

Prediction of Scrabble game players based on ARIMA and Machine Learning under big data

Xinhua Lu*

School of Computer Science, South China Normal University, Guangzhou, China, 510631

*Corresponding author: 13613083494@163.com

Abstract. Wordle is a favored puzzle. Firstly, to predict its daily player count, based on analyzing the inherent trend of the data and verifying its stationarity, the effectiveness of using the $ARIMA((2,3,5),2,(2,3,5))$ model to predict the number of reported results is proved. Before predicting, this thesis preprocesses the data from two parts: Number of reported results and Number in hard mode. Then, the 95% prediction interval for March 1, 2023, is [8626 16199]. Secondly, to predict the distribution of reported results percentages, this thesis uses 7 random forest regressors, their feature variables have 5 dimensions (Contest number, Frequency of word, Number of repeated types, Number of letter types, Maximum number of repetitions), and the response variable takes one of the percentages (7 dimensions) in turn. Results show that the distribution of reported results percentages of "SALSA" on March 1, 2023 is [0, 2, 15, 35, 31, 15, 2].

Keywords: Big Data, ARIMA, Random Forest Regressor, Word Feature Mining.

1. Introduction

Wordle is a popular puzzle that players are supposed to solve the puzzle by guessing a five-letter word in six tries or less. Most players solve the puzzle based on the frequency of the word, spelling difficulty, and word attributes, which are also the sources of puzzle difficulty.[1]

In order to explain the daily number of reported results and create a prediction interval for the number of reported results in the future, and utilize the features mined in the word[2] to predict the percentage of the distribution of game attempts, e.g. 1, 2, . 6 times, 7 or more times. Our work mainly includes the following:

Based on the preprocessed historical data on daily number of reported results, an ARIMA prediction model[3] of the daily number of reported results is established to create a prediction interval for the number of reported results on March 1, 2023.

Then, the thesis establishes a random forest regression model[4], which uses the 5-dimensional data (The number of repeated types, The number of letter types, The maximum number of repetitions, Word frequency, and Contest number) as feature variables(The first three are mined from the features of words[5]), and use the 7-dimensional data of the number of word attempts as the response variable in turn to construct 7 random forest regression models. This thesis then makes predictions on the relevant percentage of attempts for the word SALSA on March 1, 2023.

2. Numerical Data Description and Cleaning

2.1. Data Description

The data is directly visualized as Figure 1, since the data is not intuitive.

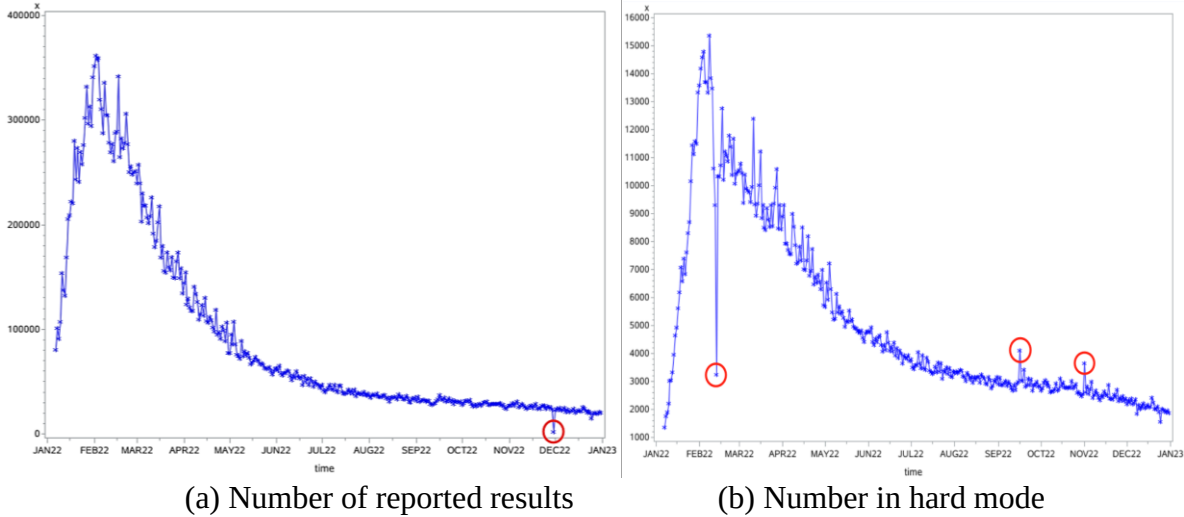


Figure 1. The Original Sequence

Figure 1(a) has an abnormal value on November 30, 2022, recorded as X_i . The solution word corresponding to X_i is study, the number of reported results is 2569, and the number in hard mode is 2405. The study is classified as elementary level in the Oxford Dictionaries[6], while the number of reported results is significantly less than the number of results the previous day. At the same time, X_i only affects the number of reported results on November 30, 2022, but not the subsequent values.

This thesis uses the PauTa criterion to judge whether to remove X_i and takes the reported result values $\{X_{i-5}, X_{i-4}, \dots, X_{i-1}, X_{i+1}, \dots, X_{i+5}\}$ of the 5 days before and after X_i respectively. Then, calculating the mean $\bar{X}$ and variance σ , if $|X_i - \bar{X}| > 3\sigma$, X_i should be removed.

Similarly, Figure1(b) has 3 abnormal values on November 30, 2022. After judgment, all of them should be removed.

2.2. Data Cleaning based on KNN Regressor

The KNN regressors[7] are a smoothing process for outliers. This thesis takes these two columns of data containing outliers as the feature variables and response variables of the model in turn to construct two KNN regressors (here $K = 5$). For each outlier, the non-missing eigenvalues corresponding to the outlier are substituted into the model to obtain the K sample points with the shortest distance (Euclidean distance), and then the model uses the average value of the corresponding features in the K sample points to replace outlier.

The images after the outlier correction are shown as Figure 2:

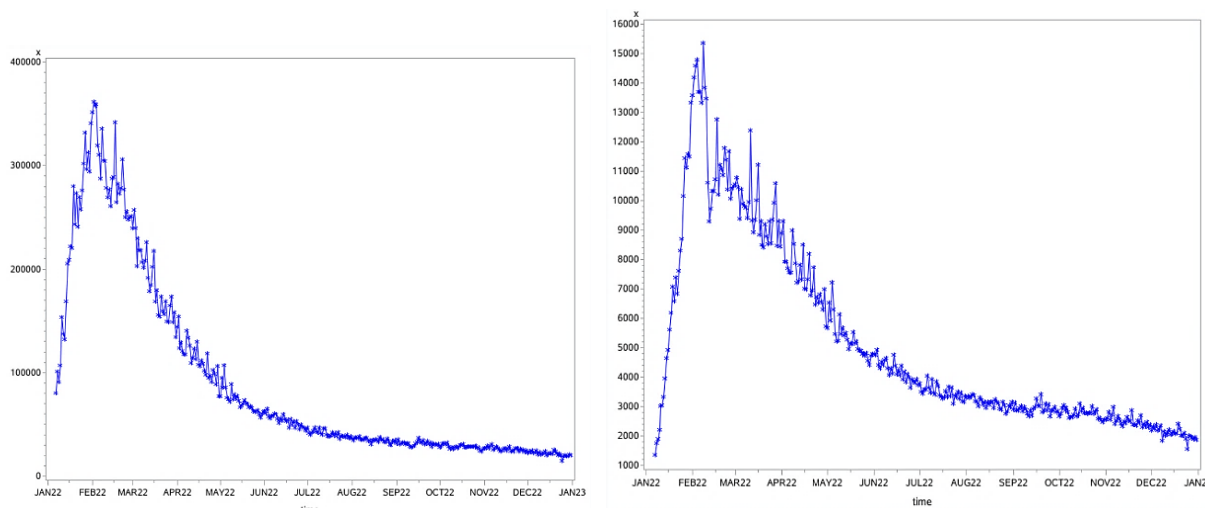
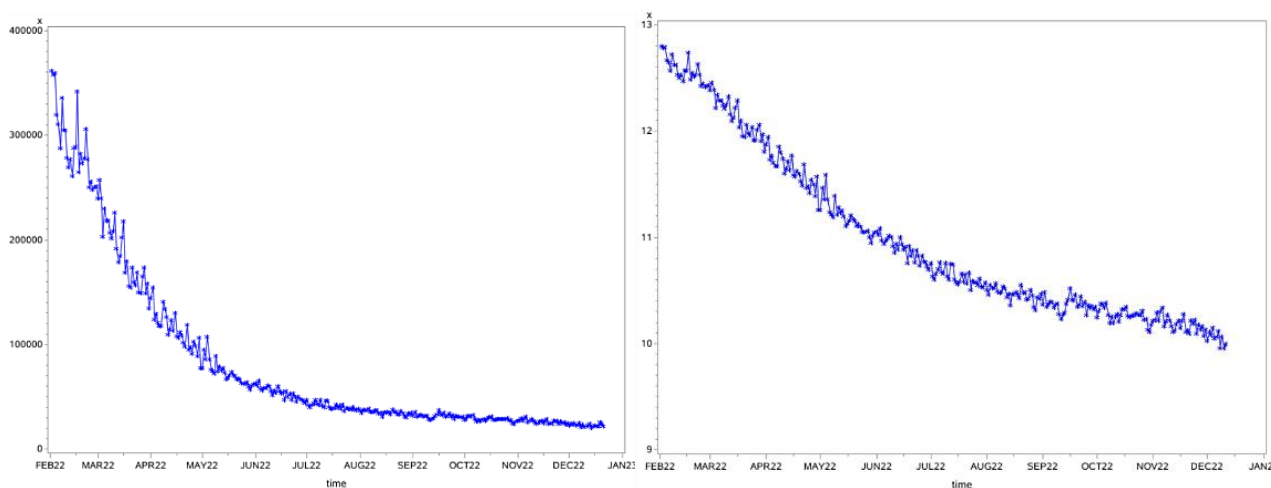


Figure 2. The Corrected Sequence

3. Daily Reported Results Prediction Model

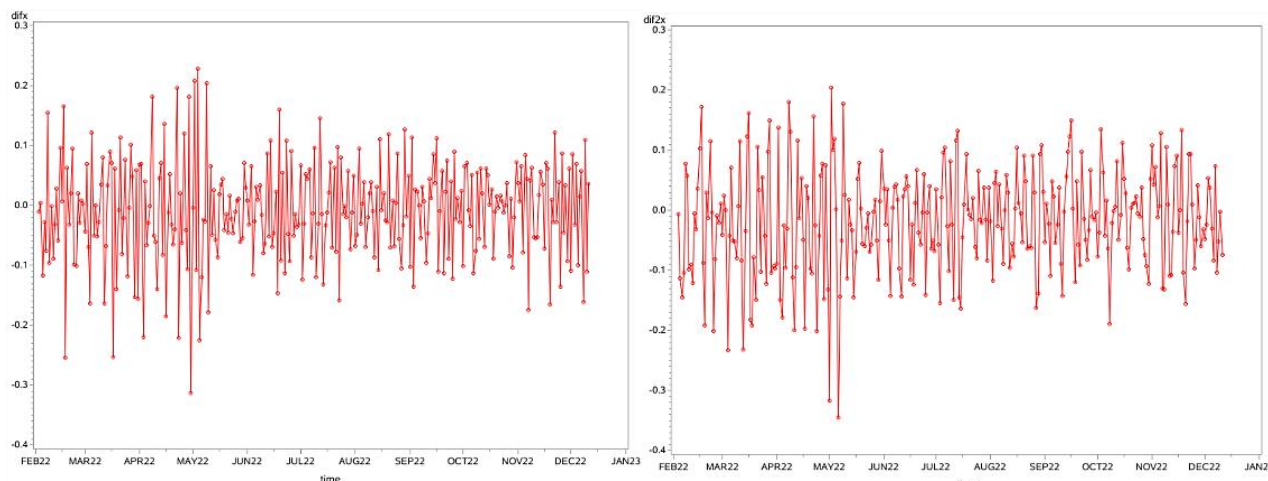
3.1. Data Stabilization

Since time series often have characteristics such as trend and heteroscedasticity, if the data is analyzed directly, it is likely to lead to spurious regression and other phenomena[8]. Therefore, before performing regression analysis on the data, the original sequence should be logarithmic to eliminate characteristics such as heteroskedasticity, and then take the first-order difference and second-order difference of the logarithmic sequence. Considering the timeliness of data, this thesis uses data from February 2, 2022, to December 11, 2022, for forecasting as Figure 3(a). The data after December 12, 2022, is used as the test set.



(a) Original Sequence

(b) Data after taking the logarithm



(c) First Difference

(d) Second Difference

Figure 3. Timing Diagram

As Figure 3(b), there is an obvious trend in the data which is after taking the logarithm, so this thesis uses the difference method as Figure 3(c) and (d) to remove the trend of the original data. The Ljung-Box test was carried out on the above sequence, and the results are shown as Table 1:

Table 1. Ljung-Box Test Results

Lag Order of Second Difference	Chi-Square	Degrees of Freedom	P-value
6	76.01	6	<.0001
12	83.47	12	<.0001
18	99.83	18	<.0001

The sequences after the difference both have passed the white noise test. This model finds that the sequence after the second difference is a white noise sequence at the 1% significance level. Since the second-order difference is more stable than the first-order difference, and the parameters are easier to find, the model performs further analysis on the data after the second difference.

Next, the model uses the ADF test to test the stationarity of the sequence data after the second-order difference, and the results are shown as Table 2:

Table 2. ADF Test Results

Type	T-statistic	P-value
Zero Mean	-15.29	<.0001
	-17.64	<.0001
Single Mean	-15.77	<.0001
	-18.78	<.0001
Trend	-15.94	<.0001
	-19.20	<.0001

3.2. Identification of the ARIMA Model

For a stationary time sequence transformed from a non-stationary time series, the random process composed of autoregressive and moving average parts is denoted as $ARIMA(p, q)$, where p and q respectively represent the maximum order of autoregressive and moving average parts. The general expression for $ARIMA(p, q)$ is

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \mu_t + \theta_1 \mu_{t-1} + \theta_2 \mu_{t-2} + \dots + \theta_q \mu_{t-q} \quad (1)$$

The number of differences is the order d in the $ARIMA(p, d, q)$ model. After the data is stabilized, the $ARIMA(p, d, q)$ model is transformed into the $ARMA(p, q)$ model. The model chooses the second-order difference, then d takes 2.

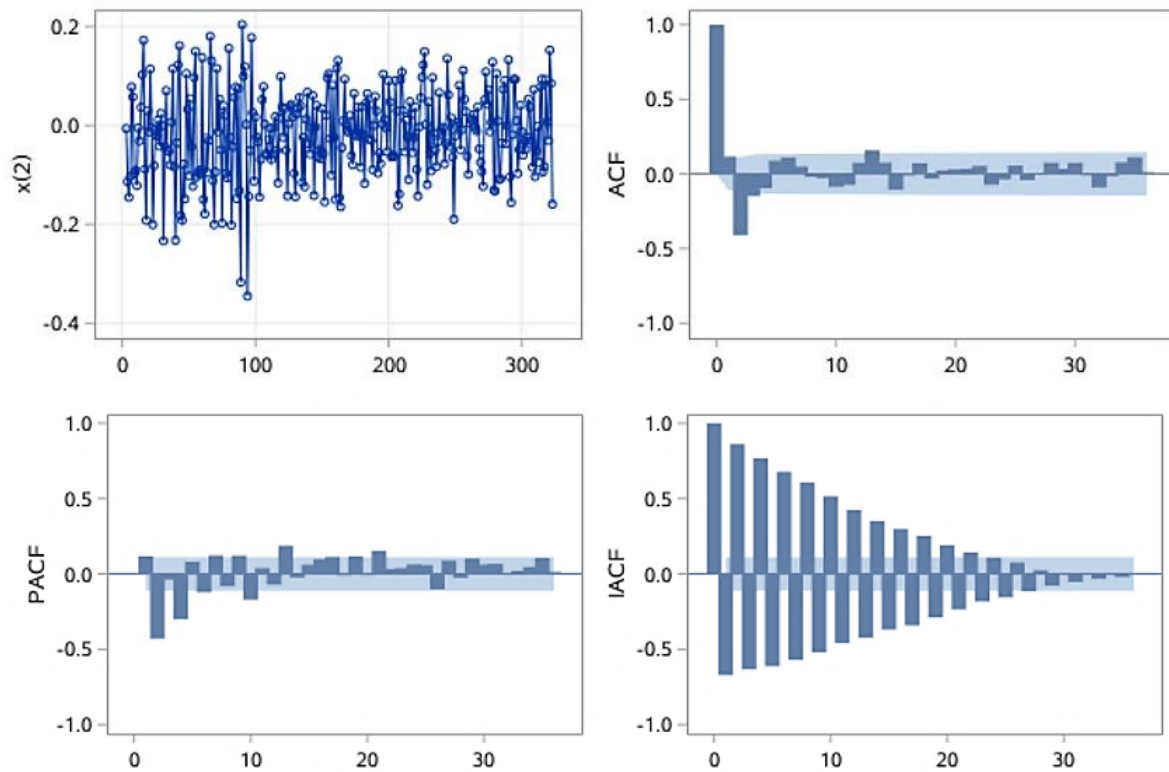


Figure 4. Trend and Correlation Analysis

From Figure 4, the PACF after the second-order difference has a serious tailing phenomenon, and the lag of 2 and 5 in this graph is greater than twice the standard deviation. In ACF, the ACF after the second order difference is basically within two standard deviations after the fourth order, but the lag of the third order in this graph is greater than two standard deviations. According to the test results, the sparse coefficient model $ARMA((2,3,5), (2,3,5))$ was selected to initially fit the model.

3.3. Parameter Estimation of ARIMA Model

The model uses maximum likelihood estimation (ML) for parameter estimation, and the results are shown as Table 3:

Table 3. Parameter Estimation of ARIMA Model

Parameter	Estimated value	P-value
$MA1,1$	0.94486	<.0001
$MA1,2$	0.97260	<.0001
$MA1,3$	-0.95101	<.0001
$AR1,1$	0.17505	0.0040
$AR1,2$	0.98085	<0.0001
$AR1,3$	-0.15698	0.0091

The P values of all the parameters in Table 4 are less than 0.05. At the 1% significance level, the above parameters are considered as 0 if they are not significant, and all parameters of the model pass the significance test.

Finally, the form of the model can be obtained as

$$\ln X_t = \frac{1-0.94486B^2-0.9726B^3+0.95101B^5}{1-0.17505B^2-0.98085B^3+0.15698B^5} \varepsilon_t, \varepsilon_t \sim NID(0,0.004886) \quad (2)$$

3.4. Model Checking

The model conducts a white noise test on the residual sequence, and the results are shown as Table 4:

Table 4. White Noise Test of Residual Series

Lag Order	Chi-Square	P-value
6		<.0001
12	31.94	<.0001
18	40.86	0.0002
24	47.50	0.001

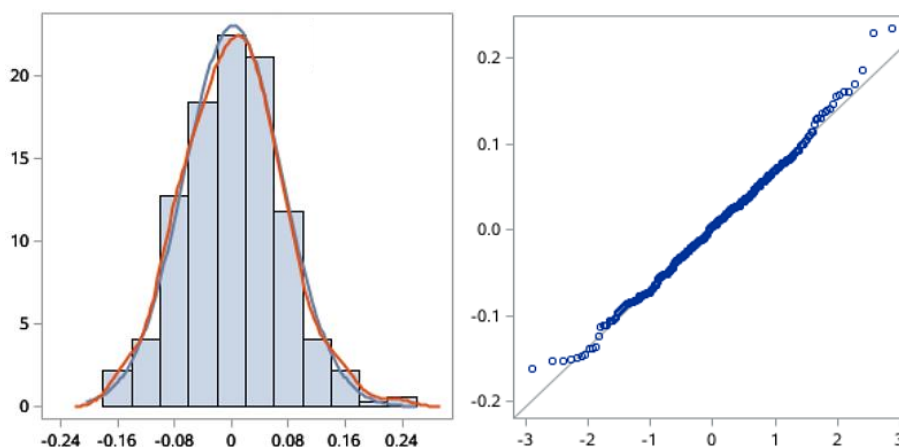


Figure 5. Diagnostic Results for Residual Series

The p-values of the white noise test of the residual sequence are all less than 0.05, and the sequence is white noise after passing the significance test. From Figure 5, the sequence distribution presents a normal distribution, and the distribution meets the model requirements.

3.5. Model Prediction Interval

The thesis gets the prediction interval of the model at the 95% confidence level. The prediction effect diagram of the model is shown as Figure 6:

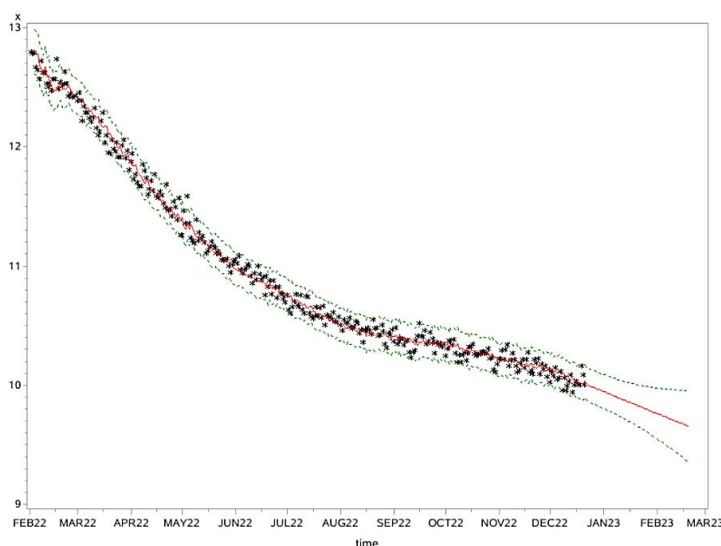


Figure 6. Model Prediction

It can be seen from the prediction effect chart of the model that the prediction and fitting ability of the model are excellent. The thesis uses this model to predict that the 95% prediction interval for March 1, 2023 is [8626,16199].

4. Percentage Distribution Prediction Model based on Random Forest Regressor

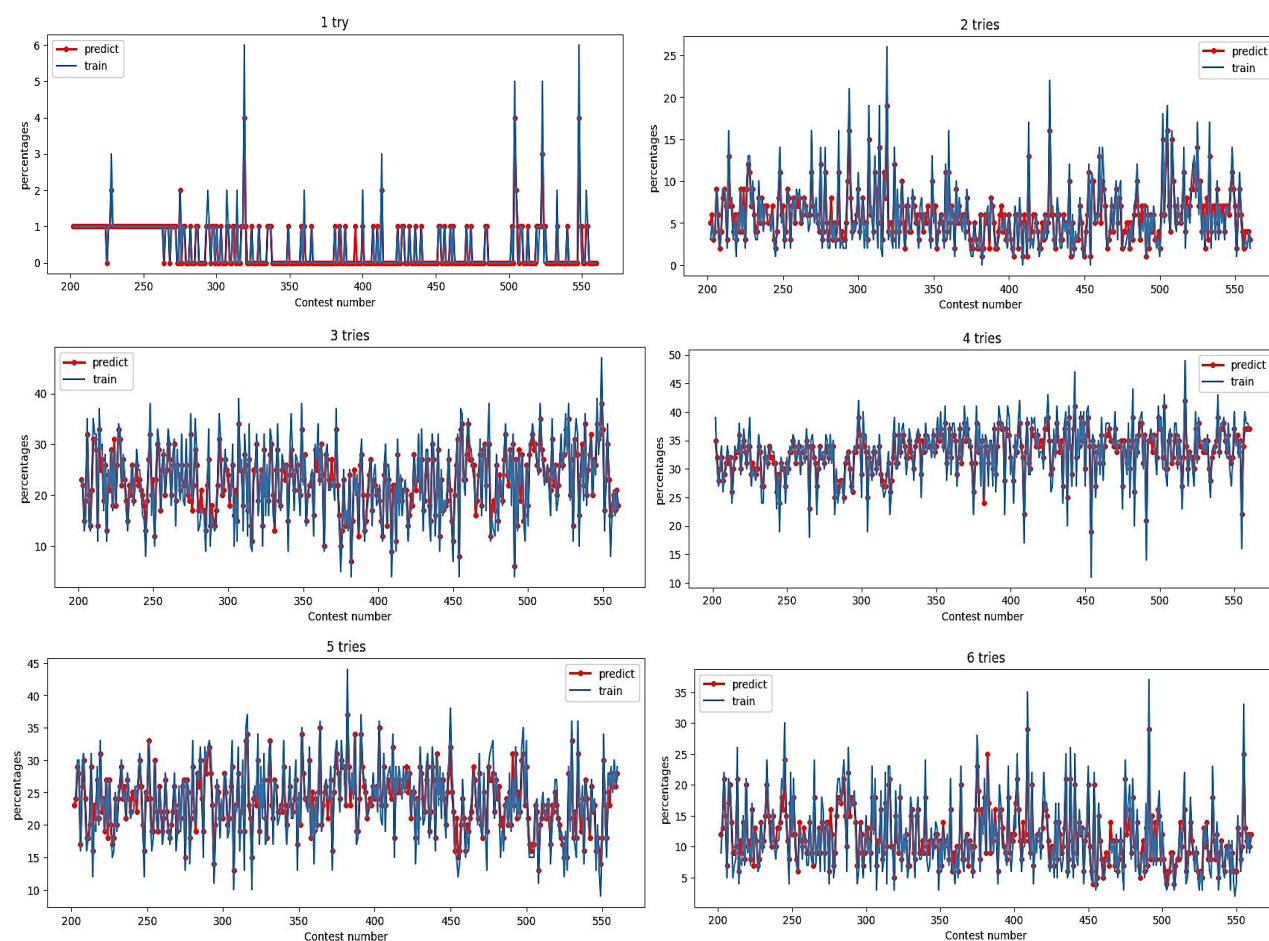
4.1. Model Building and Parameter Search

Random forest is an algorithm that integrates multiple trees through the Bagging[9] idea of ensemble learning. This learning method makes the model have a stronger fitting ability.

The feature variables of the random forest regressor is 5-dimensional, which are contest number (representing the time axis), Frequency of word(from The Corpus of Contemporary American English[10]), Number of repeated types, Number of letter types, Maximum number of repetitions(the latter three are mined from the features of words). Since the correlation between part of speech and percentage distribution is extremely low, the thesis does not select it as a feature. The model sequentially takes one dimension in the associated percentages of (1 try, 2 tries, 3 tries, 4 tries, 5 tries ,6 tries, 7 or more tries(X)) as the response variable, and construct 7 corresponding regression models respectively.

The regressor uses squared error as a measure of the segmentation quality of random forest regressors. In order to prevent the model from underfitting, the model adjusts the estimators parameters of the regressor (the number of trees in the forest); and to prevent the model from overfitting, the model uses the pre-pruning method to adjust the depth of the tree, the mini-mum number of samples required to split an internal node and other parameters. The parameters of the random forest regressor are set by means of hyperparameter search[11].

The model respectively takes the associated percentages in (1 try, 2 tries, 3 tries, 4 tries, 5 tries ,6 tries, 7 or more tries(X)) as the response variable and substitute them into the random forest regression model, and the results are shown as Figure 7:



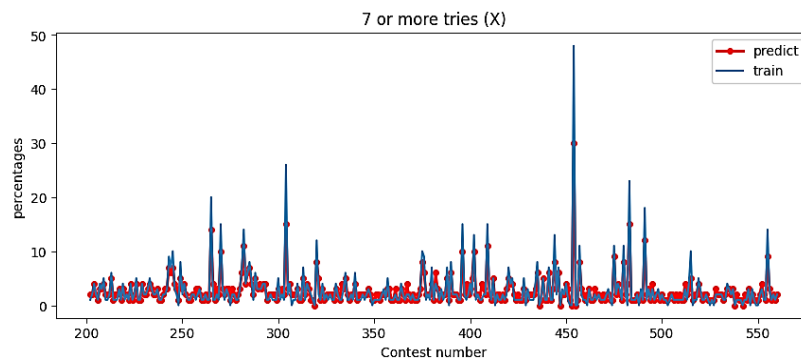


Figure 7. Model Fitting Results

The tuning parameters and goodness of fit of the seven regression models are shown as Table 5. Among them, ***N_estimators*** indicates the number of estimators, ***Max_depth*** indicates the maximum depth of the tree, and ***Min_samples_split*** indicates the minimum number of samples contained in each split point, ***R²*** indicates fitting coefficient.

Table 5. The Tuning Parameters and the Goodness of Fit

<i>Response Variable</i>	<i>N_estimators</i>	<i>Max_depth</i>	<i>Min_samples_split</i>	<i>R²</i>
1try	59	16	2	0.857
2tries	75	18	2	0.865
3tries	98	21	2	0.871
4tries	110	21	2	0.832
5tries	105	19	2	0.867
6tries	95	20	2	0.849
7 or more tries (X)	51	22	2	0.818

The mean fitting coefficient of 7 models is 0.8518, it shows that the ability of this model to fit is not bad. And the following Figure 8 is a partial picture of one of the estimators of the random forest regressor corresponding to "1 try".

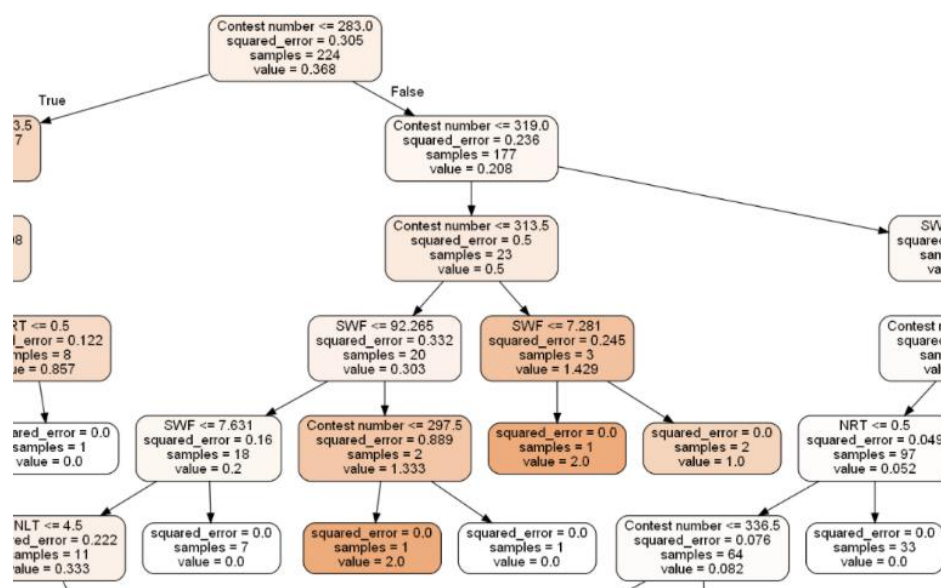


Figure 8. Partial random forest

4.2. Prediction of distribution of reported results percentages

The thesis inputs the Attributes(Contest number, Number of repeated types, Number of letter types, Maximum number of repetitions, Frequency of word) of SALSA [620,2,3,2,0.00015892] into the trained model in turn, and the predicted percentages distribution is shown as Table 6:

Table 6. Forecast Result

<i>1try</i>	<i>2tries</i>	<i>3tries</i>	<i>4tries</i>	<i>5tries</i>	<i>6tries</i>	<i>7 or more tries (X)</i>	<i>total</i>
0	2	15	35	31	15	2	100

The associated percentages of the word SALSA on March 1, 2023, are shown in the above Table 6, and the sum of the percentages is 100%, which meets the accuracy requirements, indicating that our model has a good prediction effect.

5. Conclusions

Wordle is a popular puzzle. To predict Wordle's daily player count and the distribution of reported results percentages, the thesis builds 2 models.

Model 1: The thesis preprocesses the data from 2 perspectives: Number of reported results and Number in hard mode. Then, based on analyzing the inherent trend of the data and verifying its stationarity, the effectiveness of using the ARIMA model to predict the number of reported results is proved. Then the thesis uses the historical data to fit the parameters of ARIMA, and finally determine the prediction model as ARIMA((2,3,5),2,(2,3,5)). According to the evaluation criteria, the prediction effect of the model is excellent.

Model 2: The thesis uses 7 random forest regressors, their feature dimension is 5 dimensions (Contest number, Number of repeated types, Number of letter types, Maximum number of repetitions, Word frequency), and the response variable takes one of the percentages (7 dimensions) in turn. The thesis adjusts the parameters such as estimators by hyperparameter searches. The mean fitting coefficient of 7 models is 0.8518, indicating that the model predicts well.

Finally, the thesis gets the prediction results: the 95% prediction interval for March 1, 2023, is [8626 16199]. Then for the word SALSA on March 1,2023, the distribution percentage of the number of attempts to predict it is [0, 2, 12, 22, 31, 24, 10].

References

- [1] WU Peijin. Journal of Chifeng University(Natural Science Edition), 2022,38(06):22-26.DOI:10.13398/j.cnki.issn1673-260x.2022.06.013.
- [2] Wu Jinxia. Text Difficulty Judgment for English Learning[D]. Harbin Institute of Technology, 2007.
- [3] Tuo Siwei. Global Sales Forecast of Home Game Consoles and Review of Entering China——Analysis Based on ARIMA Model and PEST Framework[J]. China Market, 2015, No.866(51):130-134.DOI:10.13939/j.cnki.zgsc.2015.51.130.
- [4] Ye Congzhou, Xiao Penglin, Qin Jun, Zhang Chengxiong, Chen Lie, Wang Ke.Research on Energy Consumption Load Forecasting Model of Super Large Buildings Based on Clustering and Random Forest Regression[J].Green Building,2022,14(05):48- 51+55.
- [5] Xu Shasha, Ma Chao. Analysis on the Writing Questions of College Entrance Examination (Zhejiang Volume) Based on Text Feature Mining [J]. Foreign Language Testing and Teaching, 2021, No.43(03):60-64.
- [6] Study Definition & Meaning | Dictionary.com
- [7] Cheng Hongliang, Liu Hong, Bai Zhaoxu, Rao Siwei, Zhang Jian. A missing data filling method based on the nearest neighbor KNN algorithm [P]. Shaanxi Province: CN107193876B, 2020-10-09.
- [8] Zhou Liang. Journal of Hunan University of Finance and Econom-ics,2018,34(01):72-78.DOI:10.16546/j.cnki.cn43-1510/f.2018.01.008
- [9] Li Xiufang, Huang Zhiguo, Chen Xiaowei. Application Research of Bagging Integration Method in Insurance Fraud Identification [J]. Insurance Research, 2019, No.372(04):66-84.DOI:10.13497/j.cnki.is. 2019.04.006.
- [10] English-Corpora: COCA
- [11] Liu Ziliang, Ju Xiang, Zhang Yongfang, etc. Random Forest Parameter Tuning Optimization Based on Improved Random Search Algorithm [J]. Network Security Technology and Application, 2022, No.256(04):49-51.