Problem Chosen	2023	Team Control Number
C	MCM/ICM	2300358
	Summary Sheet	

Funny Wordle, Math Hidden

Abstract

Wordle, a free word-guessing game, is taking the world by storm. It's been bought for millions of dollars and has more than 2 million players worldwide.

Question C is an interesting question. For the first question, we first used **grey correlation analysis** to analyze the relationship between the number of attempts and the score. The score will be explained below. For the first question of the first question, our team used ninth-order polynomial regression prediction and exponential regression prediction, and found that the effect was not good. The reason is that polynomial regression can have a good fit at the given points, but we cannot judge the monotonicity of the function outside the points, and it is difficult to study it. Results of polynomial fitting We give results in the discussion section of problem one, note that they are unreasonable. Then we used exponential regression, taking into account the herd effect, the game was hot at the beginning, and then the number of people dropped rapidly. The result of exponential regression was very bad in the first half, but the effect was very good in the last half, but the predicted result was 7013, which was very inconsistent, possibly because of the explosive monotony of the index[1]. After that, we also tried grey prediction, and the prediction accuracy was not high, which was explained later. Finally, we choose **Xgboost** and **Lightgbm** machine learning, and get good simulation regression prediction results. We predict that the number of result reports on March 1, 2023 is 20626-20623 (Lightgbm) and 20539-20540 (XgBoost), which we think is reasonable. The problem1.2 requires studying the word characteristics. For this question, we use the grey correlation model and chi-square test to analyze whether difficult mode showed significant difference with normal model.

In response to the second question, we use **ARIMA model** to make multivariate time series predictions and get good results. We believe that the uncertainty factor of our model is **ARIMA model** and we predict that the distribution of the word EERIE on March 1, 2023. The result will be seen in Task 2.

Thirdly, we use decision tree machines learning regression and get surprising results. According to the classification of parts of speech, nouns are much easier to guess than other words, and there is no obvious difference in other words, which may be because we prefer to use nouns. There is a big difference in part of speech, but it is not linear. That doesn't mean the lower the frequency, the harder it is to guess. Through our model, we believe that the difficulty of the word "EERIE" is relatively difficult . We believe that our model will be more accurate when the data is larger or more accurate. However, we believe that the accuracy of our model is acceptable at present.

In response to this last problem, we find an interesting fact, the first is that the Date and Number of reported results do not conform to the Poisson distribution, and the other is that in percentages, 5 tries to conform to the normal distribution while the others do not.

We attached the memo to the New York Times at the end.

Key Words	Prediction	Xgboost	Lightgbm	LSTM	Grey Correlation Analysis	The Characteristics of Word

[1] Michael Rosander, Oskar Eriksson,
Conformity on the Internet – The role of task difficulty and gender differences, Computers in Human Behavior, Volume 28, Issue 5,2012,Pages 1587-1595,ISSN 0747-5632,https://doi.org/10.1016/j.chb.2012.03.023.

Contents

Introduction

1.1 Background

The New York Times has a popular puzzle called Wordle, which the player needs to guess a five-letter word six times or less, each guess gets feedback, and the color of the title changes after the player submits his word. Yellow tiles indicate words misplaced but present, green tiles indicate that the letter is present in the guess word and in the correct position, and gray indicates that the letter is not present in the word. Players cannot guess about words that are not recognized by the match.

The game is divided into normal mode and difficulty mode, where the difficulty mode requires the player to continue in subsequent speculation if they find the correct letter.

MCM generated a daily results file from January 7 to December 31,2022 including the date, race number, words of the day, the number of scores reporting the day, and the number of players in difficult mode. The document also includes the percentage of people guessing the word in one, two, three, four, five, and six times.

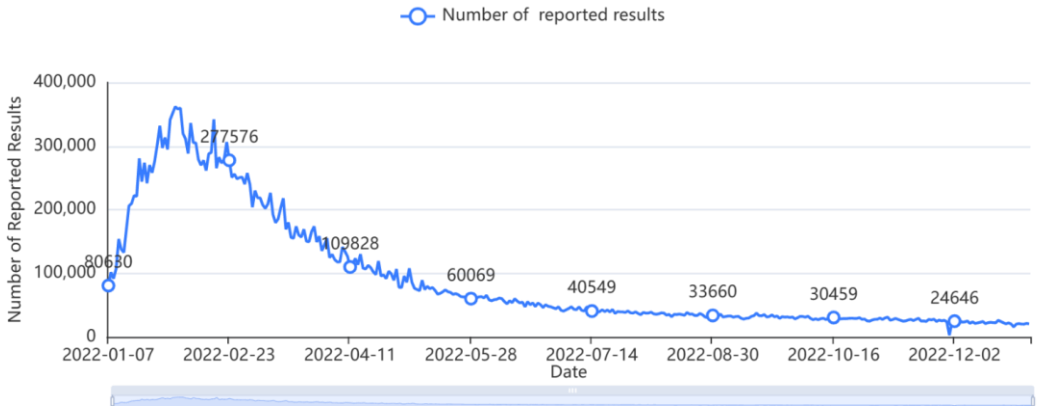


Figure 01 Date-Number of Reported Results

1.2 Problem Statement and Analysis

A known daily results file from January 7 to December 31,2022 including the date, contest number, the word of the day, the number of scores reporting on the day, and the number of players using difficult mode. The document also includes the percentage of guessing the word or failing to solve puzzles in one, two, three, four, five, and six times including X. Asking to address the following issues.

For Problem 1, problem 1.1 requires building a model to explain the daily variation in the number of reported results and uses a model to establish a prediction interval for the number of results reported as of March 1,2023. We use Polynomial Fitting, Exponential Function Fitting, Grey Prediction, Xgboost, and Lightgbm. The result shows that machine learning results are reasonable. The problem1.2 requires studying the word characteristics. For this question, we use the grey correlation model and chi-square test to analyze whether difficult mode showed significant difference with normal model.

Problem 2 requires predicting the relevant percentage of a future day, and considering the uncertainties of our model and predictions, and predicting the word EERIE of 1 March 2023, asking our confidence in the results of the model. We use the LSTM to analyze.

Problem 3 requires the development and summary of a model, to classify the problem-solving vocabulary according to the difficulty, and to find out the attributes of a given word related to each

classification. We use random forest regression and determine the difficulty of the word characteristics. The established random forest classification model is applied to the training and test data to obtain the classification evaluation results of the model.

Problem 4 we use SPSS to describe some other interesting features of the dataset.

2. General Assumptions and Justification

- It is reasonable that we choose word frequency and part of speech attributes;
- Since the data samples given by the authorities are relatively small, there are few training samples of machine learning. It can be considered that machines have not evolved close to human beings, rather than become more powerful through training.
- Our optimization of XGboost and Lightgbm machine learning models is reasonable and appropriate;
- The word frequency data we collected were reasonable, the 5-letter words we collected were accurate, and the part-of-speech tagging is accurate.
- Score and the number of try.

Number of Try	**Score**
1 try	7
2 tries	6
3 tries	5
4 tries	4
5 tries	3
6 tries	2
7 tries and X	1

3. Variable Description

Variable	*Description*
μ	The ratio of gray correlation between difficult mode and ordinary mode and score

4. Task 1

4.1 Conclusion

- We analyzed the relationship between Data and Number of Try using grey correlation analysis, which can roughly explain the distribution of reported results for Twitter on July 20, 2022. It probably means Data and Number of Try have some relations.

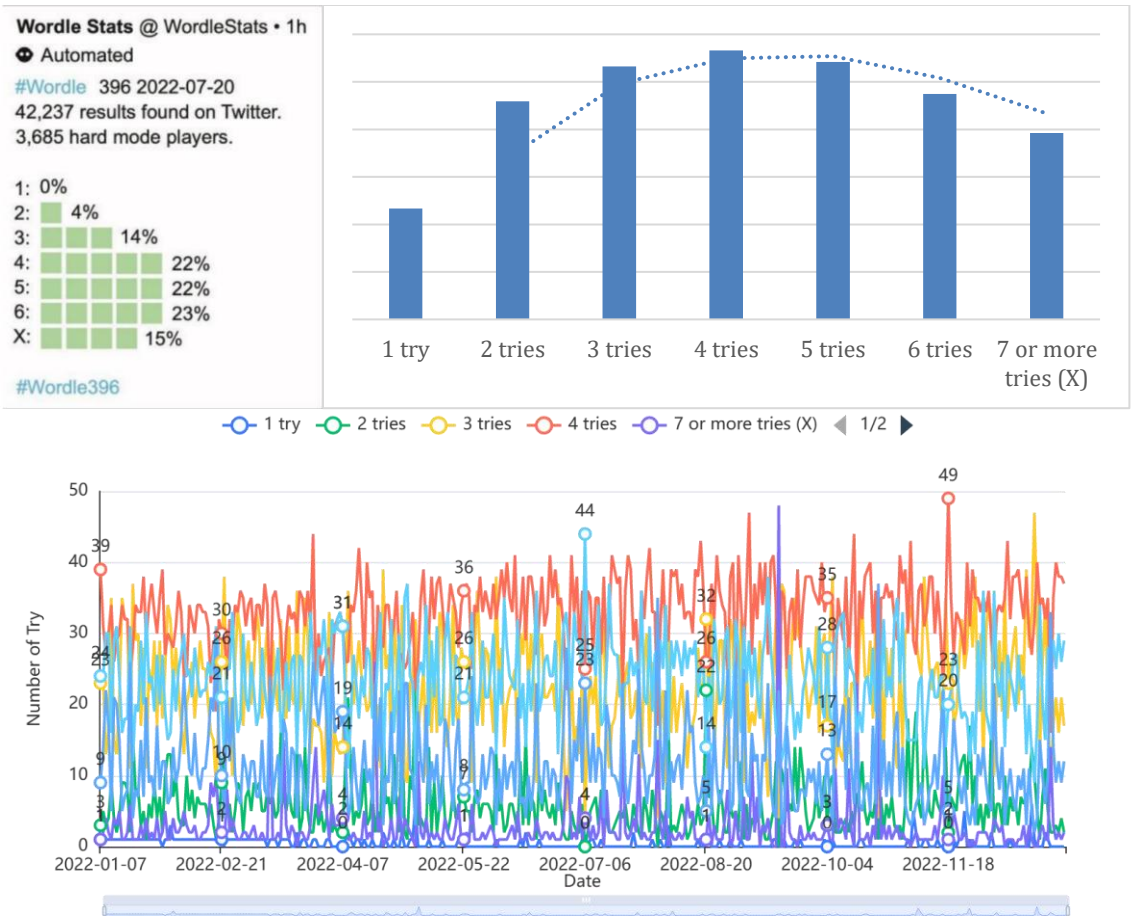


Figure 1 The Relation Between Data and Number of Try

CORRELATION RESULT

EVALUATION ITEM	Correlation degree	Ranking
4 TRIES	0.943	1
3 TRIES	0.938	2
5 TRIES	0.938	3
6 TRIES	0.921	4
2 TRIES	0.92	5
7 OR MORE TRIES (X)	0.899	6
1 TRY	0.849	7

- Number of reported results on 1, March, 2023 is **20626-20623**, the predicted result is **20626.8285329377** by Lightgbm Regression.
- Number of reported results on 1, March, 2023 is **20539-20540**, the predicted result is **20539.842** by XGboost Regression.
- We find that Characteristic or property of a certain word attribute little effect on the percentage of fractions performed in difficult mode, reasons as following:
  i. Difficult patterns are about 1.11 times more difficult than normal patterns.
  ii. The P value of pearson Chi-square test is 0.334, and there is no significant difference in Class and Number of reported results. The P value of pearson Chi-square test is 0.686, and there is no significant difference between Class and Number in hard mode data.

- We find that word frequency is strongly correlated with the number of guesses, but not linearly. Additionally, the more common words don't necessarily score higher.

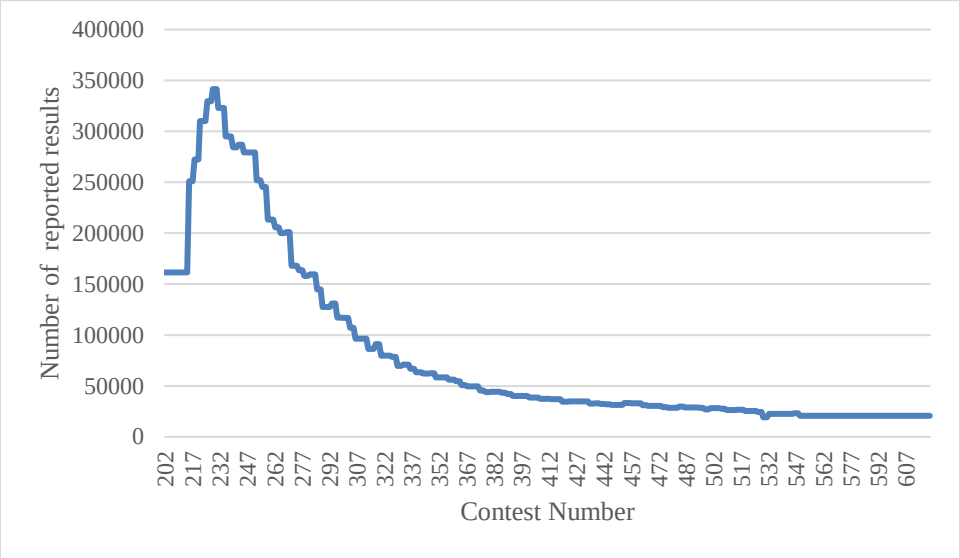


Figure 2 Prediction of Number of Reported Results by Lightgbm

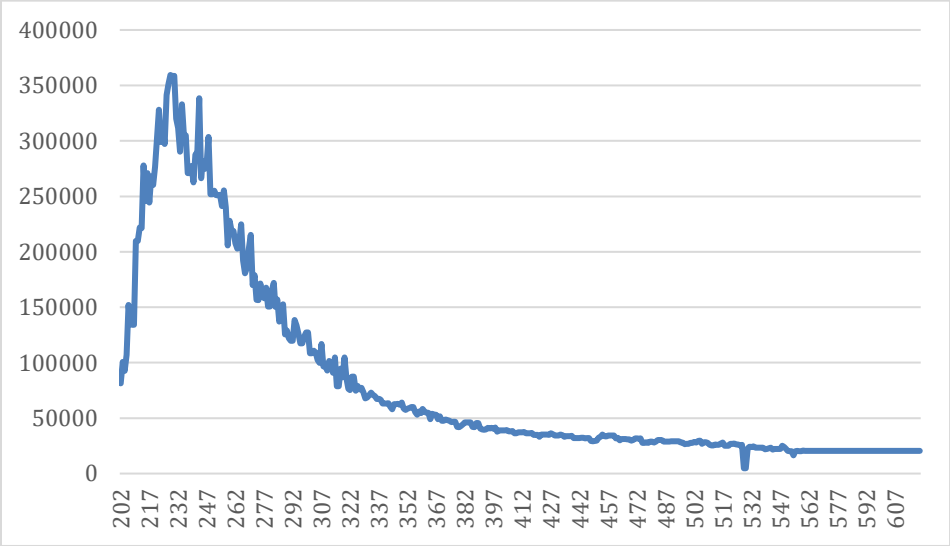
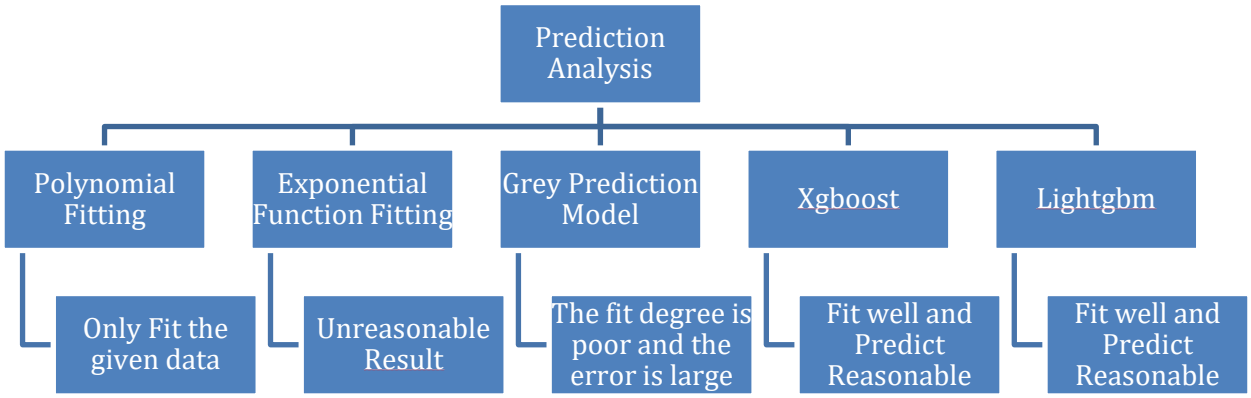


Figure 3 Prediction of Number of Reported Results by XGBoost

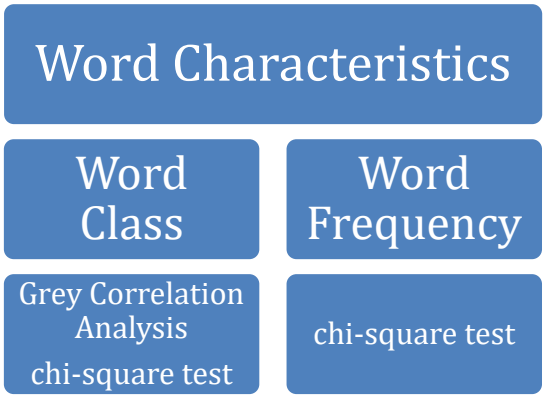
4.2 Problem Analysis

4.2.1 Prediction Analysis



The reason why we choose XGboost and Lightgbm will be stated in 4.4 Discussion Part.

4.2.2 Characteristic of Word



Word Class	noun
adjective	noun plurality
adverb	Parallel Conjunction
Comparative adjective	Past Tense of Verb
determiner	Preposition
Foreign original word	Superlative adjective
gerund	Verb Past Participle
interjection	Verb Present Singular
Modal verb	Verb prototype

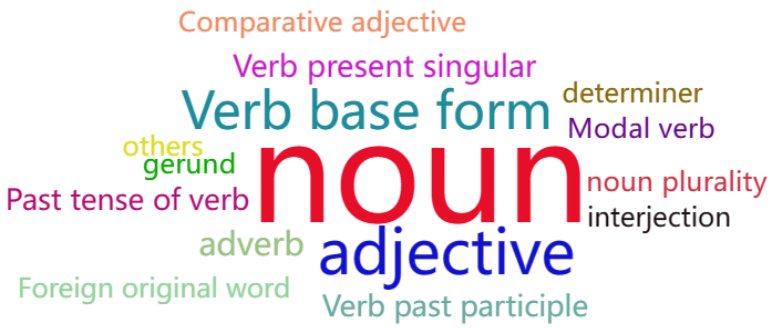


Figure 4 Word Fruency Cloud Diagram

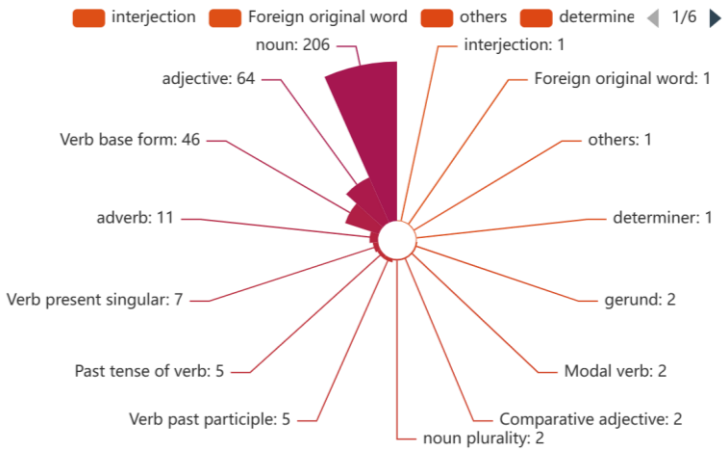


Figure 5 Word Furency Rose Map

4.3 Model Building

4.3.1 Prediction Model Buidling

Lightgbm Model

parameter name	parameter values
Training time	0.212s
Data cut	0.9
Data shuffle	yes
Cross validation	10
Base learner	gbdt
Number of base learners	100
Learning rate	0.1
L1 Regular term	0
L2 regular term	1
Sample trait sampling rate	1
Sampling rate of tree characteristics	1
Node splitting threshold	0
Minimum weight of the sample in the leaf node	0
Maximum depth of the tree	10
Minimum sample number of leaf nodes	10

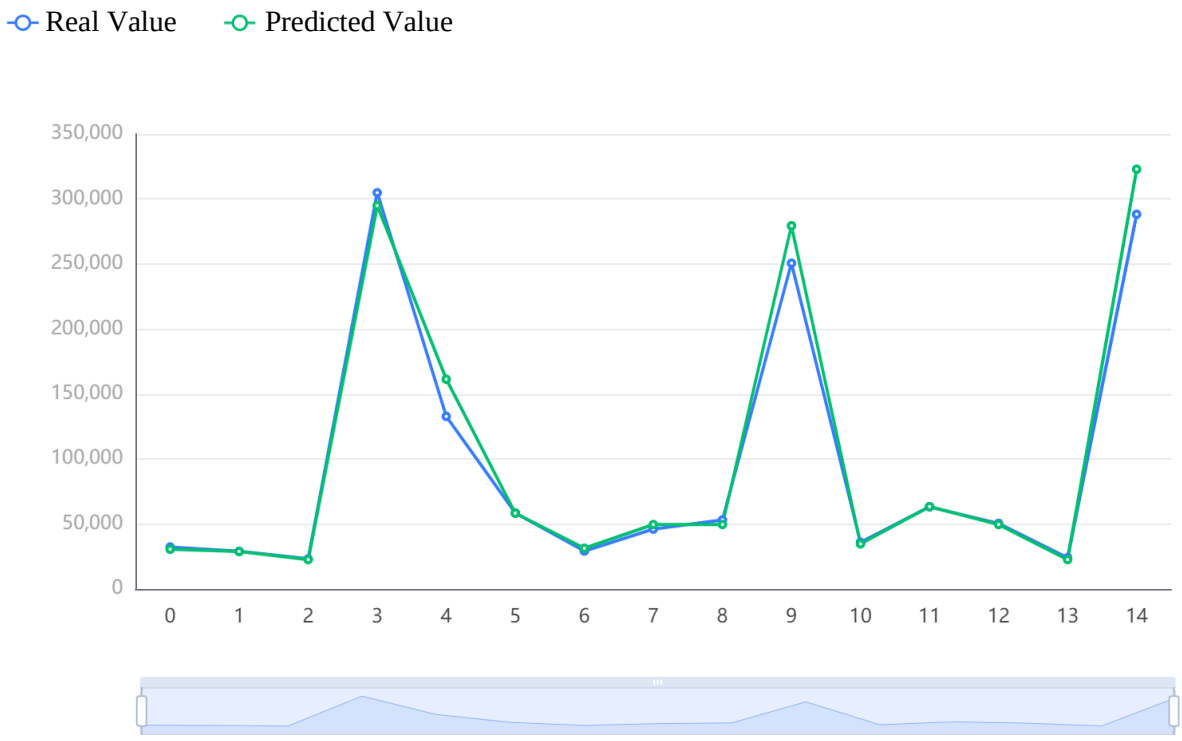


Figure 6 Test data prediction Fig

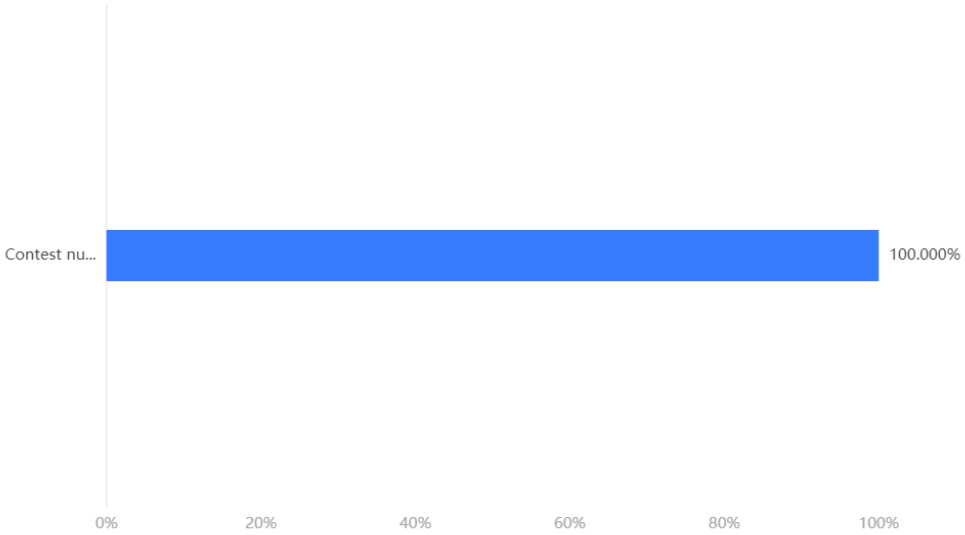


Figure 7 Characteristic importance

	MSE	RMSE	MAE	MAPE	R^2
Training set	182560768.337	13511.505	5645.762	5.615	0.977
Cross validation set	333263871.477	16104.582	7945.388	7.265	0.959
test set	195690134.304	13988.929	8259.073	6.659	0.974

evaluating indicator	evaluating indicator
MSE	148104275.46996272
RMSE	12169.810001391259
MAE	6600.744028575751
R^2	0.9813651722125035
MAPE	8.712185811616443

XGboost Model

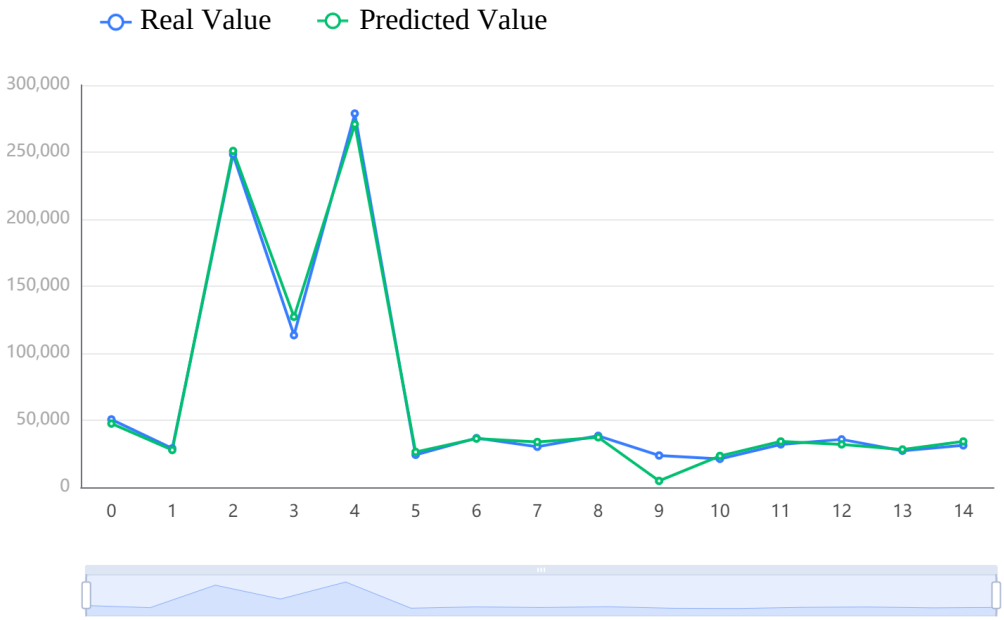


Figure 8 Test data prediction Fig

Model parameter

parameter name	parameter values
Training time	1.162s
Data cut	0.9
Data shuffle	yes
Cross validation	10
Base learner	gbtree
Number of base learners	100
Learning rate	0.1
L1 Regular term	0
L2 regular term	1
Sample trait sampling rate	1
Sampling rate of tree characteristics	1
Node feature sampling rate	1
Minimum weight of the sample in the leaf node	0
Maximum depth of the tree	10

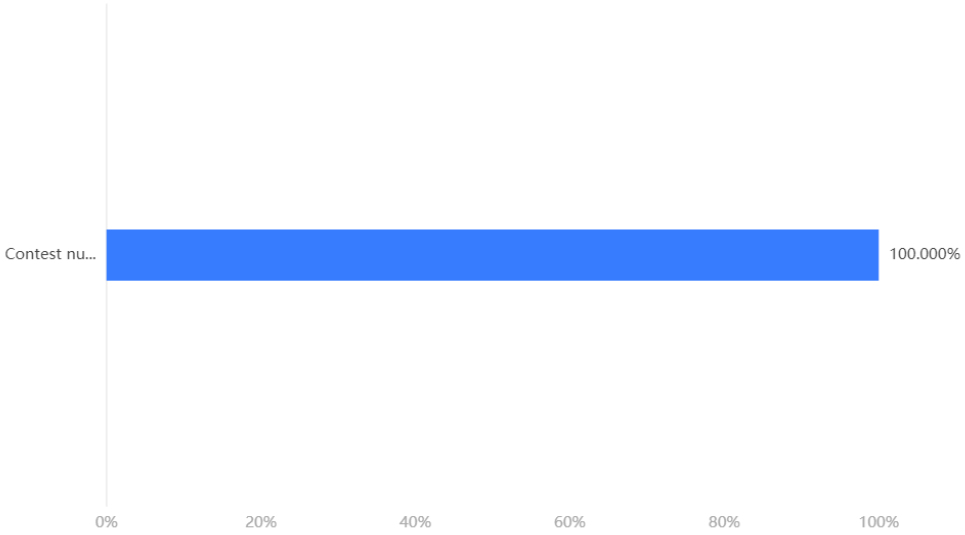


Figure 9 Characteristic importance

Model evaluation results	MSE	RMSE	MAE	MAPE	R^2
Training set	1118762.369	1057.716	758.608	1.377	1
Cross validation set	160396274.429	12180.939	6996.049	6.845	0.98
Test set	82254400.256	9069.421	5101.003	17.305	0.984

4.3.2 Characteristic of Word Model Building

Grey correlation analysis

Grey correlation coefficient:

	Correlation coefficient result Number of reported results	Number in hard mode
noun	0.7116261147638918	0.764168050438253
Verb base form	0.6423154323478485	0.6864833219410267

Verb present singular	0.5999618296672302	0.6390507793706818
noun	0.6855943559054615	0.737750177541348
adverb	0.6654390467200852	0.7167374949723485
noun	0.6514850990026824	0.7056693679658356
adjective	0.5877351992401186	0.6198922151026877
adverb	0.668739505657313	0.7160939290103272
Verb base form	0.6784701401145797	0.7346731276773213
adverb	0.6610904468809244	0.7145749229958386
adverb	0.6083177466219095	0.6482208474706941
adjective	0.6804214288433426	0.7387109771733704
adjective	0.6856750097731696	0.7488884758118337
adjective	0.6803243017863478	0.7356660205301765
adjective	0.653672014924121	0.7027122292685816

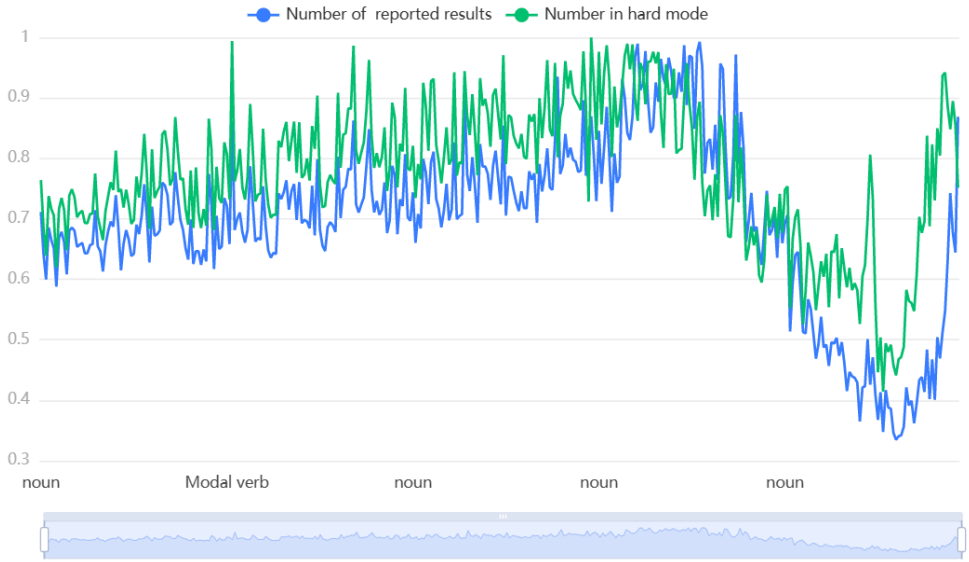


Figure 10 Correlation coefficient diagram

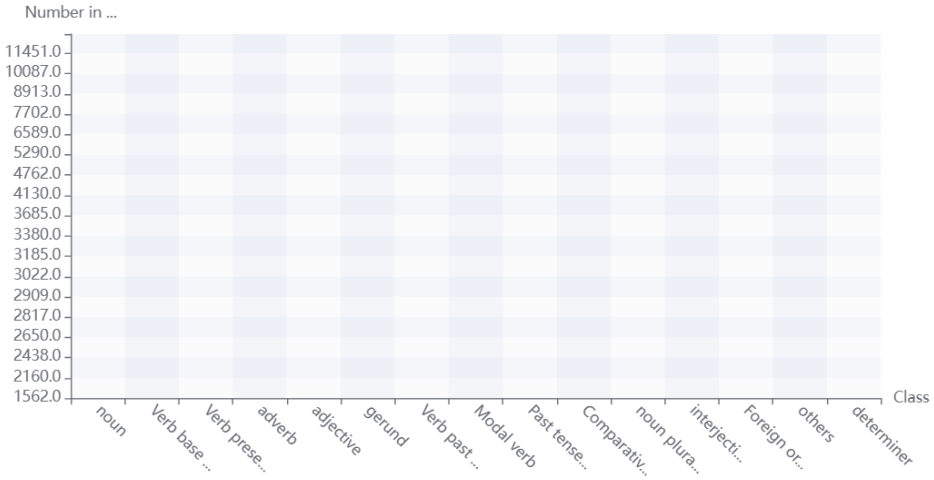
Correlation coefficient diagram

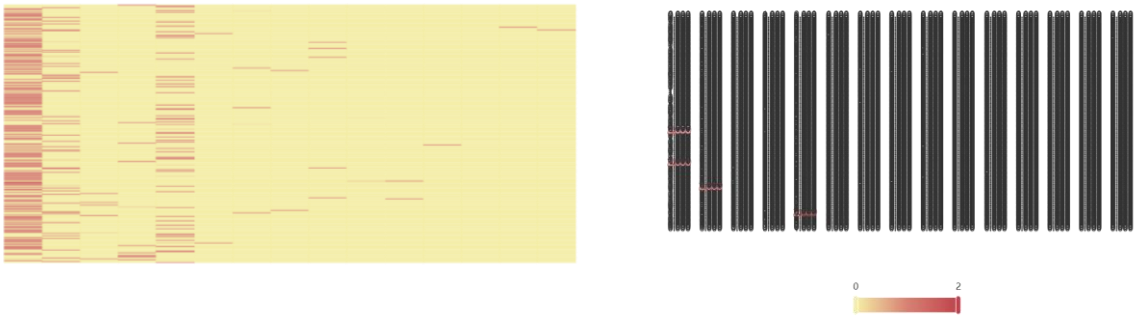
Correlation result		
Evaluation item	Correlation degree	Ranking
Number in hard mode	0.776	1
Number of reported results	0.698	2

$$\mu = \frac{the\ corelation\ in\ \text{Number in hard mode}}{the\ corelation\ in\ \text{Number of reported results}} = \frac{0.776}{0.698} \approx 1.11$$

chi-square test

Algorithm configuration:
Algorithm: Chi-square test analysis
Variable: variable X:{Class}; Variable Y:{Score, Number of reported results, Number in hard mode}
Parameter: Type :{P, e, a, r, s, o, n, card, square, check, check}
Class-Number in hard mode Thermal map





The relationship between word frequency and score (number of guesses)
Cramer's V value is 1.0, so the degree of difference between the number of times and the sequence number is a degree of extreme difference.

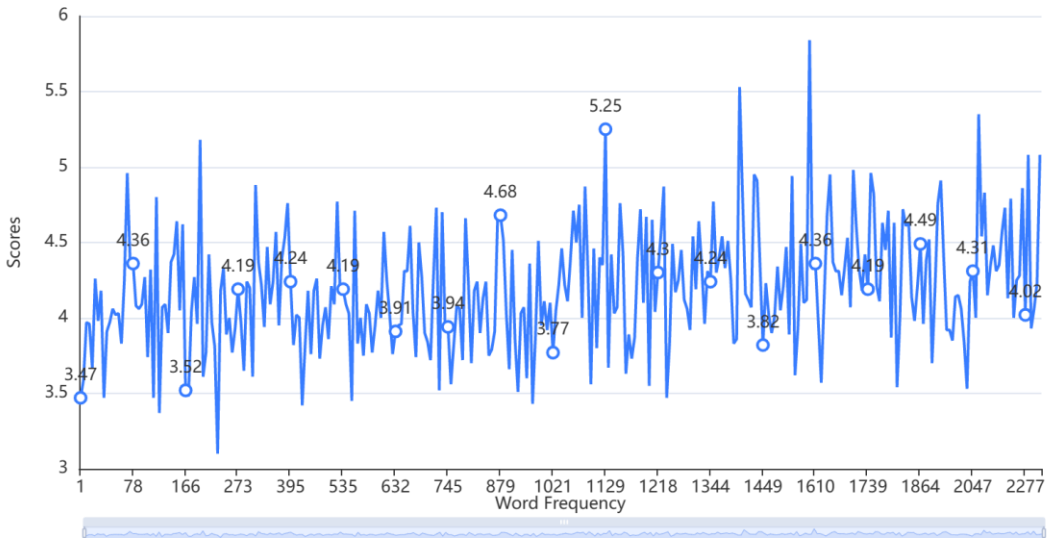


Figure 11 The Relation Between Given Word Frequency and Scores

4.4 Discussion

Our initial choice to use polynomial regression has many advantages, such as a wide range of functions that can accommodate it, polynomials that are basically suitable for a wide range of curvatures, and polynomials that provide the best approximation of the relationship between dependent and independent variables. However, the disadvantages of using polynomial regression are that they are too sensitive to outliers, and the existence of one or two outliers in the data can seriously affect the results of nonlinear analysis. In addition, unfortunately, fewer model validation tools are available to detect outliers in nonlinear regression than in linear regression. In this problem, it is difficult to find patterns to predict. We first use the grey prediction model and found that the error is very large. Then we considered using polynomial fitting. We import the data and used MATLAB to fit the ninth order polynomial function but find that the predicted value is obviously incorrect. In addition, it is also difficult to study the ninth order polynomial function.

In addition, exponential regression has a large error in the early stage, it has a good fit in the later stage, but the predicted result is not reasonable. We try to use machine learning including xgboost and LightGBM to predict, the predicted result is reasonable.

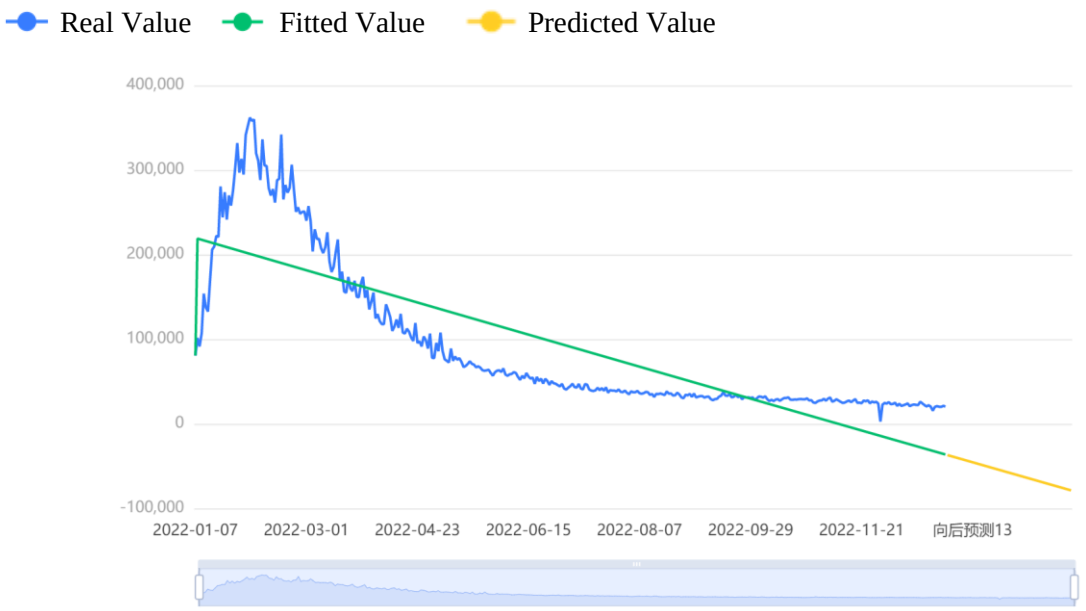
Here we give our work on Polynomial Prediction and Exponent Prediction:

Number of reported results on 1, March, 2023

Polynomial (Wrong)	**Result**
Middle Prediction	167181356.473081856
Lower Prediction	-1,435,532,622,654.47392
Upper Prediction	1,208,018,807,897.185553

Exponent (Wrong)	**Result**
Prediction	7013

Grey Prediction Model:



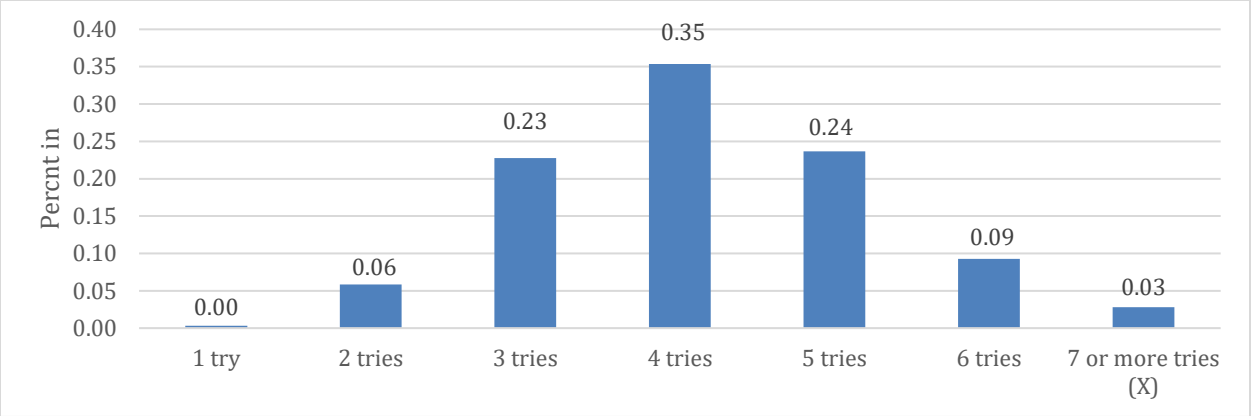
The average relative error of the model is 73.91%, which means that the model fitting effect is not good.

Why does part of speech not affect fractions but word frequency does?
Through the following analysis, after classifying the parts of speech and scores, we find that the guessing degree of nouns is much easier than that of other parts of speech, because we believe that the parts of speech will not affect the scores, while for word frequency, due to the influence of education, personal factors and other reasons, it will greatly affect the scores, but the relationship between word frequency and score is non-linear. The reason is that we do not use words solely based on word frequency, nor do we have the concept of active cognition of word frequency. In addition, in SAT and GRE, there are also some advanced words, which can be considered as low-frequency words in daily life. As a result, the word frequency of each person is different, and we only analyze groups rather than individuals. Due to the small number of samples, a large number of factors in the groups we analyzed were affected by individual factors, thus reflecting nonlinear differences.

5 Task 2

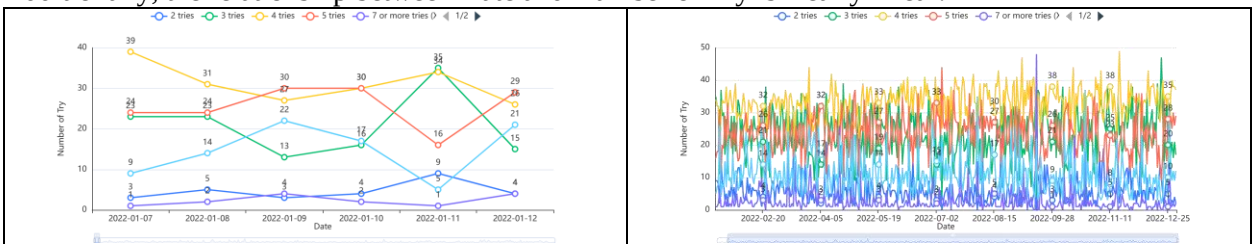
5.1 Conclusion

	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (X)
Predicted Result	0.297911	5.844011	22.72702	35.31373	23.6392	9.282584	2.805014
Normalization	0.002982	0.058493	0.227476	0.353457	0.236606	0.09291	0.028076



5.2 Problem Analysis

Our team used the ARIMA time series prediction model to predict the values of different numbers of Try. We predicted 60 units backward, which is the situation on March 1, 2023 as required by the question, and sorted the results into the table above.
Additionally, the relationship between Date and Number of Try is Nearly linear.



5.3 Model Building

The ARIMA(p, d, q) model is an extension of the ARMA(p, q) model. ARIMA(p, d, q) model can be expressed as:

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t = \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t$$

Where L is the Lag operator,$d \in Z, d > 0$

5.4 Discussion

We used the ARIMA time prediction model to predict each Number of tries separately, which may be related to each other, which we did not consider. The advantage of the ARIMA model is that the model is very simple, requiring only endogenous variables without the help of other exogenous variables. The disadvantage is that the time series data is required to be stable, or stable after differential differentiation, which can only capture linear relationship rather than nonlinear relationship in essence. But we think that because of the small amount of data, our use of the ARIMA model was reasonable and the results were accurate.

6 Task 3

6.1 Conclusion

We found it hard to guess except for nouns. And for EERIE, the level of difficulty is little easier than other word class but much higher than noun words.

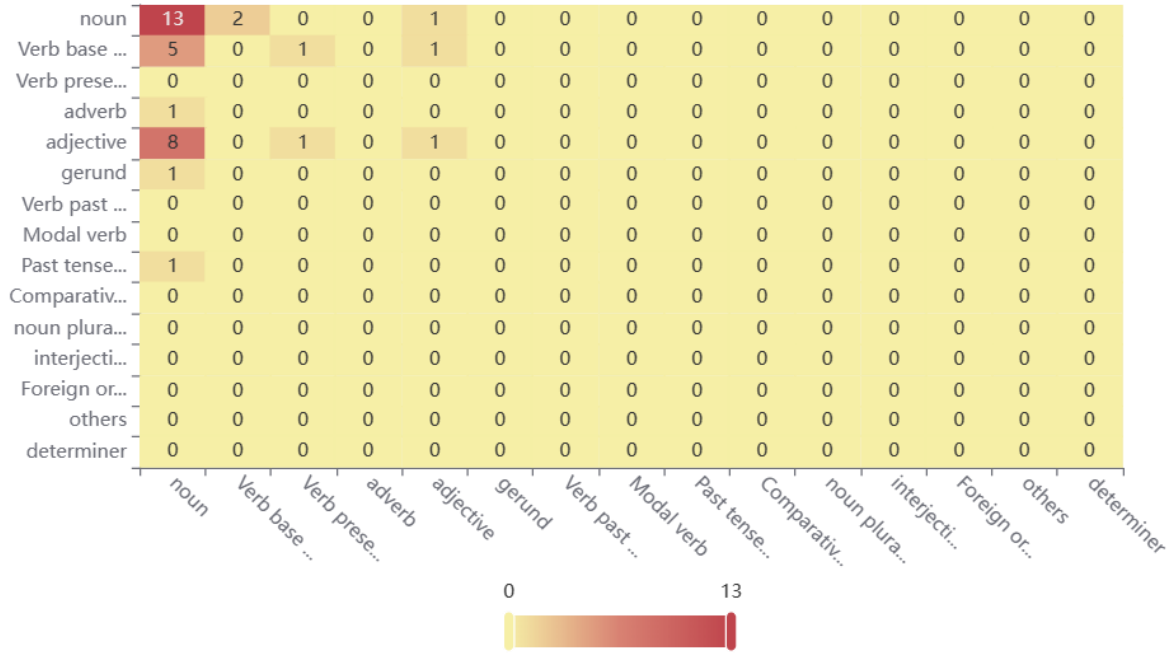


Figure 12 Thermal Map of Word Class

6.2 Problem Analysis

We use random forest classification to classify the data given by the authorities and calculate the importance of features through the established random forest. In addition, Our group believes that word frequency cannot be used as an indicator to analyze difficulty, because word frequency is related to education level, and we cannot guarantee that education level is an endogenous variable.

6.3 Model Building

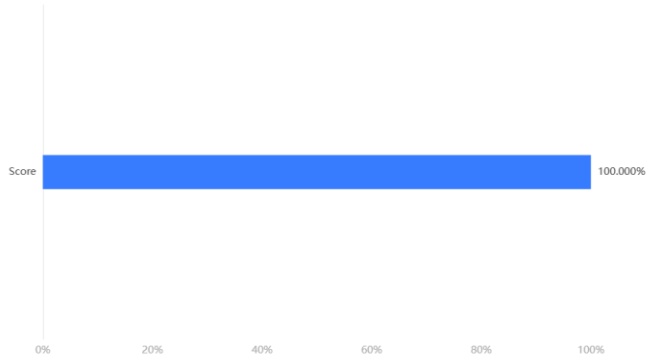
Algorithm: random forest classification
Variable: variable X:{Score}; Variable Y:{Class}

Model parameter

Parameter name	Parameter value
Training time	0.192s
Data segmentation	0.9
Data shuffle	True

Cross validation	False
Evaluation criteria for node splitting	gini
Number of decision trees	100
There are put back samples	true
Out-of-pocket data testing	false
The maximum proportion of features considered when dividing	auto
Minimum number of samples for internal node splitting	2
The minimum number of samples of leaf nodes	1
The minimum weight of the sample in the leaf node	0
The maximum depth of the tree	10
Maximum number of leaf nodes	50
Threshold of node partition impurity	0

Feature importance



Model evaluation result

	Accuracy	Recall rate	Accurate rate	F1
Training set	0.691	0.691	0.693	0.633
Test set	0.389	0.389	0.292	0.3

6.4 Discussion

Since the amount of data given is relatively small, and people usually prefer nouns to guesses, we believe that nouns are the easiest to guess, which is trivial. However, for other words, due to the small sample size, we cannot get specific classification results, that is to say, our model classification may be unreliable due to the small amount of data.

7 Task 4

7.1 Discussion

A. We found that the Date and Number of reported results does not match the Valbertine distribution。

Poisson test table

Results of Poisson distribution test					
Name	Option	Actual frequency	Expected frequency	X^2	P
Date	2022-01-07	80630.000	0.000	None	0.000***
	2022-01-08	101503.000	0.000		
	2022-01-09	91477.000	0.000		
	2022-01-10	107134.000	0.000		
	2022-01-11	153880.000	0.000		
	2022-01-12	137586.000	0.000		
	2022-01-13	132726.000	0.000		
	2022-01-14	169484.000	0.000		
	2022-01-15	205880.000	0.000		
	2022-01-16	209609.000	0.000		
	2022-01-17	222197.000	0.000		
	2022-01-18	220950.000	0.000		
	2022-01-19	280622.000	0.000		
	2022-01-20	243964.000	0.000		
	2022-01-21	273727.000	0.000		

Note: ***, ** and * represent the significance level of 1%, 5% and 10% respectively

B. The number of attempts does not satisfy the normal distribution except 5 tries.

The normal test is based on S-W test or K-S test to obtain the results:

- 1 try sample N < 5000, S-W test is adopted, and the significance P value is 0.000***, showing horizontal significance, rejecting the null hypothesis, so the data did not meet the normal distribution.
- 2 tries sample N < 5000, S-W test is adopted, and the significance P value is 0.000***, the horizontal significance is presented, rejecting the null hypothesis, so the data does not meet the normal distribution.
- 3 tries sample N < 5000, uses S-W test, and the significance P value is 0.034**, the horizontal significance is presented, rejecting the null hypothesis, so the data does not meet the normal distribution.
- 4 tries sample N < 5000, S-W test is adopted, the significance P value is 0.000***, the horizontal significance is presented, rejecting the null hypothesis, so the data does not meet the normal distribution.
- **5 tries sample N < 5000, tries sample N < 5000, uses S-W test, the significance P value is 0.151, the level is not significant, can not reject the null hypothesis, so the data meets the normal distribution.**
- 6 tries sample N < 5000, uses S-W test, the significance P value is 0.000***, the horizontal significance shows, rejects the null hypothesis, so the data does not meet the normal distribution.
- 7 or more tries (X) Sample N < 5000, S-W test is adopted, the significance P value is 0.000***, horizontal significance is presented, rejecting the null hypothesis, so the data does not meet the normal distribution.

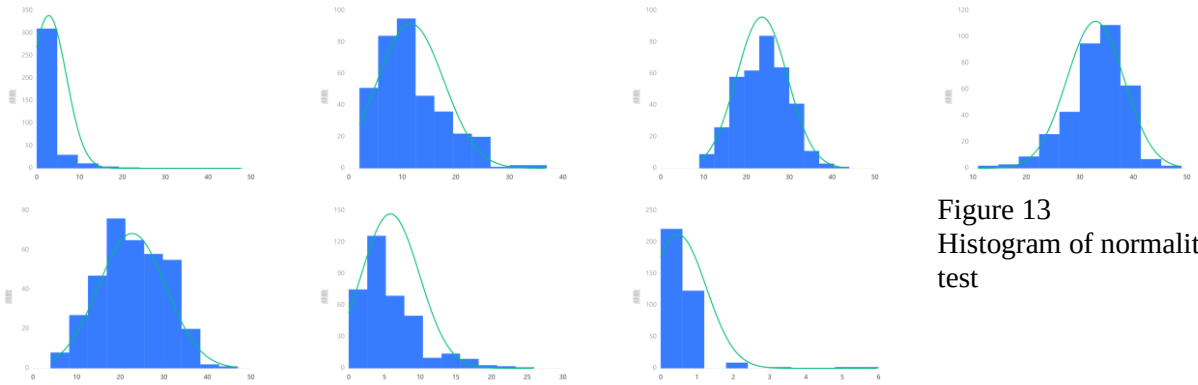


Figure 13
Histogram of normality test

- *Normality Test (2023/2/19 10:27:22)*

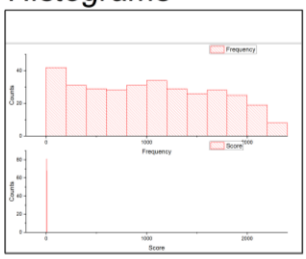
NormalityTest

Shapiro-Wilk

	DF	Statistic	p-value	Decision at level(5%)
Frequency	330	0.95956	6.54357E-8	Reject normality
Score	357	0.97593	1.13887E-5	Reject normality

Frequency: At the 0.05 level, the data was not significantly drawn from a normally distributed population
Score: At the 0.05 level, the data was not significantly drawn from a normally distributed population.

Histograms



We find that there is no normal distribution between word frequency and score

8. Memo

620 8th Ave
New York
United States

18 February 2023

Dear New York Times,
It's nice to have the opportunity to solve your Times crossword question. The puzzle is so interesting and challenging that it has attracted so many people to solve it and has been translated into more than 20 languages. There has also been a certain amount of discussion on Chinese social media, with many people also keen on the puzzle and some making videos to discuss the issue. Because it is easy to understand and practice English level.

However, the rules of this popular game are not complicated, which is one of the reasons for its popularity. It includes both normal mode and hard mode. The hard mode is more than easy and cannot be tried for an infinite number of times, and players can share the results on Twitter, including the number of successful attempts, which also increases its social and attracts more players. The idea of the game is simple and not simple, and the fact that we can't draw conclusions about it easily is what makes this question so fascinating. After playing a game or two we began to have the fun in the game. We were happy to analyze the data and draw some meaningful conclusions. Modeling this topic was interesting for us. I also want to thank the New York Times for coming up with such interesting games. You have also improved our interest in English learning and deepened our understanding and mastery of English spelling.

Now I'd like to introduce our understanding and explanation of the problem. This is the part we particularly want to highlight. They give us four questions, each of which is very interesting and valuable. Each question is about the puzzles. The question gives us a known document and asks us to analyze and solve the problem. We need to solve all the little questions in all the questions and for each question we have thought deeply, we have used the right models, so that we can reasonably explain the questions and come up with a satisfactory answer, We use Polynomial Fitting, Exponential Function Fitting, Grey Prediction, Xgboost, and Lightgbm., We use polynomial fitting, exponential function fitting, grey prediction, Xgboost, and Lightgbm., We use LSTM to analyze the second problem. For the third problem, we used random forest classification and other methods. Problem 4 we use SPSS to describe some of the interesting features of the data set. Each model is the result of careful thinking, and we find that they fit so well with the problem, which is why these models are used. In addition, for each problem, we have made a variety of charts to facilitate understanding. They are the presentation of our thoughts and the intuitive presentation of our data. We believe that these charts are helpful to show our results and facilitate readers' understanding. As for the writing of the paper, we were careful in every part and divided the problem into different steps, from the surface to the inside and from the depth to the shallow. We also obtained a lot of surprises, which also helped us to solve the problem better. In the process of writing the paper, we kept generating interesting new ideas, which also helped us to understand the problem better. Some of the difficulties also made us think for a long time, but the good news is that we have solved these problems well. The whole article is the result of our careful thinking and we have well organized and presented our ideas in the article. This is the result of the division of labor and thinking of our whole team. We also owe the result to our instructor. We believe this is a completely neat answer.

The results are as follows:

A. We analyzed the relationship between Data and Number of Try using grey correlation analysis, which can roughly explain the distribution of reported results for Twitter on July 20, 2022. It probably means Data and Number of Try have some relations.
B. Number of reported results on 1, March, 2023 is **20626-20623**, the predicted result is **20626.8285329377** by Lightgbm Regression.
C. Number of reported results on 1, March, 2023 is **20539-20540**, the predicted result is **20539.842** by XGboost Regression.
D. We find that Characteristic or property of a certain word attribute little effect on the percentage of fractions performed in difficult mode, reasons as following:
    a) Difficult patterns are about 1.11 times more difficult than normal patterns.
    b) The P value of pearson Chi-square test is 0.334, and there is no significant difference in Class and Number of reported results. The P value of pearson Chi-square test is 0.686, and there is no significant difference between Class and Number in hard mode data.
E. We find that word frequency is strongly correlated with the number of guesses, but not linearly. Additionally, the more common words don't necessarily score higher.
F. Our resume's model predicts percentage the word EERIE on March 1, 2023 as following table,

1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (X)
0.00	0.06	0.23	0.35	0.24	0.09	0.03

G. We found it hard to guess except for nouns. And for EERIE, the level of difficulty is little easier than other word class but much higher than noun words.
H. We found that the Date and Number of reported results does not match the Valbertine distribution。
I. The number of attempts does not satisfy the normal distribution except 5 tries.

We would be delighted if our results were valuable and appropriate to you. We would also appreciate hearing from you.

Yours sincerely

Team# 2300358

9. Advantages and Disadvantages

9.1 Advantages

- We used a variety of methods to predict the reported results on March 1, 2023, and chose machine learning through comparison and reasonable analysis, and proved that polynomial fitting prediction and exponential fitting prediction were unreasonable.
- We collected all 5-letter words through the corpus and obtained their word frequencies through the official website. By adding the word frequency attribute to the officially provided words, we get a more objective word frequency attribute for the percentage of reported hard mode.
- If we use the data provided by the government as training samples, the machine learning ability will not be enhanced, that is to say, we can better simulate human operation, because the sample number is small.
- We used more diagrams to make understanding the paper easier and more intuitive.

9.2 Disadvantages

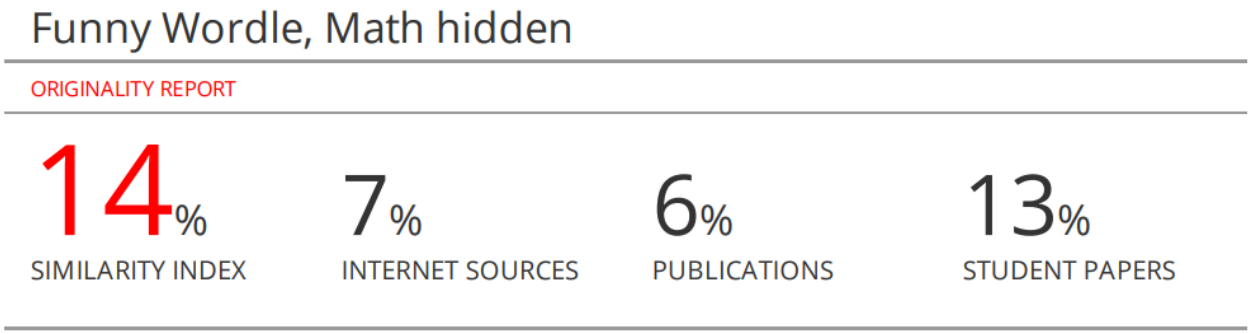
- The setting of machine learning simulation parameters may not be optimal and may have more accurate results.
- We cannot guarantee that the word frequencies we collect are up-to-date or accurate, and there may be errors in our assessment.
- Education level is also a factor worth considering. For example, the number of attempts between liberal arts students and science students is different, and those who like Sudoku may be different. However, this point can be considered as we study the whole.
- We can't guarantee that someone is hacking the game to improve their rate.

10. Reference

[1] Scientific Platform Serving for Statistics Professional 2021.SPSSPRO. (Version 1.0.11)[Online Application Software]. Retrieved from https://www.spsspro.com.
[2] Meng Q . LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 2018.
[3] Chen T , Guestrin C . XGBoost: A Scalable Tree Boosting System[J]. ACM, 2016.
[4] Zhou Zhihua. Machine Learning [M]. Tsinghua University Press, 2016.
[5] Liu Ruiyuan, Zhang Zhixia. Hypothesis testing of Binomial distribution and Poisson distribution discrimination [J]. Journal of Qinghai University (Natural Science Edition),2008,(01):44-47.2000.
[6] Zong Xuping, Yao Yulan. Rapid test of Statistical Distribution of Data using Q-Q graph and P-P graph [J]. Statistics and Decision, 2010(20):2.
[7] Mao Shisong, Wang Jinglong, Pu Xiaolong, etc Advanced Mathematical Statistics (Second Edition) [M] Beijing: Higher Education Press, 2006

Appendix

Duplicate checking:



Polynomial curve fitting prediction code：

```
function [fitresult, gof] = createFit(x, y)
%CREATEFIT(X,Y)
%  Create a fit.
%
% Data for 'model fitting prediction graph' fit:
%      X Input : x
%      Y Output: y
%  Output:
%      fitresult : a fit object representing the fit.
%      gof : structure with goodness-of fit info.
%
% Also see FIT, CFIT, SFIT.

% is generated automatically by MATLAB from 17-Feb-2023 18:47:11

```

```
%% Fit: 'Model fitting prediction graph '.
[xData, yData] = prepareCurveData( x, y );

% Set up fittype and options.
ft = fittype( 'poly9' );
opts = fitoptions( 'Method', 'LinearLeastSquares' );
opts.Robust = 'Bisquare';

% Fit model to data.
[fitresult, gof] = fit( xData, yData, ft, opts );

% Create a figure for the plots.
figure( 'Name', 'Polynomial Model Fitting Prediction Graph' );

% Plot fit with data.
subplot( 2, 1, 1 );
h = plot( fitresult, xData, yData);
legend( h, 'Number of  Reported Results vs. Contest Number', 'Model Fitting Prediction Graph', 'Location', 'NorthEast',  'Interpreter', 'none' );
% Label axes
xlabel( 'Contest Number', 'Interpreter', 'none' );
ylabel( 'Number of  Reported Results', 'Interpreter', 'none' );
grid on

% Plot residuals.
subplot( 2, 1, 2 );
h = plot( fitresult, xData, yData, 'residuals' );
legend( h, 'Model Fitting Prediction Graph- residuals', 'Zero Line', 'Location', 'NorthEast', 'Interpreter', 'none' );
% Label axes
xlabel( 'Contest Number', 'Interpreter', 'none' );
ylabel( 'Residuals', 'Interpreter', 'none' );
grid on
```