Research on Word Attributes and Word Meaning Difficulty Based on Time Series Analysis

Zhenbiao Miao *

College of Materials and Chemical Engineering, West Anhui University, Lu'an, China, 237012

* Corresponding Author Email: 18715674741@163.com

Abstract. In recent years, natural language processing has increasingly focused on revealing the relationship between word attributes and word meaning difficulty through data mining and statistical modeling to provide new perspectives for lexical cognition research. In this study, the daily result data from January 7 to December 31, 2022, classify word attributes into 30 categories of related attributes. A time-series model was built to derive the number of reported results on March 1, 2023, to be about 12,295 items. It was found that 29 attributes such as the commonness of words are not related to the difficulty pattern and the meaning of words are related to the difficulty pattern. Then corresponding to the data on the number of attempts to solve different subcategories, a gradient boosting tree (GBDT) [1] regression model was built to derive the correlation percentages under the number of attempts 1, 2, 3, 4, 5, 6, and X for players solving the EERIE word case as 1%, 4%, 18%, 28%, 28%, 16%, and 5%. In this paper, it is found that in solving EERIE words, players have a higher probability of solving the word under 4 and 5 attempts and a smaller probability of solving the word under 1 and 6 attempts.

Keywords: Time Series Analysis (ARMA), Gradient Boosted Tree (GBDT) Algorithm, Correlation Analysis.

1. Introduction

In the field of natural language processing and vocabulary learning, the study of word attributes [2] and word sense perception has been an important direction. With the development of data analysis and machine learning techniques in recent years, how to reveal the relationship between word perception and difficulty through large-scale data mining and statistical modeling has become one of the key issues for researchers.

For this study to solve the change in the number of reported results and predict the future distribution of word results, this paper utilizes time series models and gradient-boosted tree algorithms to construct a prediction system. In solving real-world problems, time series models and gradient-boosting tree algorithms play an important role in data analysis and prediction. Time series models are good at capturing the time dependence and trends of data and are suitable for financial analysis, sales forecasting, etc. They can process historical data and predict future trends. Gradient Boosting Tree algorithms, such as GBDT and XGBoost, are widely used in machine learning fields such as credit scoring and ad click prediction for their high accuracy and robustness, which can automatically process feature relationships and are robust to noise and outliers. Both methods are indispensable tools in data analysis and prediction.

This paper revolves around the original data, the proposed time series analysis ARIMA for forecasting, the data is given in the number of reports from January 7 to December 31, 2022, and other related data, the study predicts that the future data under similar conditions should be based on the original data using the ARIMA [3] model to first process the data from January 1 to March 1, to establish a relevant model, and through the model to predict the number of reports on March 1, 2023, It was also required to determine whether any attributes of words affected the percentage of reported scores played in difficult mode. Firstly, multiple attributes of words should be identified as data, then the data should be tested for normal distribution and processed through Spearman correlation analysis to find out whether any attributes of words affect the percentage of scores played in difficult mode. Next, the requirement was to predict the relevant percentage of players attempting to solve the word

at 1, 2, 3, 4, 5, 6, and X on March 1, 2023, and this paper first considered that the relevant percentages should be modeled for 1 attempt, 2 attempts, 3 attempts, 4 attempts, 5 attempts, 6 attempts, and 7 or more attempts, respectively. Based on the seven models established, the percentage of relevant attempts is determined using the relevant word attributes of EERIE.

2. Data source and preprocessing

The data in this article was obtained from https://www.comap.com/, to solve the problems related to Wordle's [4] predictions, some problems and errors were found in this article for the data from January 7 to December 31, 2022, in the given table, so before solving and analyzing the problems, this article preprocesses the data for the results of the popular puzzle game provided by the New York Times daily.

According to the wordle [5] rules, puzzles are words consisting of five letters, and for data consisting of four or six letters that do not conform to the rules, after relevant verification and analysis, the erroneous words should have originally consisted of five letters, and in this paper, we have chosen to make several corrections to the words that appeared to be erroneous. Words that incorrectly do not conform to the rules will be restored to their letter five. For December 16 and December 26, the word "rbrobe" was changed to "probe", and on April 29, "tash" was changed to "trash".

This paper compares and analyzes the number of people who reported their scores on that day and finds that the number of people who reported their scores on that day was between 20,000 and 350,000 people. For November 30th, the number of people reporting their scores was 2,569, which is a large error, so the paper decided to remove the November 30th data. The study analyzed the sum of the percentages of players who solved the puzzles in one to seven or more attempts and found that on March 27, the sum of the percentages of players who solved the puzzles in one to seven or more attempts was 126%, which is grossly disproportionate.

Finally, it was found that nearly half of the data in the table showed that the sum of the percentages of players who solved the puzzle in one to seven or more attempts was less than 100%. This paper analyzed this twice, once by removing any error data where the percentage was not 100%, and again by retaining small errors (e.g., between 98% and 102%) and removing only the serious errors on March 27th. However, after deleting all data with errors, a large error was made in predicting the number of reports for the first problem on March 1, 2023.3 This error was not a significant one. Due to the large amount of data with minor errors, deleting all the data severely affected the direction of the prediction, which finally resulted in a negative number of reports on March 1, 2023, as predicted. Therefore, the team of this paper decided to use the second discussed option: keep the minor error data and delete only the serious error data on March 27th.

3. Time Series Forecasting and Study of Lexical Difficulty Patterns

3.1. Forecasting the number of results on March 1

In this paper, through the time series model, the first January 1 to March 1 each prediction interval is divided into 7-14 days, every 10 days for a total of six forecasts. Take the first prediction from January 1 to January 10 as an example, while the ARIMA [6] model requires the sequence to meet the smoothness, view the ADF test table, first of all, according to the size of the t-value of the analysis ($p<0.05$), to analyze whether it can be significantly rejected the hypothesis that the sequence is not smooth. As shown in Table 1 ADF test table:

Table 1. ADF test table

variant	index	t	P	AIC	thresholds		
					1%	5%	10%
Number of	0	-4.247	0.001***	7941.052	-3.448	-2.869	-2.571
results re	1	-4.434	0.000***	7935.891	-3.448	-2.869	-2.571
posted	2	-8.077	0.000***	7916.647	-3.448	-2.869	-2.571

Note: ***, **, and * represent 1%, 5%, and 10% significance levels, respectively.

First of all this paper reports the results based on the variable Number: at the difference order of 0, the significance p-value of 0.001*** is significant at the level where the hypothesis is rejected and the series is a smooth time series. And on the first order difference, the significance P-value is 0.000***, which is significant at the level where the hypothesis is rejected that the series is a smooth time series. As shown in Figure 1 the number of reported results (0th order differences):

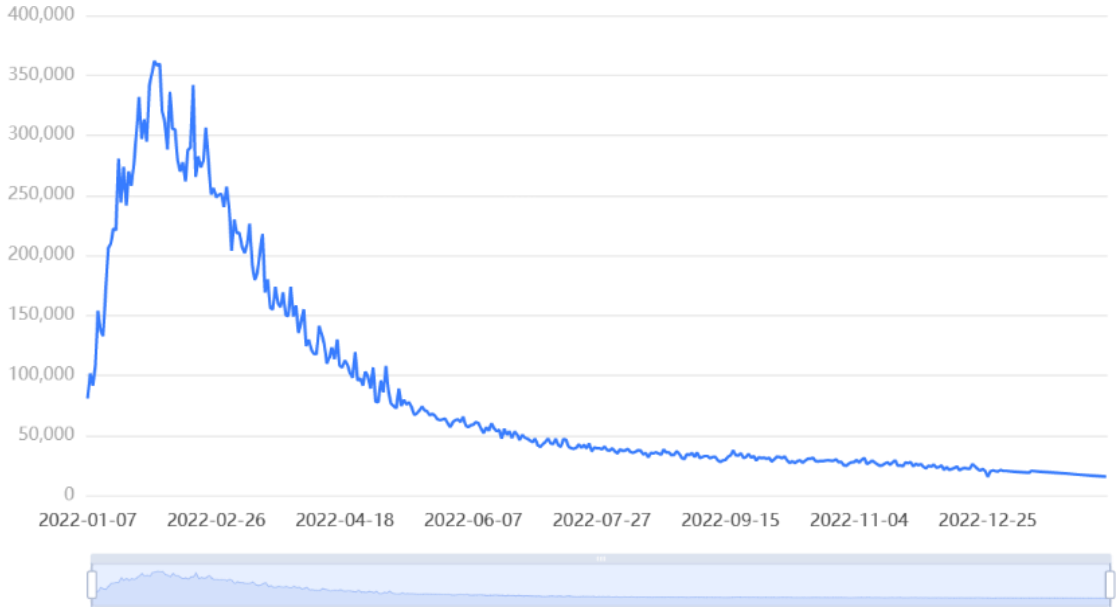

Figure 1. Number of results reported (0th order difference)

Table 2. ARIMA model (1,1,0) test table

Item	Symbols	Value
	Df Residuals	394
Sample size	N	397
Q-statistic	Q6(P value)	0.038(0.846)
	Q12(P value)	26.886(0.000***)
	Q18(P value)	57.955(0.000***)
	Q24(P value)	97.227(0.000***)
	Q30(P value)	119.024(0.000***)
Information Guidelines	AIC	8524.794
	BIC	8536.738
Goodness of fit	R^2	0.983

Note: ***, **, and * represent 1%, 5%, and 10% significance levels, respectively.

Shown by Table 2: ARIMA model (1,1,0) test table.Secondly, this paper knows the accuracy through the table of model parameters: the corresponding value of the information criterion AIC is 8524.794 and the corresponding value of the information criterion BIC is 8536.738, the square of the fit R is 0.983, which is very accurate and shows that the treatment using the model of time series analysis (ARIMA) is very accurate. Then the result of the above model is a test table of the ARIMA model (1,1,0), based on the variable: the number of reported results, from the results of the analysis

of the Q-statistic, we can get that: there is no significance on the level of Q6, and the hypothesis that the residuals of the model are a sequence of white noise can not be rejected, and the model's goodness-of-fit R^2 is 0.983, and the model has an excellent performance, and the model meets the requirements.

Finally, this paper divides the period from January 1 to March 1 into 10 days as an interval, first of all, we use the time series forecasting model for January 1 to January 10 to carry out the above-related series analysis and get the relevant number of people, and use the same operation as the first interval above for January 10 to January 20, January 20 to January 20, January 20 to January 30, February 1 to February 10, February 10 to February 20, February 20 to February 30, and the relevant remaining five intervals use the same operation as the first interval above to get the number of people on March 1 according to the time series forecasting model. January 10, February 10 to February 20, February 20 to February 30, and the remaining five intervals of interest using the same operations as the first interval above, yielding a March 1-day headcount of 12,295 based on the time-series forecasting model.

3.2. The effect of word attributes on difficulty mode scores

In this paper, the attributes of English words are first divided into a total of 30 categories, the first major category is word meaning (how many official explanations of the meaning of a word there are), the second major category is word lexicality (the result of classifying words mainly based on their grammatical features, including syntactic functions and morphological changes, taking into account the meaning of the word), the third major category is the length of the word (the number of letters that make up a word), the fourth major category (the number of times a word has been viewed in Google the number of times a word is viewed in Google), and in the fifth major category, the number of times the five letters that make up an English word appear in the 26 letters of the alphabet. A normality test and Spearman's correlation analysis were used to determine if any of the attributes of the words affected the percentage of scores in the difficult mode. Secondly, according to the rule, the puzzle letters were five, so the length attribute of the word was not related to the percentage scored in the difficult mode. Then the data was subjected to Shapiro-Wilk [7] (small data samples, typically sample sizes of -low 5000) or Kolmogorov-Smirnov [8] (large data samples, typically sample sizes of 5000 or more) tests to check for significance. If it does not show significance ($p>0.05$), this means that it conforms to a normal distribution and vice versa. As in Figure 2: a normality test histogram and as in Figure 3: b normality test histogram.

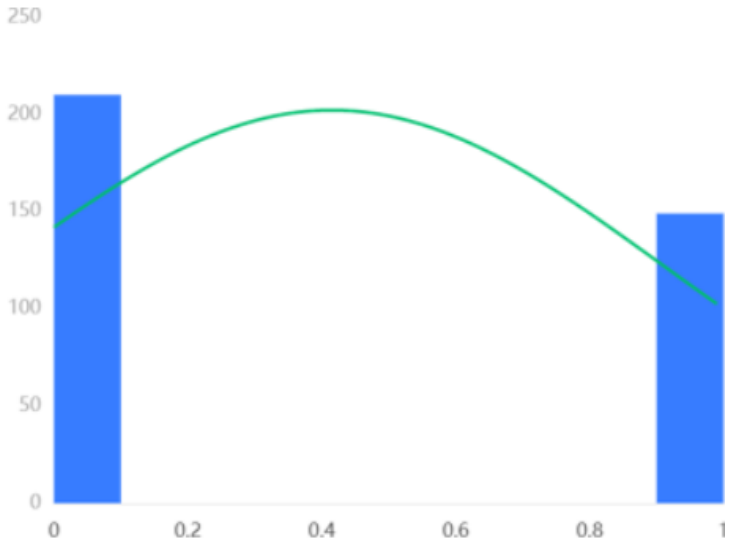

Figure 2. A histogram of the normality test

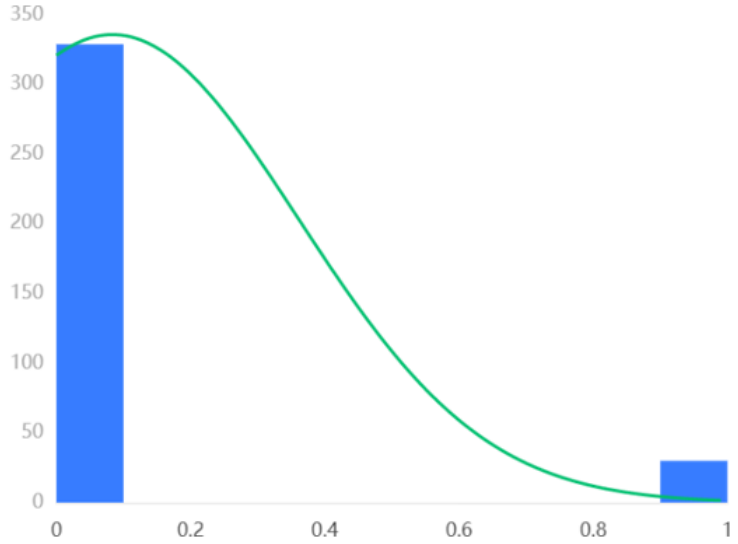

Figure 3. Histogram of the b-normality test

This paper analyzes the properties of the 26 letters of the alphabet to produce 26 histograms of normality tests [9]. Only the normality test histograms associated with a and b are shown above, the normality test histograms for the remaining 24 letters yielded the same operations as a and b. The histograms for the remaining 24 letters are shown above. It was finally concluded that the data did not meet the normality test based on the mean and variance, normality test histogram, and other data. The data was then analyzed by Spearman [10] correlation. It was tested whether there was a statistically significant relationship between XY ($p<0.05$) and then analyzed the direction of the correlation coefficients, both positive and negative, as well as the degree of correlation. As shown in Table 3: Table of correlation coefficients:

Table 3. Table of correlation coefficients

	a	b	c	d	e	f	g	h
Difficult Percent-age	0.041 (0.436)	0.044 ((0.410)	-0.058 (0.275)	0.047 (0.376)	0.082 (0.122)	0.001 (0.982)	0.028 (0.593)	-0.09 (0.087*)
	i	j	k	l	m	n	o	
Difficult Percent-age	0.042 (0.430)	0.042 ((0.430)	-0.085 (0.107)	0.034 (0.519)	-0.038 (0.473)	-0.027 (0.613)	-0.11 (0.038**)	
	p	q	r	s	t	u	v	
Difficult Percent-age	0.004 (0.943)	0.029 (0.579)	-0.103 (0.052*)	-0.09 (0.088*)	0.000 (0.996)	0.012 (0.816)	0.069 (0.191)	
	w	x	y	z	Word Views	Word Meaning	Lexicalitof words	
Difficult Percent-age	-0.068 (0.201)	0.078 (0.141)	0.058 (0.269)	0.1 (0.058*)	-0.093 (0.077*)	-0.154 (0.003***)	-0.049 (0.357)	

This paper first tests whether there is a statistically significant relationship between XY to determine whether the p-value presents significance ($p<0.05$). If it presents significance, it means that there is a correlation between the two variables and vice versa, there is no correlation between the two variables. Analyze the positive and negative direction of the correlation coefficient and the degree of correlation. The research team conducted a Spearman [9] analysis on 29 data to determine whether the word attributes are related to the difficulty pattern by analyzing the two variables based on ($P<0.05$). The research team concluded that there was no relationship between word genericity, word lexicality, and the number of letters in the 26 letters that make up the word and the difficulty mode. The lexical meaning of words (p less than 0.05) was related to the difficulty mode. The more types

of meanings a word has, the more people will be able to use the word in different contexts and the more they will be able to use the word. The more types of word meanings, the more people remember the different types of meanings of the word in real life, and the more the brain recognizes the word. So, word meanings are related to patterns of difficulty.

4. Vocabulary Prediction and Uncertainty Analysis Based on the GBDT Model

4.1. Predictions for the term EERIE

In this paper, the Gradient Boosted Tree (GBDT) regression algorithm is used to solve this problem as the percentages to be predicted belong to the numeric category of data. In this paper, the 30 words summarized previously with the genus of the relevant attributes are used to build the model, which are the percentage of the first result, the percentage of the second result, the percentage of the third result, the percentage of the fourth result, the percentage of the fifth result, the percentage of the sixth result, and the percentage of the seventh result that has no result, and reclassified, which produces seven models. When it is necessary to predict the seven relevant percentages for a future date, it is sufficient to analyze the attributes of the next 30 words. Enter them into each of these seven models to be able to get the seven percentages needed for this paper. Let this paper take the first resulting percentage as an example and implement it in the following way. As in Fig. 4: The importance of features is shown by the calculation of the built Gradient Boosted Tree (GBDA):

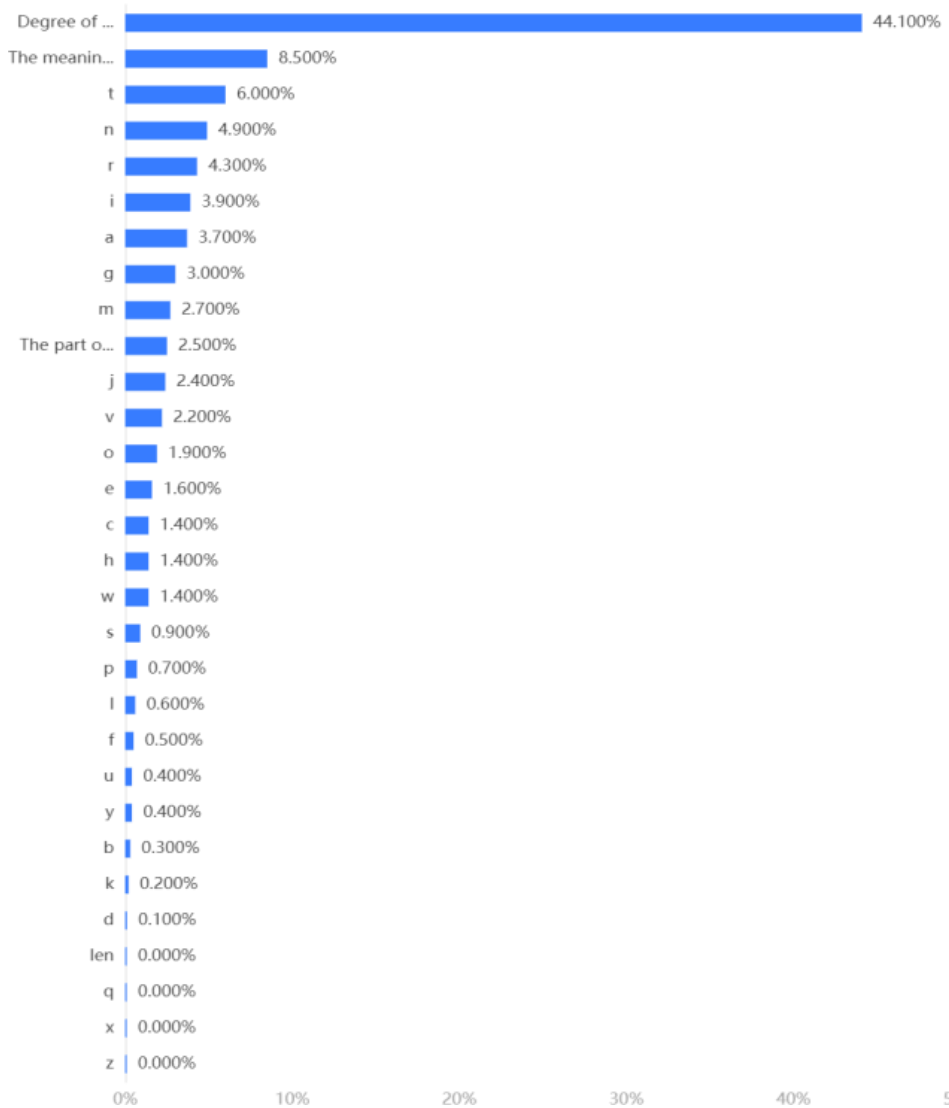

Figure 4. The importance of features is calculated from the built Gradient Boosting Tree (GBDT)

Gradient Boosted Tree (GBDT) is stochastic and the results are different for each operation, if this training model is saved, future data can be directly uploaded to this training model for computation and prediction, resulting in the relevant prediction results Table As shown in Table 4: Forecast Results:

Table 4. Predicted results

Predicted results Y	1 try	Len	a	b	c	d	e	f	g	The meaning of the word	The part of speech of the word	h	i	j
0.051341059894493946	0	5	0	0	0	0	0	0	1	10	2	0	0	0
0.710685866658167	1	5	0	0	0	0	0	0	0	5	1	1	0	0
1.8734302270847583	1	5	1	0	1	0	1	0	0	4	1	0	0	0
1.6125750086257404	1	5	0	0	0	1	1	0	0	17	3	0	1	1
0.586760505188961	1	5	0	0	0	0	0	0	1	8	1	0	1	1
0.5590479241167966	0	5	0	0	1	0	1	0	0	4	1	1	0	0
0.11605680587768676	0	5	0	0	0	0	1	0	0	4	1	1	0	0
0.019677532951419507	0	5	0	0	0	0	1	0	0	14	2	0	1	1
-0.0004917056126836541	0	5	0	0	0	0	0	0	1	15	2	0	0	0
0.06770404112700167	0	5	0	0	0	0	1	0	0	9	1	0	0	0
0.2228763521595766	0	5	1	0	0	0	0	0	0	4	2	0	0	0
0.39427586140886933	0	5	1	0	0	1	0	0	0	19	1	0	1	1
1.6225413816722858	1	5	0	0	0	0	1	0	0	17	1	0	1	1
0.00047053745057956393	1	5	1	0	0	0	0	0	0	4	2	0	1	1
0.10200393271186661	5	5	1	0	0	1	1	0	0	8	1	0	0	0

By searching for 30 attributes of the word eerie and bringing them into the above model, a table of percentage values for the derived results can be obtained As shown in Table 5: Percentage values of derived results

Table 5. Percentage values of derived results

Predicted results Y	len	Commonness	Word meanings of words	Etymology of words	a	b	c	d	e	f	g	h	i	j
0.5190347134981548	5	21	2	1	0	0	0	0	3	0	0	0	1	0

Similarly, the thirty attributes of the word eerie can be carried over into the remaining six models to obtain six percentages, which, after rounding, are shown in the following table

Table 6. Proportion related to number of attempts

Related percentages	1	2	3	4	5	6	X
Predicted value	1	4	18	28	28	16	5

As shown in Table 6: Proportions related to the number of attempts: This paper was tested to obtain a table of percentages associated with the number of attempts, and the sum of the individual predictions is 100%, which meets the requirements of the question with a high degree of accuracy. The final percentages associated with the number of times the word EERIE solved play 1, 2, 3, 4, 5, 6, and X were 1%, 4%, 18%, 28%, 28%, 16%, and 5%.

4.2. Uncertainty associated with models and forecasts

4.2.1. Accuracy of raw data provided by headings

The data for the test set in the model used in this paper comes from the table provided in the title Percentage of one guess, two guesses, three guesses, four guesses, five guesses, six guesses, and unsolvable puzzles (denoted by X), which is part of the title that describes the source - " These results were obtained by mining Twitter" and "many (but not all) users on Twitter report their scores on Twitter". This suggests that the accuracy of this data has yet to be verified and that there may be inaccuracies in the percentage of people reporting scores, in which case using word attributes for prediction may make the fit less effective and affect the accuracy of the model.

4.2.2. Accuracy of word attributes collected in this paper

Some of the word attributes in the model of this paper are taken from some official website data, such as the commonness of words, which is for all age groups and can be considered average, but people playing this game are more inclined to young people, and for young people, the commonness of words will be somewhat different from the average, which will cause some bias, and the prediction using such data The prediction of the percentage of guessing one, two, four, five, six, and unsolvable puzzles (denoted by X) using such data to predict the percentage of guesses of one, two, three, four, five, six, and unsolvable puzzles (denoted by X) would be more or less subject to some error.

4.2.3. Ratio of the number of new to old players

In this problem, this paper uses the attributes of the words to predict the percentage of each category, but I think this is not comprehensive it is not only the attributes of the words that affect these relevant percentages, but also other factors such as the change in the percentage of new and old players playing the game, assuming that an increase in the percentage of old players in this game will result in a decrease in the number of successful puzzles being solved and a corresponding increase in the percentage of attempts. And vice versa. Therefore, the ratio of the number of new and old players affects the percentage of guessed words in one attempt, two attempts, three attempts, four attempts, five attempts, six attempts, and unsolved puzzles (denoted by X) to some extent, which is one of the uncertainties of this topic.

5. Conclusion

In this paper, a model is developed to explain the changes in the number of reported results by performing a time series analysis of daily results data throughout the year 2022. It was found that the number of scores reported by players showed a trend of rapid growth followed by a gradual decrease. Through detailed analysis of the data, the team predicted that the number of reported results on March 1, 2023 would be approximately 12,295 results. In an in-depth analysis of word attributes, the team identified 30 relevant attributes and explored how these attributes impacted player performance in hard mode. The team's modeling analysis revealed a number of interesting findings: 1. word commonness: the team found that word commonness was not directly related to player performance in the difficulty mode. 2. number of word meanings: the team found that the number of word meanings was related to the difficulty mode. Specifically, the more meanings a word has, the more the player recognizes it.

In this paper, a Gradient Boosted Tree (GBDT) regression model was used to predict the relevant percentage of games completed on a given date. By modeling the data for the year 2022, the team was able to predict the percentage of players solving EERIE words for different numbers of attempts. The team's model predictions showed that: 1. 1% of players solved on one attempt. 2. 4% of players solved on two attempts. 3. 18% of players solved on three attempts. 4. 28% of players solved on four attempts. 5. 28% of players solved on five attempts. 6. 16% of players solved on six attempts. 7. 5% of players solved on seven or more attempts. This paper also takes into account uncertainties in the data sources, including the accuracy of the raw data, the accuracy of the word attributes, and changes in the proportion of new and old players. Through these detailed analyses and predictions, this paper provides The New York Times with insights into the player performance of Wordle puzzle games, and offers valuable references for future puzzle design and game improvements.

In this paper, a time series model and a gradient boosting tree algorithm are used to effectively predict the change in the number of reported outcomes and the future distribution of word solving in Wordle games. The study fully utilizes the advantages of time series models in capturing time dependence and the accuracy of the gradient boosting tree algorithm and obtains reliable prediction results through data preprocessing and feature engineering. In addition, this paper explores the potential of this approach for applications in other domains, such as frameless applications and natural language processing, providing new ideas for solving similar problems.

References

[1] Feng Naixing, Wang Huan, Zhu Zixian, et al. Perfectly matched single-layer intelligent algorithm based on limit gradient boosting for efficient absorption of aero transient electromagnetic problems [J]. Physics Letters, 2024, 73 (06): 173 - 81.

[2] Liu Enchen, Yang Xiaolong, Luo Xinwu. The Application of Logical Thinking Training in Professional English Teaching: Starting from the Representation of "Properties" of Things by Words [J]. Overseas English, 2021, (02): 100 - 1.

[3] DU Zhiyong, YANG Fan, YANG Wenjie. Study on daily maximum temperature prediction for long time series based on Conv1D-LSTM hybrid model [J]. Journal of Beijing Institute of Printing, 2024, 32 (09): 52 - 7.

[4] SUN Le Zheng, ZHAO Meng Di, DU Shaobo, et al. Prediction of game difficulty and influencing factors based on multivariate linear regression model [J]. Digital Technology and Application, 2023, 41 (12): 85-7.

[5] Cai Zhongzhe, Zeng Riwei, Lin Chengcheng, et al. Prediction and Analysis of Wordle Answering [J]. Journal of Taizhou College, 2023, 45 (06): 8 - 13+21.

[6] ZHANG Mingming, ZHANG Mingdong. Application of combined gray Verhulst-time series model in settlement monitoring [J]. Mapping and Spatial Geographic Information, 2024, 47 (09): 151 - 3.

[7] LI Junfeng, YANG Xin. Impact of water level change on water quality in Caizihu Lake [J]. Anhui Chemical Industry, 2024, 50 (02): 124 - 9.

[8] Xutuo Wang, Yuting Wei, Huanhuan Zhang. Credit scoring model selection method based on Kolmogorov-Smirnov (KS) statistic [J]. Mathematical Statistics and Management, 2024, 43 (01): 100 - 16.

[9] Liu Jindong, Wang Tencheng, Chen Yonghong, et al. Improved method based on Anderson-Darling normality test and its application [J]. Practice and Understanding of Mathematics, 2023, 53 (04): 184 - 93.

[10] Li Peng, Luo Xiangchun, Meng Qingwei, et al. Ultra-short-term load forecasting for customer-level integrated energy systems based on Spearman correlation threshold optimization and VMD-LSTM [J]. Global Energy Internet, 2024, 7 (04): 406 - 20.