neural network models, as the main method in the field of NLP, become our first choice to accomplish this problem(MIMO). The data sample of our problem is not very large, and the neural network is easily overfitted. So we avoid using complex network models and choose the base model (MLP) instead

of RNN,CNN.

Using the useful word attributes extracted from model 2, our neural network model was trained by cross-validation.Acceptable results on training set, test machine and overall.We also completed the prediction of the distribution of the scores of the word EERIE. The details of the prediction process and the results are shown in detail in main article.

Model 4: Cluster-DT Model

Classifying the difficulty of any word is our last model. Again this model is important. This model is our most complex model, but logical. We named it as Cluster-DT Model. It combines the use of model 2 and model 3.

For any given word that also meets the rules of the Wordle game, the distribution of scores can be output using the previous model.We built a clustering algorithm using an improved decision tree model, using a combination of supervised and unsupervised learning.Not only can you classify the attached words, but you can also output classification rules to facilitate the classification of any given word, such as EERIE Classified as Difficulty Level: Average.

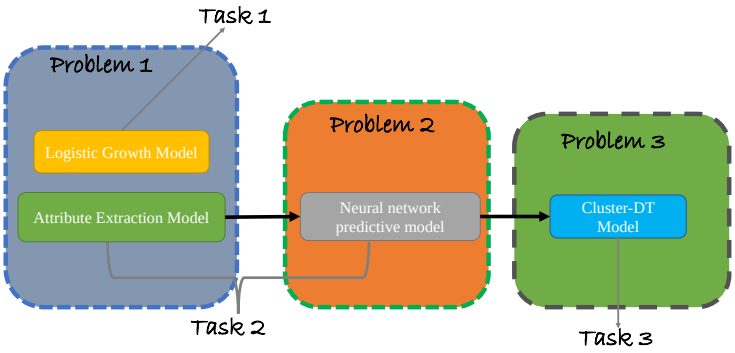


Figure 15: Relationship between different models

These are the four models we have built, Figure 15. Finally we share some interesting findings in the modeling process and give some tips for game development:

1. From model 1, we analyze that the game has entered a stable period at the moment and the heat has decayed. Make some improvements to the game to bring Wordle back into the public eye, Otherwise, the popularity may be further down. For example, add multiplayer matchmaking mode or friend ranking, etc.

2. At the same time, we note that on weekends, the number of players may be weakened due to the tiredness of working days or players going to play other games, so some new game activities are added on weekends to attract players in their free time. We believe that this will be an effective way to increase the number of players.

We hope that our work and models is helpful to you. we sincerely wishes Wordle games will develop better and be loved by more people.

References

[1] Meaning, Jianwei Gao, Tianhui Zhu, Jiaming Zhu (2019). Logistic regression-based prediction of public transportation mobile payment user volume. *Journal of Jiaozuo University*, 33(03), 84-87.

[2] *Wordle frequency of words*, 2022, from `https://github.com/amuellerastro/3b1b_wordle/blob/master/_2022/wordle/data/freq_map.json`

[3] Lan Yuexin,Liu Bingyue,Zhang Peng(2017). Research on the Dynamic Prediction Model of Online Public Opinion Fever Oriented to Big Data.*Intelligence Magazine*, 36(06):105-110+147.

[4] Winsor,C.P. (1932) A Comparison of Certain Symmetrical Growth Curves *Procedings of the Washington Academy of Science*, **38**,1-59.

[5] Theil,H. (1969) A Multinomial extension of the linear logit model. *International Economic Review*, **10**,251-259.

[6] LI Yu, YANG Yating, LI Xiao, MI Chenggang, DONG Rui.(2019) Research on neural network language model for the Chinese-to-Uyghur machine translation.*Journal of Xiamen University(Natural Science)*.2019,58(02):189-194.

[7] Charoenporn T, Kruengkrai C, Theeramunkong T, Sornlertlamvanich V .(2007)An EM-based approach for mining word senses from corpora. *IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS*, 4, 775-782.

[8] Varnadore AE, Roberts AE, McKinney WM.(1997) Modulations in cerebral hemodynamics under three response requirements while solving language-based problems: A transcranial Doppler study. *NEUROPSYCHOLOGIA*,35(9), 1209-1214.