Prediction and feature importance analysis for wordle strategy based on machine learning models

Zhiming Fan

Department of Mathematical Science, University of Nottingham (China), Yingzhou District, Ningbo, China

smyzf1@nottingham.edu.cn

Abstract. The word guessing game, Wordle, has attracted considerable attention from researchers in computer studies and mathematics due to its significant relevance. Many studies have been taken on Wordle to explore its potential in simulating human behaviors during gameplay and devising optimal strategies for players. In this paper, strategies based on mathematics theories or machine learning technologies are compared by their feature importance in order to explore whether there is a loss of feature meaning in machine learning methods. In the whole study, the effect of features and their relations to the expectations of tries of game Wordle are also explored. More specifically, this paper introduces machine learning models such as linear, random forest, bagging as well as gradient boosting decision trees to visually provide a relation between word’s features (such as repeat frequency, vowels, consonants) and the expectation of attempts in a game. After establishing machine learning models, the feature importance is derived by feature engineering techniques. The importance is then compared with spearman statistical correlations based on the dataset to hence draw the conclusion of the change of feature meaning in machine learning methods. Study results indicate that there is loss of meaning and effect of features in the better fitted prediction models (gradient boosting decision trees, random forest) compared to the statistic approaches.

Keywords: machine learning, wordle, feature importance, statisical correlation

1. Introduction

Wordle, a word guessing game introduced by New York Times, has captured the minds of the internet since its inception in last year. Because of its simplicity but strong correlation to computer studies, Wordle rapidly gets millions of people immersed in it, not only players, but also researchers in different fields. In this game, players are given at most six attempts for guessing a five-letter-formed word (words are different every day), at each attempt, they will be given relevant hints with colored tiles to help guessing (color tiles will be specifically introduced in next section) [1]. Given the game's profound relevance to informatics, mathematics, and computer science, research on it provides a great meaning for applications in algorithm simulating human behaviors in this type of games [2], and promotes studies in machine and deep learning technologies in game prediction [3].

In the previous studies of Wordle, researchers have devoted in discovering an optimal strategy to win this game and aiding in the analysis and optimization of individuals' decision-making processes during gameplay. One of the approaches was using mathematical and statistical theories to analyze the guessing process and make prediction of possible situations, refer to knowledges of probability and combinatorics [4, 5], combining with Wordle's linguistics backgrounds [1]. This prediction was frequently completed by various mathematical derivations, like fully discernible set, character statistics label [6], and vector rank one approximation [7]. These all provide a rigorous solution to main problem when considering an efficient strategy, including abstractly deducing the process of guess, and translating word information into mathematical languages [8]. Consequently, these mathematical approaches greatly improve the accuracy of making guess, efficiently reducing the expectation of total tries, for example, in [7], with an initial guess of "SLATE" the method, the author solves puzzle in 4.04 guesses on average, with success rate of 98.7%. In [1], internal key words and external words were selected in order to calculate determine combination with highest probability, as a result, the p-optimality method (find a best probability p) and p-FDS method (best probability p based on fully discernible sets) have expectation of 3.773 and 3.769. In addition to math, with the increase of Wordle's data, more strategies based on machine learning were proposed. Different from mathematical derivations, machine learning (especially reinforce learning) simulated the whole process and grabbed the main characters during each process to estimate parameters on-line, avoiding lack of practicality and making full use of big data. In [9], the authors apply Reinforcement Learning (RL) based on approximation in value space and the rollout algorithm to translate the process of Wordle into a Partially Observable Markov Decision Process (POMDP) problem and derive an adaptive control approach. The expectations of guess reduced 0.19 from 3.32 to 3.13. While in [2] and [10], epsilon-RL method was used when considering Wordle a NP-hard problem. In other research, the Convolutional Neutral Network with NP problem was also applied [11]. All these above indicate the progress in establishing the best strategy. Nevertheless, so far there is no specific research that discusses how features of the words under linguistics backgrounds affect its true expectation of tries. Because when using machine learning methods based on big data, sometimes some of the features may disappear or diverge from its original effect, which might be meaningless [12, 13]. Thus, this study aims to explore the effect of features and make comparison of their performance difference between math and computer approaches, when making prediction of expectations.

Under this consideration, this study designs to analyze relations between word features and this true expectation from Wordle's recorded data and establish machine learning models to visually present the results. To discover and transform word features, the specific standardization method is deployed. The machine learning models used for predicting composed of linear, random forest, bagging and Gradient Boosting Trees (GBDT). Finally, a comparison of feature importance in statistics and machine learning is conducted.

2. Methodology

2.1. Wordle rules (color tiles information)

The specific rules of Wordle are as follows: players need to guess for the correct five-lettered word in six attempts. After each attempt, colored tiles are given: the grey blocks represent the non-existence of letters, while the green blocks represent the existence of letters with actual correct positions in the word. The yellow ones mean that there are such letters, but they are in wrong positions.

The following Figure 1 shows a whole process: In first attempt (line 1), the grey blocks indicate that there is no letter "A" and "I" in this word, while yellow ones indicate there is "R, S and E" in this word, but not at positions as first guess. In second attempt (line 2), the green "R" means the existence of "R" at beginning of this word (actually, the word is rebus).

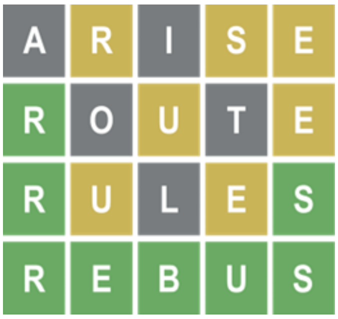

Figure 1. Ample game process.
Photo/Picture credit: Original

2.2. Dataset description and preprocessing

This study utilizes the revised dataset sourced from the Wordle Tweets dataset available on Kaggle [14]. The original dataset consists of all reported and published results since wordle 210. It records everyday player's data in 5 columns, including id, date, usernames, and answer matrix of wordle. Among them, the id and answer columns are extracted, which has 34,705,452 results. In other words, the total number of samples is 34,705,452, separated by different dates and wordle words.

The processing consists of three parts. Firstly, the whole dataset is categorized by date and title words. This classification is performed to facilitate the exploration of relations between word features and this true expectation, it is necessary to synthesize data with the same word together to obtain the whole data for each word such as the number of attempts and success rates. This procedure involves sorting out dataset by different date, from 2022-1-7 to 2023-4-30, because by the rule, the answer word for Wordle keeps the same every day. During this step, based on derived data, this study also lists different columns of data to describe every day's situation, these include the total numbers of tries, percentages for different tries (from 1 to more than 6), expectations of tries and success rate (success is defined to be the number of tries that are less than or equal to 6). Secondly, the dataset is revised since there are a few data not well recorded, some of their percentage (of tries) sum exceed 100%, while few of them has only 4-letter formed words. Some date, might be due to technique problems, has not been recorded. After taking a series of statistical tests, these data are scaled or removed, since the sample size is sufficiently large and there is no significant change under revision. The size of samples becomes 477 × 14 (477 words with 14 columns of feature or success rate indicators). Finally, and the most important, this study derives possible relevant features from words and then translates them into digital descriptions in dataset. A word, for example, "apple" can have the following linguistics features: five letters, vowels and consonants, repeated letters, different meanings. There should be more relevant characteristics, but the most influential and explicit ones are considered in this study. More specifically, in word apple, letters "a, p, p, l, e" are recorded by their position orders in alphabet, "a" is the 1st one, "p" is the 16th. Then for vowels, letters "a, e, i, o, u" are recorded, and others are recorded as consonants, while repeating frequency are counted as well ("apple" has repeating frequency of 2). The time features are considered as well.

2.3. Proposed approach

Machine learning models offer valuable tools for visually present relations between features and expectations [15-17]. Therefore, they are introduced in this study to choose the optimal model for predicting expectations of attempt for Wordle. The specific models used consists of linear, random forest, bagging and Gradient Boosted Decision Trees (GBDT). The following section introduces these models briefly.

Linear regression fits a linear model with coefficients of $a = (a_1, a_2, a_3 \dots)$ to minimize residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.

The random forest regression model establishes a forest of decision trees and combines all trees together, each tree in the ensemble is built from a sample drawn with replacement from the training set. Individual decision trees typically exhibit high variance and tend to overfit, by taking an average of individual tree's predictions, some errors can cancel out, and whole result can obtain a reduced variance.

Bagging method builds several instances of a black-box estimator on random subsets of the original training set and then aggregate their individual predictions to form a final prediction. It can improve single model and reduce the risk of overfitting.

GBDT applies several loss functions on decision tree regressor to ensure accuracy, which has various uses in machine learning. It adds newly fitted trees to minimize the sum of losses. By default, the initial model is chosen as the constant that minimizes the loss. Therefore, at each iteration, the estimator is fitted to predict the negative gradients of the samples. The gradients are updated at each iteration, which can be considered as some kind of gradient descent in a functional space.

After establishing models, every feature's performance in those computer science method is analyzed, and compared with statistical performance, like Spearman and Pearson correlations. This study then presents difference and similarity of feature importance in both math and computer method of strategies to make brief discussion.

2.4. Evaluations

Evaluation of prediction models are based on the calculation of R^2 score between the predicted results and training dataset:

$$SS_{res} = \sum_i (y_i - f_i)^2 \tag{1}$$

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \tag{2}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{3}$$

where y_i represents the predicted value of given data, f_i is the true value and $\bar{y}$ is the mean value of data. The score represents the degree of the model can explaining the observed data, the higher [18], the better it fits.

3. Results and discussion

3.1. Prediction performance

After applying machine learning algorithms on the dataset, models for predicting expectation are automatically fitted and a table containing calculated evaluations, such as R^2 scores and Mean-squared-error (MSE) is shown in Table 1.

Table 1. Prediction performance based on diverse models with evaluations of R^2score and MSE.

Models	R^2 score (Training set)	R^2 score (Testing set)	MSE
Linear	0.4531	0.3532	0.07635
Random Forest	0.9458	0.6244	0.04435
Bagging	0.9237	0.5862	0.04885
GBDT	0.9018	0.6808	0.03769

The table indicates that among four models, GBDT exhibits the highest R^2 score on the testing set of 0.6808 and the lowest MSE of 0.03769. In contrast, the training set performance demonstrates notable proficiency across all models, except for the linear model, with R^2 score exceeding 0.9. Since the performance on the testing set is more crucial when considering the power of prediction, GBDT and Random Forest is more convincing when making predictions.

In order to facilitate a more straightforward and lucid interpretation, a series of plots shown in Figure 2, Figure 3, Figure 4, and Figure 5 related to predicted values by different models are shown below, compared with true values.

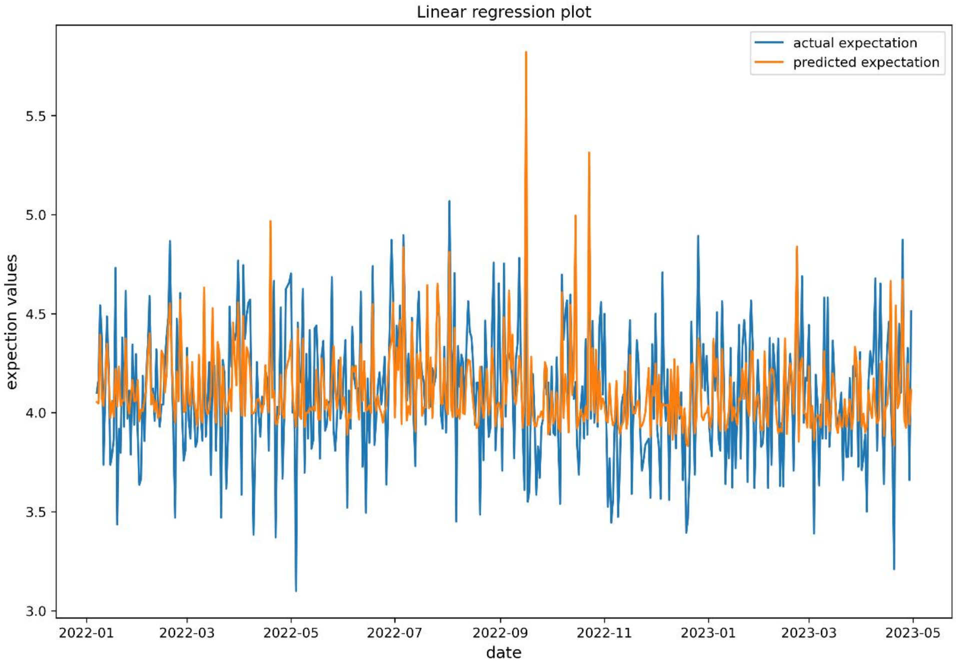

Figure 2. The predicted performance of Linear model.
Photo/Picture credit: Original

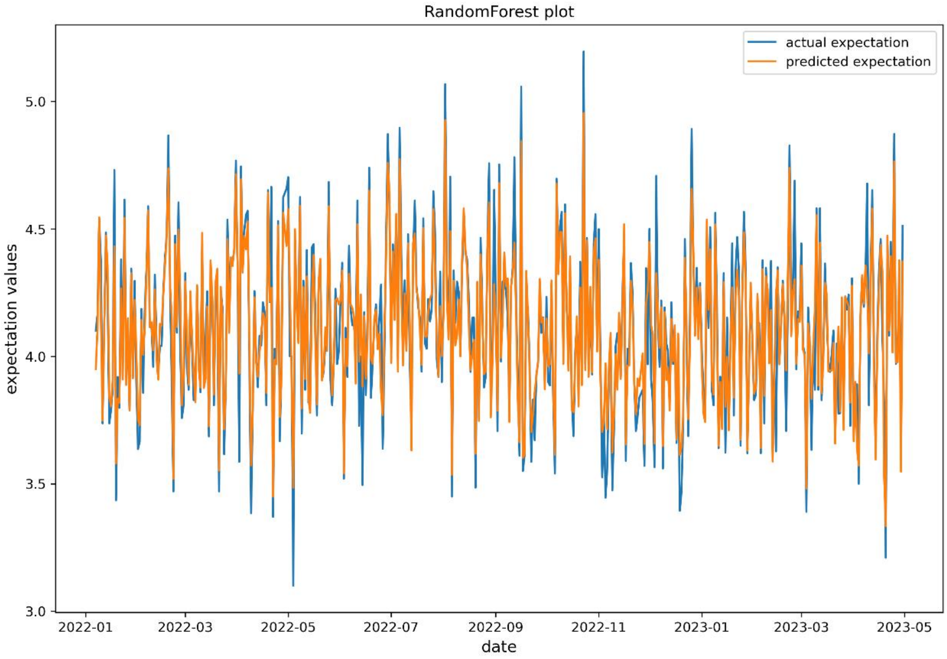

Figure 3. The predicted performance of Random Forest model.
Photo/Picture credit: Original

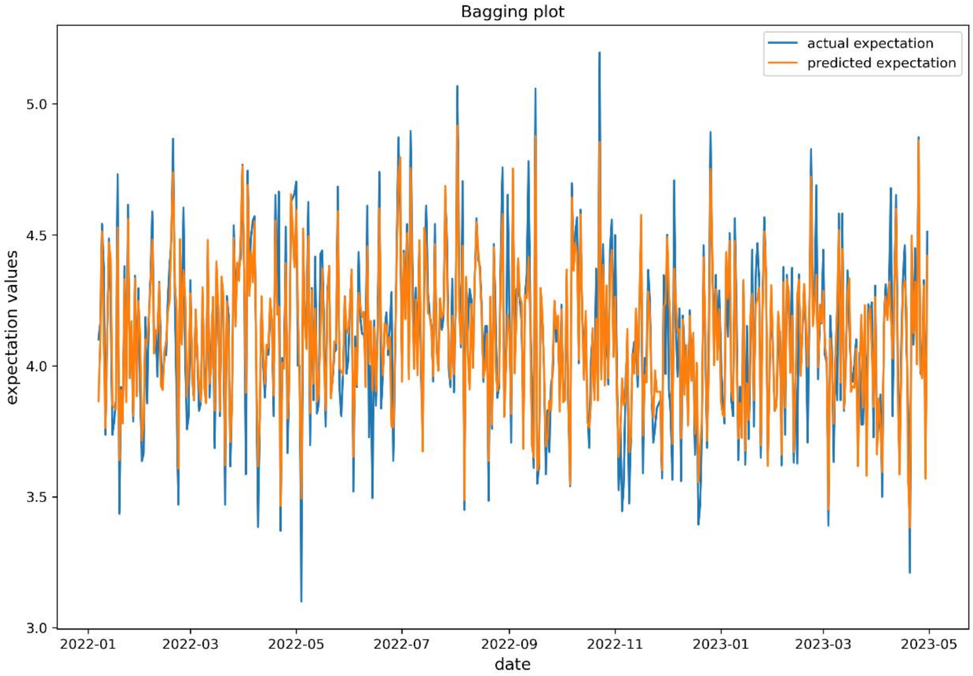

Figure 4. The predicted performance of Bagging model.
Photo/Picture credit: Original

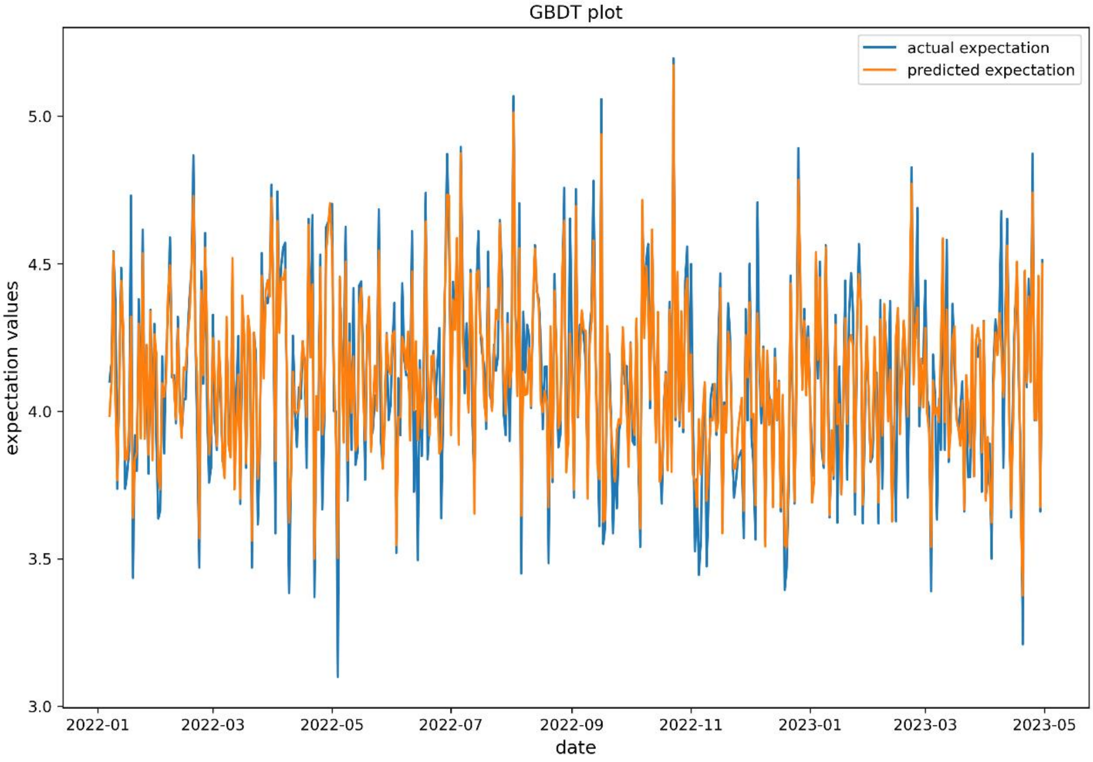

Figure 5. The predicted performance of GBDT model.
Photo/Picture credit: Original

The line charts offer a more comprehensive examination of the models' performance. Upon analyzing these plots, it becomes apparent that GBDT exhibits superior efficacy in handling challenging scenarios characterized by highly difficult or straightforward words, as compared to Random Forest and Bagging. This is reflected in the plots that GBDT fits most peak and valley points well. Moreover, all models present different degrees of delay of prediction, but also GBDT acts to the rapid change of trends better. These may not be sufficient to verify that GBDT is the best model to predict expectation of attempts, but it is better among the four algorithms.

3.2. Feature importance of various models

Based on the model result, a comparison of feature importance in statistics and machine learning is conducted, demonstrated by Table 2. The statistical method is spearman correlations while Random Forest and GBDT are chosen as machine learning methods since they have better prediction performance.

Table 2. Feature importance and correlation (to expectation) of statistical methods and machine learning models.

Features	Spearman correlation with expectation	Random Forest	GBDT
Success rate	-0.761	0.625	0.662
Repeat frequency	0.436	0.054	0.081
Position 1	-0.051	0.037	0.040
Position 2	0.002	0.029	0.028
Position 3	0.098	0.045	0.037
Position 4	-0.092	0.040	0.038
Position 5	0.088	0.043	0.043
Consonants	0.007	0.006	0.001(2)
Vowels	-0.007	0.007	0.001(1)

The table provides evidence that the distinction emerges at the repeated frequency, where all variables share the common dominant feature of success rate. Nevertheless, the statistical correlation associated with the repeated frequency is 0.436, whereas it diminishes to 0.054 and 0.081 in machine learning models, respectively. This is due to the basic algorithm of those models, since they tend to choose the dominant features and then sharply decrease the significance of other features to guarantee the accuracy of predicting. In addition to the decreasing importance of repeating letters, there is also difference in the features of position letters. They present a lower variance in machine learning approaches, which seems that all positions have no significant difference. Nevertheless, there is a big gap between positions in statistical analysis. The inverse situation appears for vowels and consonants, they are same important because of symmetry, but it is not the case in fitted models. All above shows that indeed a loss of feature meanings when fitting models. These behaviors can weaken the convincing of prediction models to different degrees, and further affect the accuracy of proposed strategies using these models.

4. Conclusion

In this study, relations between word features and this true expectation from Wordle’s recorded data have been discovered. The comparing analysis between statistics and machine learning models is also held in order to investigate whether there is a loss of feature meaning when using these technologies. To achieve this goal, machine learning models are, and feature engineering technologies are used. Study results demonstrated that the Gradient Boosted Decision Trees model has better performance in predicting expectations. Study results also presented that some of the features lose or diverge from its original effect in the fitted models, which might weaken the power of predicting. In the future, a further study can be designed to investigate how the loss of meaning of features can affect the accuracy of machine learning models in previous strategies. Additionally, more feature relations will be added to complete this research, providing more results of these analysis, offering a broaden view.

References

[1] Short, M. B. (2022) Winning Wordle Wisely—or How to Ruin a Fun Little Internet Game with Math. The Mathematical Intelligencer, 44(3), 227-237.

[2] Anderson, B.J., & Meyer, J.G. (2022) Finding the optimal human strategy for Wordle using maximum correct letter probabilities and reinforcement learning. arXiv.

[3] Gu, W., Foster, K., Shang, J., & Wei, L. (2019) A game-predicting expert system using big data and machine learning. Expert Systems with Applications, 130, 293-305.

[4] Lahiri, A., Shah, N., Agarwal, S., & Nandakumar, V. (2023) Deterministic Algorithmic Approaches to Solve Generalised Wordle. arXiv.

[5] Hamkins, J. D. (2022) Infinite Wordle and the Mastermind numbers. arXiv.

[6] De Silva, N. (2022) Selecting Seed Words for Wordle using Character Statistics. arXiv.

[7] Bonthron, M. (2022) Rank one approximation as a strategy for Wordle. arXiv.

[8] Cunanan, M., & Thielscher, M. (2023) On Optimal Strategies for Wordle and General Guessing Games.

[9] Bhambri, S., Bhattacharjee, A., & Bertsekas, D. (2022) Reinforcement Learning Methods for Wordle: A POMDP/Adaptive Control Approach. arXiv.

[10] Lokshtanov, D., & Subercaseaux, B. (2022) Wordle is NP-hard. arXiv.

[11] Rosenbaum, W. (2022) Finding a Winning Strategy for Wordle is NP-complete. arXiv.

[12] Liu, B., Zhang, Y., & Zhang, S. (2023) Explore the difficulty of words and its influential attributes based on the Wordle game. arXiv.

[13] Liu, C. (2022) Using Wordle for Learning to Design and Compare Strategies.

[14] Hamner, B. (2022) Wordle Tweets, A daily sample of Wordle results tweets. Kaggle, https://www.kaggle.com/datasets/benhamner/wordle-tweets.

[15] Li, J., Pan, S., Huang, L. (2019) A machine learning based method for customer behavior prediction. Tehnički vjesnik, 26(6): 1670-1676.

[16] Yu, Q., Chen, P., Lin, Z. et al. (2020) Clustering Analysis for Silent Telecom Customers Based on K-means++, 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE,1: 1023-1027.

[17] Zhang, H., Shi, Y., Tong, J. (2021) Online supply chain financial risk assessment based on improved random forest. Journal of Data, Information and Management, 3: 41-48.

[18] Gareth, J., Daniela, W., Trevor, H., et al. (2013) An introduction to statistical learning: with applications in R. Spinger.