ISSN 2616-5805 Vol. 4, Issue 4: 17-21, DOI: 10.25236/AJMS.2023.040403

Report on the Analysis and Prediction of Wordle Data Based on the SIR Model

Hangyu Zeng

School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, 102206, China

Abstract: *Wordle is a New York Times crossword puzzle that has become very popular recently. In this paper, a SIR infectious disease model was developed to explain the reasons for the variation in the number of daily reported results and to predict the number of future reports. By building the SIR contagion model, this paper explains that the main reason for the change in the number of daily reports over time is Twitter publicity, and predicts that the number of reported results on March 1, 2023 will be approximately 9411. Further, by analyzing the correlation and feature importance ratings from decision tree regression, we conclude that words with common letters and high frequency are more likely to be guessed by players with fewer guesses, whereas words with repeated letters were less favorable. Also, the presence or absence of common letters in words had the most significant effect on the difficulty of the game.*

Keywords: *Wordle, SIR Infectious Disease Model, Decision Tree*

1. Introduction

Crossword puzzles have a long history and a wide audience in English-speaking countries and are expressed in many forms, of which crossword puzzles are the most iconic. Wordle is a crossword puzzle that has recently become very popular in the New York Times crossword app. The app incorporates the New York Times' extensive portfolio of puzzles and guessing games, a type of word puzzle.

Wordle debuted in October 2021 and has quickly become a global phenomenon. The addictive game, which can only be played once a day, has become a daily obsession for many people. According to the New York Times, 10% of active players worldwide have played 145 or more Wordle games as of July 2022. And it was recently reported that Wordle has brought in tens of millions of new subscribers to the New York Times.

As statistical reports are to be disclosed to the public, the New York Times' data analysis of game results needs to be accurate and efficient. It was therefore particularly important to build a model to analyses and predict the results of the report. We have selected data from https://www.comap.com/contests/mcm-icm. Regarding the rules of Wordle, they seem to have been greatly simplified in terms of difficulty and format compared to crossword puzzles. The game is updated daily and the player's only objective is to guess a five-letter word in six attempts. The game interface is an array of 5 x 6 squares. When the player enters a guess via the keyboard below, the game colors the letter squares to show the accuracy of the guess.

- Green: the letter appears in the answer and is in the correct place.

- Yellow: the letter appears in the answer but not in this position.

- Grey: the answer does not contain the letter.

- The player then continues to try based on the hints given to him until he guesses the correct answer, or until he has used up his six chances.

At the end of the game, the player is given a tally sheet showing the correctness of each move, the number of days played in a row and a countdown to the next online game.

As the number of participants in Wordle grows, on the one hand, the New York Times needs to ensure the accuracy and efficiency of the existing data analysis; on the other hand, it should build on the existing data analysis to provide the New York Times with ideas for the design of future games and reduce fluctuations in participation results, which are contributed to by different external factors, such as the

ISSN 2616-5805 Vol. 4, Issue 4: 17-21, DOI: 10.25236/AJMS.2023.040403

nature of the words, personal characteristics, etc.

2. The basic fundamental of SIR models

The susceptible infection recovery model is a common mathematical model to describe the spread of infectious diseases, which divides the population into the following three categories." As the number of reported results varies over time and tends to rise sharply to its highest value and then fall to a stable level at the beginning of the game's release, we believe that the main influences on this trend are official and player publicity for the following reasons.

Firstly, after seeing word-of-mouth publicity on social media, many people will become players with the intention of trying the game, mostly out of curiosity rather than any real interest in the content genuinely interested. These new players will tweet and attract more people to try the game, which in turn will drive the number of players to continue to grow. Over time, the number of players will stabilize, and these loyal players are those who are genuinely interested in the content itself, and those who see the game as a daily pastime [1-3].

Secondly, aspects such as the number of tweets posted by players and the difficulty of the text will also cause small fluctuations in the overall trend in the number of players per day [4]; these factors will have an impact on the number of people playing the game, but are not the main influencing factors during the game's popularity [5-6].

Therefore, we treat the spread of games among internet users as a typical infectious disease transmission process and use this to model the SIR epidemic. Tweets posted by players announcing their scores are considered 'viral'; users who have not played the game are considered 'healthy'; current gamers are considered 'infected', a group that can Those who used to be gamers and now choose not to play are considered 'recovered' and are less affected by the tweets in a given future time period. These groups are less likely to be influenced by Twitter propaganda and less likely to become gamers again than those who are 'healthy'. A simultaneous search for the parameters NPR, OPR, LR, and TR*TSR in the interval [0, 1] yielded the best fit of the model to the original data, at which point NPR was approximately 0.23, OPR was approximately 0. 11, and LR was approximately 0. 18 [7].

The number of users not playing, the number of players reporting results and the number of users no longer playing (denoted here by S, I and R respectively) were taken to be the moment of 0 based on the results of the parametric search over time, and similar to the number of days thereafter, as 7 January 2022. This is shown in the graph. It can be seen that the number of players shows a clear trend of rising and then falling until 200 days, and reaches a maximum around 50 days. After 200 days, the number of current players and the number of groups not playing and no longer playing gradually stabilizes [8-10].

$$\frac{dS}{dt} = S(t-1) - I(t-1) \times TR \times TSR \times NPR \times S(t-1) / AU$$

$$\frac{dI}{dt} = I(t-1) + S(t-1) - S(t) + R(t-1) - R(t)$$

$$\frac{dR}{dt} = R(t-1) + I(t-1) \times LR - I(t-1) \times TR \times TSR \times OPR \times R(t-1) / AU$$

$$S(0) = S_{0.} \quad I(0) = I_0$$

$$AU = S_0 + I_0$$

3. Results

3.1 Data simulation

Simultaneous searches of the parameters NPR, OPR, LR, TR*TSR in the interval [0, 1] yielded the best fit of the model to the original data when NPR was about 0.23, OPR about 0. 11, and LR about 0. 18, as shown below in Fig 1.

ISSN 2616-5805 Vol. 4, Issue 4: 17-21, DOI: 10.25236/AJMS.2023.040403

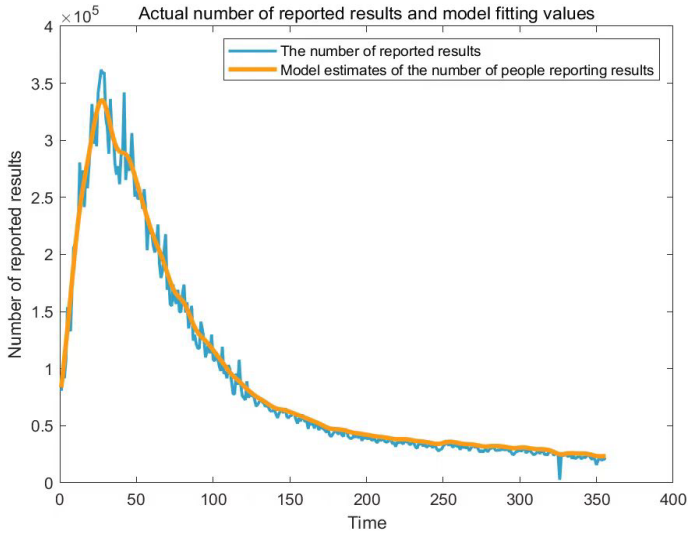

Figure 1: Actual number of reported results and model fitting values

3.2 Analysis of experimental results

The number of users who have not played the game, the number of players who report the results and the number of users who no longer play the game (here denoted by S, I, R, respectively) are plotted over time ac- cording to the parameter search results, and the moment of 0 is taken as January 7, 2022, and analogously with the number of days thereafter. This is shown in the figure. It can be seen that the number of players shows a significant upward and then downward trend until 200 days ago and reaches a maximum at about 50 days. After 200 days, the number of current players as well as the number of unplaced and no longer playing groups gradually stabilizes. The results are similar to the original data and the model is reasonable, as shown in Fig. 2.

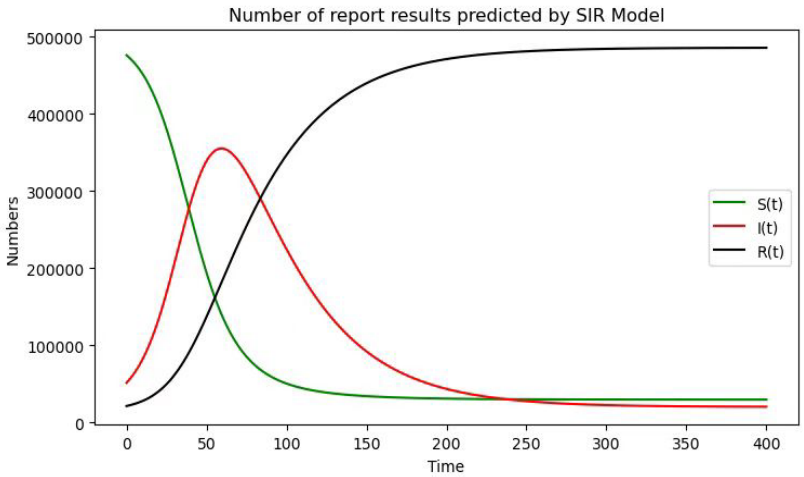

Figure 2: Number of report results predicted by SIR Model

Therefore, in response to the change in the number of reported results, we believe that the promotional effect of posting tweets is the main reason for the change in numbers. Among them, about 23 percent of users who have not played the game will become players after seeing the tweets, 18 percent will stop playing the game for various reasons, and 11 percent of those who no longer play the game will become players again. Using the SIR infectious disease model constructed in this paper, we predict the number of outcomes reported on March 1, 2023 to be 9410.67, with a confidence interval of (8553.43, 11132.59).

Based on the above-constructed metrics, this paper first analyzes the correlation between the selected metrics and the percentage of each attempt in the score report, from which the direction of the influence of different attributes of words on the number of user guesses is derived. The results are shown in the figure 3 below.

ISSN 2616-5805 Vol. 4, Issue 4: 17-21, DOI: 10.25236/AJMS.2023.040403

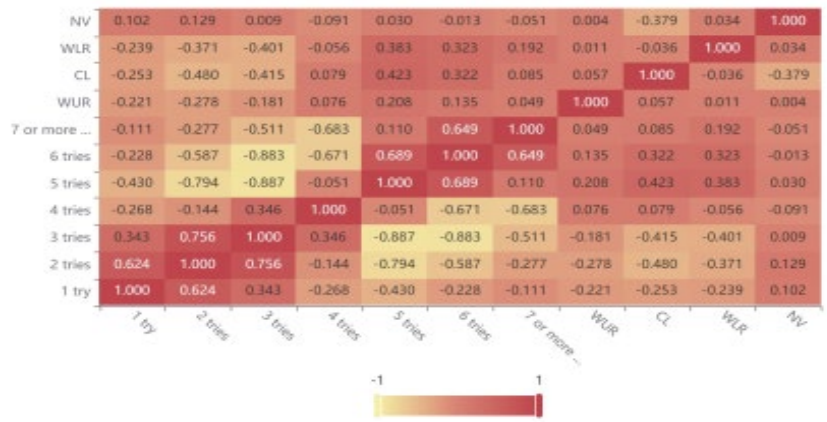

Figure 3: Correlation coefficient between variables

As can be seen in Figure 3, the absolute value of the correlation between the number of vowel letters in a word and the number of different attempts is less than 0.15, and this variable has very little effect on the difficulty of the player guessing correctly. In terms of frequency of word use, WUR (the higher the frequency of word use, the smaller the WUR value) is negatively correlated with the proportion of guesses made in 1-3 attempts, and it is clear that players are more likely to guess common words correctly. In terms of how common the letters of the word are and whether the word contains repeated letters, both CL and WLR were negatively correlated with a smaller percentage of guesses, suggesting that words containing more common letters (e, t, a, etc.) were more likely to be guessed, whereas for words containing repeated letters, players were less accurate in guessing the word due to the extended range of letter positions and possible psychological factors.

Based on the results of the correlation analysis above, we know that the number of vowels in a word has no significant effect on the difficulty of guessing, while common words, common letters, and words that do not contain repeated letters all reduce the difficulty of players' guesses. However, the relative effects of different factors cannot be directly derived from correlations alone. For example, words containing common letters (CL) and words containing repeated letters (WLR) were similar in terms of the percentage of times these two factors were correlated, both between 0.25 and 0.5. Therefore, in order to compare the extent to which the different factors chosen influenced the percentage of scores under the difficulty model, we used a decision tree regression model to compare the relative importance of the different features. We substituted the number of attempts from 1-6 and the percentage of people who could not solve the puzzle as dependent variables into the model to obtain the importance of each feature in the different cases, as shown in Figure 4 below.

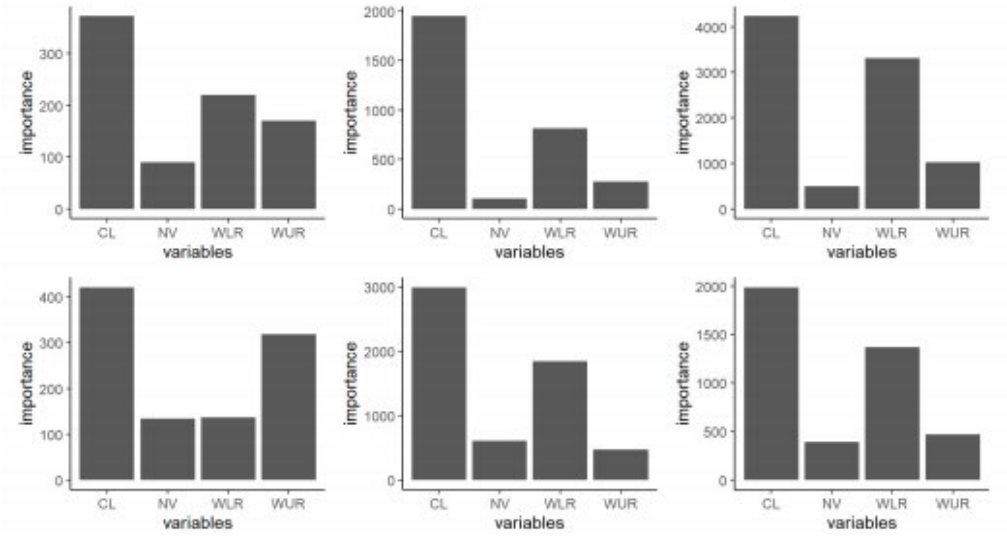

Figure 4: The decision tree regression model compares the relative importance of different features

ISSN 2616-5805 Vol. 4, Issue 4: 17-21, DOI: 10.25236/AJMS.2023.040403

4. Conclusions

In this paper, we explain the changes in the number of reports by building a SIR contagion model analogous to the fluctuations in the number of players and predict the number of reported outcomes on a given date. For the analysis of word attributes, we analyzed correlations to obtain that factors such as common words and common letters have a positive effect on reducing the difficulty of player guessing, and constructed a decision tree regression model to conclude that words containing common letters have the most significant effect on the proportion of reports.

References

[1] Yang Chaobo, Xie Weihong, Wang Lizang. Research on optimization and intervention of SIR model for online public opinion [J]. Frontiers of Data and Computing Development, 2023, 5(01): 115-127.
[2] Qu Y, Zhai JW. Analysis of project implicit knowledge transfer factors based on improved SIR model [J]. Industrial Engineering, 2023, 26(01): 146-152.
[3] Zheng M. M., Zhang W. N., Liu Y., Wu X. T. Research on the propagation process of pedestrian crossing violation based on migration-based SIR model [J]. Journal of Dalian Jiaotong University, 2023, 44(01): 53-57+63.
[4] Yang JH, Li X, Liu H. Diagnostic efficacy of serum sTREM-1 combined with SIRS score for sepsis in burn patients with complications [J]. Chinese general medicine, 2023, 21(02): 234-237.
[5] Chen J, Xiong Y, Tong J, et al. Analysis and prediction of COVID-19 in the US based on the time-varying parameters SIR model[J]. Journal of Physics: Conference Series, 2020, 1678(1):012082 (8pp).DOI:10.1088/1742-6596/1678/1/012082.
[6] Sang R, Zhang L, Wu H. A two-patch SIRS contagion model for media-induced migration rate change [J]. Journal of Xinjiang University (Natural Science Edition) (in English and Chinese), 2023, 40(01): 49-56+60.
[7] Zhao Yanjun, Sun Xiaohui, Su Li, Li Wenxuan. Qualitative analysis of stochastic SIRS infectious disease models with logistic growth and Beddington-DeAngelis incidence [J]. Journal of Mathematical Physics, 2022, 42(06): 1861-1872.
[8] Zhao Xiaoqiang, Luo Weilan, Liang Haopeng. Bearing fault diagnosis based on SIR multilevel residual connected dense network [J]. Journal of Lanzhou University of Technology, 2022, 48(06): 46-54.
[9] Cang Linqing. Modeling and research of Si-SIR rumor propagation model on complex social networks [D]. Nanjing University of Posts and Telecommunications, 2022.
[10] Chen Wenhao, Cui Ruiwen, Liu Mengna. Research on the cross-contagion mechanism of risk among banks in China - based on SIRS contagion model [J]. North China Finance, 2022, (11): 24-34.