2023

Mathematical Contest In Modeling®

Certificate of Achievement

Be It Known That The Team Of

Nathan Schambach
Bryce Drynan
Brechan Allison

With Faculty Advisor

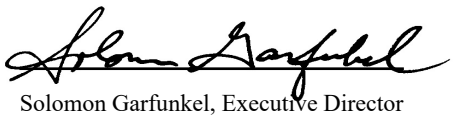
Daniel Cicala

Of

Southern Connecticut State University

Was Designated As

Successful Participant



Solomon Garfunkel, Executive Director

Administered by

COMAP

With support from

AMS ASA informs MAA MATHEMATICAL ASSOCIATION OF AMERICA siam

Steven B. Horton, Contest Director

Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2322659

Abstract:

Wordle is an incredibly popular five-letter word guessing game. Here we present our approach modeling how many people submit their scores on a given day, what the distribution of how many tries it took for a given word, and how to classify the words based on difficulty. Our methods utilize multiple different regression techniques as well as a K-Nearest-Neighbors classifying algorithm. We found that a logarithmic regression curve can model the predicted score submission amount and that on March 1, 2023, we predict there will be 14650 people who submit their scores to Twitter. We found that a normal distribution is best for predicting the amount of tries on a given day. We were able to use linear regression as well as K-Nearest-Neighbors to predict the likelihood of a given try category. Additionally based on the attributes on the word “eerie” will be a medium difficulty word. We also discuss many refinements that could be made to our model such as increasing the number of features extracted from a word and using different distances in the K-Nearest-Neighbors model.

Table of Contents:

Introduction

Background

Wordle is a word guessing game developed by software engineer Josh Wardle for him and his partner to play. After playing Wordle for a few months, and it taking over his family group chat on WhatsApp, Wardle would go on to release the game to the public in mid-October 2021. On November 1, 2021, the game had 90 players, and by the time January 2, 2022, rolled around 300,000 people had played the game.

On January 31, 2022, the New York Time acquired Wordle, where it is currently offered as a daily puzzle. Players play the game by trying to solve for a five-letter word. Per the directions on the New York Times' website, every player gets six guesses to get the word, getting feedback with every guess. Every guess must also be a valid English five-letter word. The feedback is provided by each guessed tile changing color. If the tile turns grey, it means that the letter is not in the word. If the tile turns yellow it means that the tile is in the word, but in the wrong spot. Finally, if the tile turns green it means that the letter is in the word and in the right spot. **Figure 1** is an example solution that was guessed in three tries from July 21, 2022.

Figure 1: Example Solution of Wordle Puzzle from July 21, 2022

If the regular version does not provide enough challenge, there exists a toggleable “Hard Mode” that can be played. “Hard Mode” makes the game more difficult by making it so that if you guess a letter and it turns yellow or green, you are required to use those letters in your subsequent guesses.

Many people who play Wordle post their results to ***Twitter***. The data set provided to us by MCM is a file containing results from January 7, 2022, to December 31, 2022. The dataset includes the date, contest number, word of the day, the number of people reporting scores that day, the number of players on hard mode, and the percentage that guessed the word in one try, two tries, three tries, four tries, five tries, six tries, or could not solve the puzzle.

Restatement of the Question

We have been tasked by the New York Times to do an analysis of the data provided to us and answer 4 questions. The first question has asked us to create a model for an interval of players per day and then estimate the interval for March 1, 2023. In addition, we were tasked with seeing if any attributes of the word affect the percentage of people reporting in Hard Mode on any given day. The next task that we were given asked us to, for a given date and word, develop a model which predicts the distribution of the reported results. Then, use the model to predict the distribution of the word eerie on March 1, 2023. Another task we were given was to develop and summarize a model which could classify a solution word by difficulty, then use the model to rank the word eerie. Finally, we were tasked to report any interesting observations that we have about the dataset.

Global Assumptions

i) Contestants do not cheat by looking up hints or the solution ahead of time. If they did, it would skew the data to less tries

ii) Contestants only report once a day. Reporting your score more than once a day would make the dataset that we were proved with inaccurate.

iii) The valid guess word pool has remained constant over all contest days.

iv) All contestants complete the wordle independent of one another. People cooperating would result in the skewing of the distribution for tries it would take.

v) All contestants understand English.

vi) All contestants play the game optimally.

vii) The more difficult a word, the more tries it will take to guess it.

viii) All contestants know all five letter words

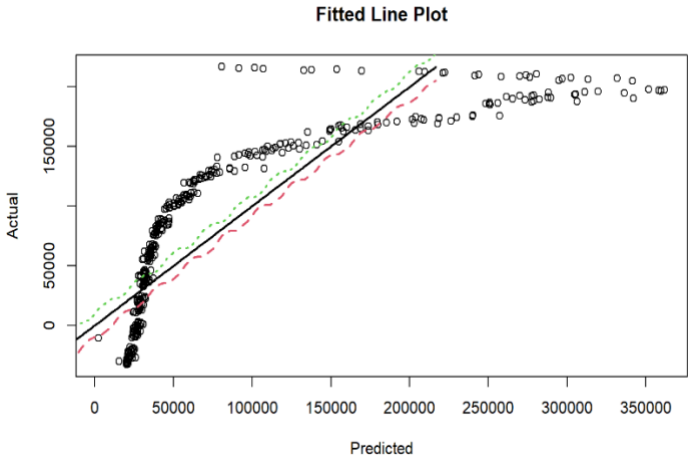
Model 1-Players on any day:

Assumptions

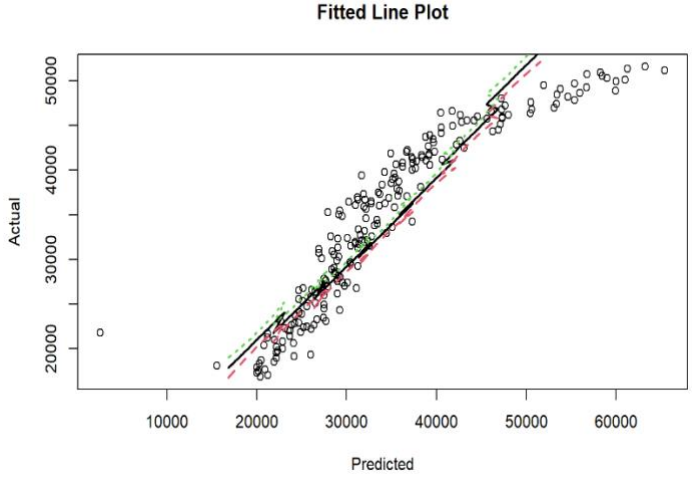
The model can be used to predict results on any day.

Method

We analyzed the data and realized any data before June 1st, 2022 should be considered an outlier. This was determined because there were heavy outliers when plotting the predicted values of our model.



We created our first model by running a linear regression from the date of June 1st, 2022 to December 31st, 2022.



Number of Reported Results = 80281.23 – 4748.80 (Month) – 208.10 (Day)

Adjusted R^2 = 0.85, S= 1.0037

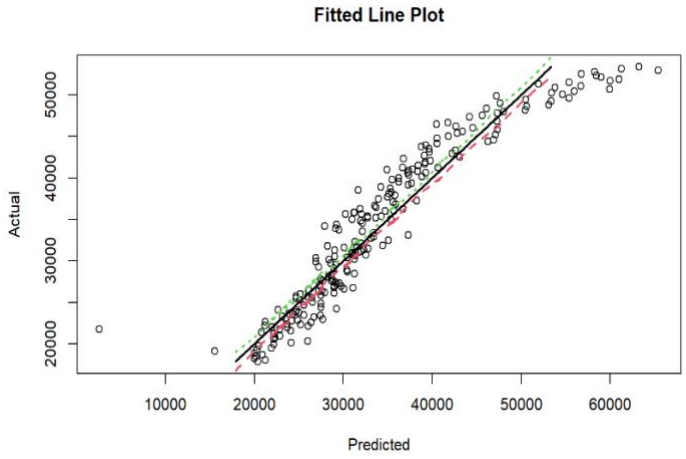
The R^2 value indicates there is a 0.85 or 85% of the variability in the amount of reported results is explained by the linear relationship between months passed since June 1st, 2022 and days in the month. The standard deviation of the residuals was 1.0037.

Based on this model, the predicted number of reported results for March 1st, 2023, is

$$80281.23 - 4748.80(15) - 208.10(1) = 8841.13$$

which in our opinion is an underestimation by the model. We moved onto a logarithmic approach because of the plot of the residuals and what we believe is an underestimation by the model for the number of reported results for March 1st, 2023.

We created our second model by running a logarithmic regression from the date of June 1st, 2022, to December 31st, 2022.



Number of Reported Results = 129293.37 – 42257.48 log(Month) – 207.53(Day)

Adjusted R^2 = 0.889, S = 1.004

The R^2 value indicates there is a 0.889 or 88.90% of the variability in the number of reported results is explained by the logarithmic relationship between months passed since June 1st, 2022 and days in the month. The standard deviation of the residuals was 1.004.

Based on this model, the predicted number of reported results for March 1st, 2023, is

$$129293.37 - 42257.48log(15) - 207.53(1) = 14650.46$$

which we believe is a much more accurate prediction of what the potential number of reports for March 1st, 2023.

Based on common practice we proceeded to acquire a 95% confidence interval for the expected number of reported scores for March 1st, 2023. We are 95% confident that the true number of reported scores for March 1st, 2023 is between 13209.58 and 16091.35. There are multiple reasons to rely on the logarithmic model's output. First, the output of the logarithmic model is more reasonable at 14,650 reported results compared to the linear model's projection of 8,841. Secondly, the R^2 value is higher at 0.899 compared to 0.85. Finally, the actual vs predict shows a logarithmic curve to it.

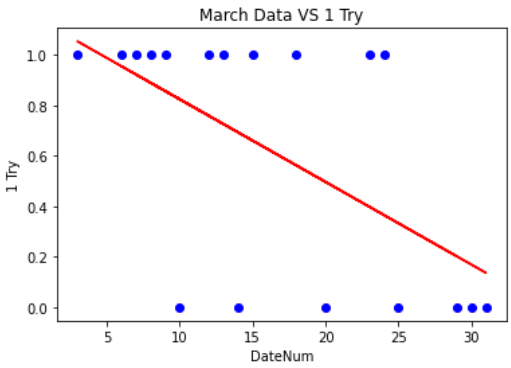
We were also asked to analyze whether any attributes of the word affected the percentage of scores reported that were played in Hard Mode. To find the answer to this we calculated the expected value of each word. Then we found whether there was a correlation between that and the number of reported results for hard mode on a given day. The formula for the correlation coefficient is as follows:

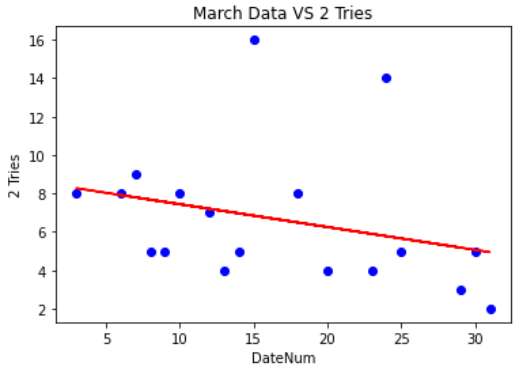
$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

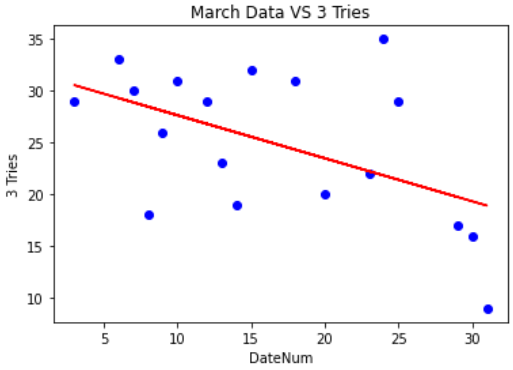
After using the software R to calculate the coefficient, we found that the correlation coefficient was 0.0615. This signifies there is a very weak relationship between the two variables.

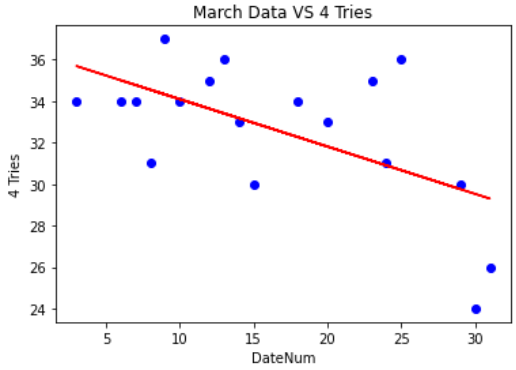
Model 2-Distribution of a word's tries:

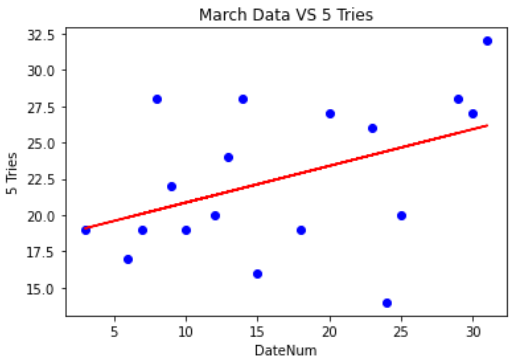
In order to predict the results of the associated values for a given try at a future date, we can use Linear Regression in order to generate a line of best fit for the data.

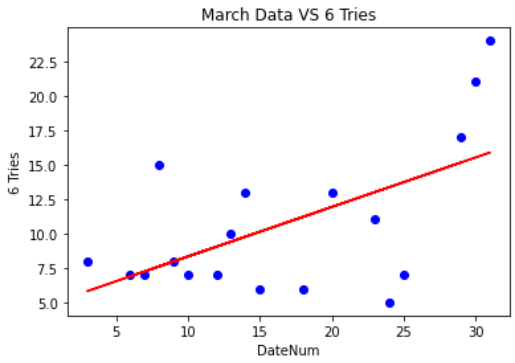


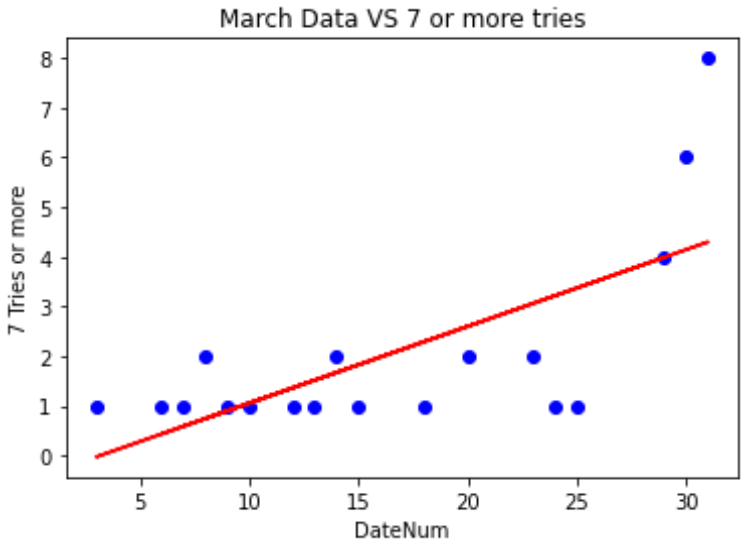






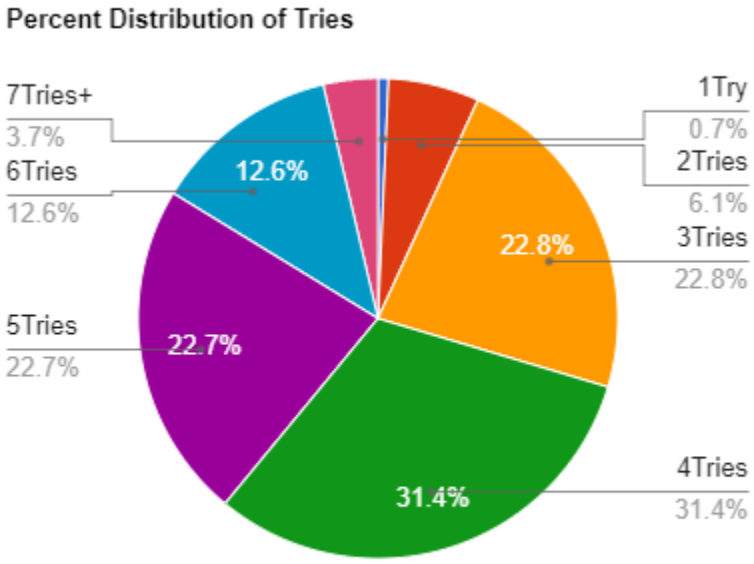




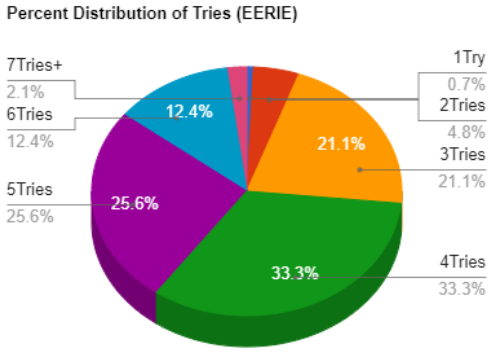


Here we will be using March 1st 2022 as our example for predicting tries. We compared the number of days in the month (DateNum) to the number of tries for that given category. There are some trends that emerge, with the first four categories having a moderate linear negative relationship while tries 5, 6 and 7+ have moderate positive linear relationships. To answer the question of predicting the percentage of tries for a given date, there are many uncertainties with the model. Particularly when only working with one dataset from 2022, there will be no way to compare year over year change in metrics when looking at a monthly difference. So instead, we can create the best model on paper based off of 2022 and use that to

predict for 2023.



Given the Linear Regression for each category of 'try', we have the given percentages in which we would predict a date of March 1st, 2023 would score. More specifically speaking, we can use this model to predict the difficulty.

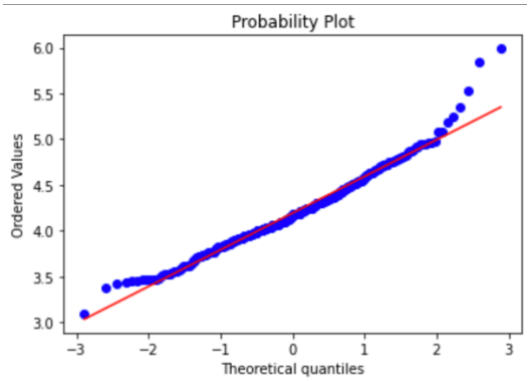


As we will talk about in the next segment, we have classified the word EERIE as a 1, a medium difficulty word. Using this as our basis for prediction, we can see the differences between the tries compared to the previous average prediction. Tries number 4 and 5 have increased but all other tries have decreased in percentage. However due to the very limited size of our sample data, we were only able to generate 57% accuracy in our model.

Model 3-Difficulty of a word:

To create a model for ranking the difficulty of a word we decided to train a K-Nearest-Neighbors model, a machine learning model, to predict whether a word would fall into one of three difficulty categories, 0 being the easiest, 1 being medium, and 2 being the hardest words to guess.

To create these difficulty categories, we used the frequency distribution provided for the number of tries it takes to guess a word. We then took the expected value of each distribution to obtain how many tries each solution provided should take. Looking at the expected values of all the solution words we can see that they are normally distributed.



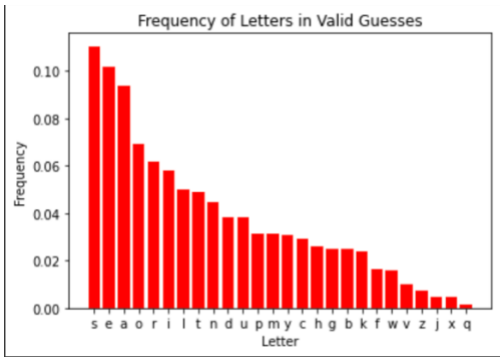
QQ-plot of the expected values

The mean of the expected values is 4.1923 tries, with a standard deviation of 0.4037 tries. We assumed that most words should be of the medium level so using the mean and standard deviation we made easy difficulty an expected value of 1 try up to 3.7886 tries (mean – standard deviation), medium is an expected value of 3.7886 tries up to 4.5960(mean + standard deviation) tries and finally hard from 4.5960 tries up to 6 tries.

The attributes of solution words that we chose to look at were the frequency of each letter in its position, the frequency of each letter in a five-letter word in all positions, the ratio of vowels to consonants, whether the word has more than one of the same consonant, and whether the word has more than one of the same vowel.

Letter Frequencies

We decided to use a dictionary of all five letter words that wordle accepts as guesses, as of January 15, 2023, to be what we based our letter frequencies from. For the frequency of each letter, we looked at how many times it appeared in any word, in any place to get the result below.

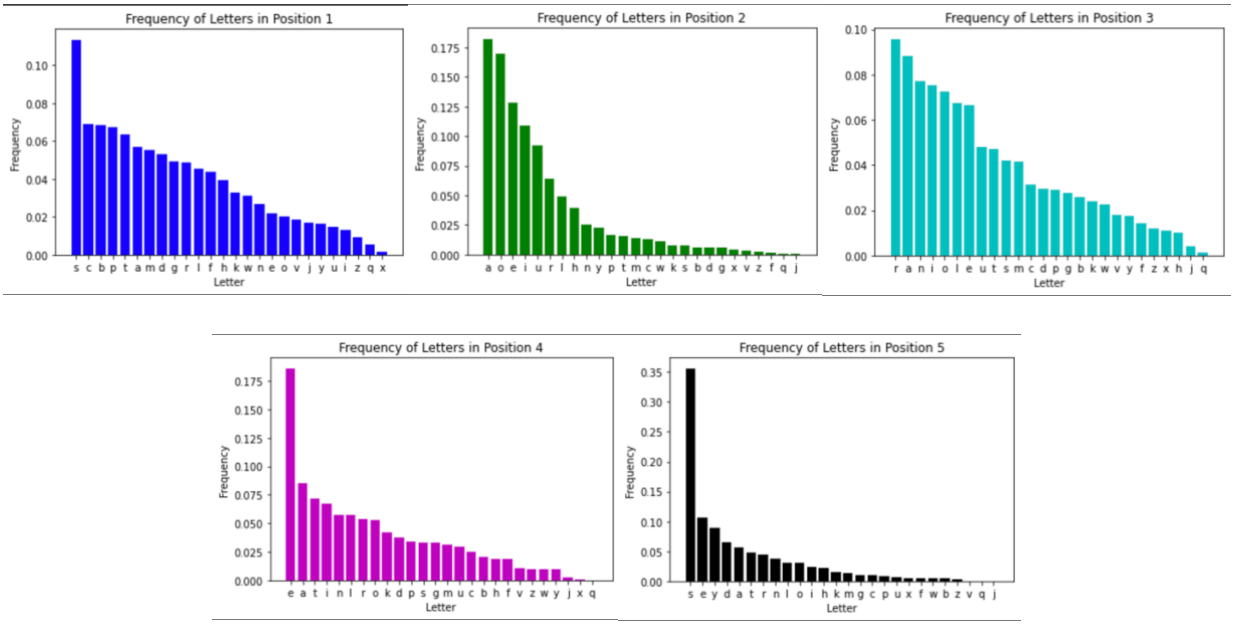


Frequency of letters in valid guesses

To then calculate the score of a word we summed the frequencies.

$$freq_{any\ position} = f_{l1} + f_{l2} + f_{l3} + f_{l4} + f_{l5}$$

Where l1, l2, l3, l4, l5 are the letters at positions 1, 2, 3, 4, 5. To calculate the per letter frequencies, we broke up each word by letter, then looked at how often a specific letter appeared in a specific letter slot. The frequencies by letter position are listed below.



Frequencies by letter position

To calculate the score for a given word we summed the frequencies for each position.

$$freq_{in\ position} = f_{l1\ in\ p1} + f_{l2\ in\ p2} + f_{l3\ in\ p3} + f_{l4\ in\ p4} + f_{l5\ in\ p5}$$

Analysis of Model 3

To test our model, we have decided to rank the word “eerie”. The word’s characteristics are provided below.

total_freq_score	position_freq_score	>1 of same Consonant	>1 of same Vowel	Ratio
0.42529	0.420025	1.0	0.0	4.0

Attributes of the word eerie

Where total_freq_score is the frequency score for a letter in any position, position_freq_score is the frequency score for letters in their position, and Ratio is the ratio of vowels to consonants. Our model believes that “eerie” has a difficulty rating of 1 which means that it has medium or average difficulty.

Since our model is a supervised learning model, we can get more information on its ability to classify the difficulty of words correctly. Our model was able to achieve on average an accuracy of .6421 or 64.21% of the time it will correctly classify a word into its correct difficulty level.

Strengths and Weaknesses of our Models

Model 1

The first model was designed to explain the future reported results on a given day. We believe that the model successfully does that as 14,650 is a reasonable projection. We were asked to produce a prediction interval for the given day of March 1st, 2023, and we also believe that the model does that with a confidence interval between 13209.58 and 16091.35. The model does not just do this for March 1st, 2023, but can do this for any future given date. Furthermore, the model could produce a correlation coefficient producing sufficient evidence that there is a very weak relationship between the attributes of the words and the number of reported results on Hard Mode.

The first model could have been improved in multiple areas. We believe, while reasonable, that the model is under-predicting the number of reported results for March 1st, 2023. This is due to when we used the model’s equation for December 31st, 2022, it underestimated the reported results by around 1000. The model had a large portion of the

given dataset removed under the consideration that they were outliers, and we believe that is why it is underestimating values. The model is only reliant on data from June 1st, 2022, to December 31st, 2022. Also, we would have liked the logarithmic model's line of best fit to have more of a curve to it to follow the predicted versus the actual more. The visual display of what the model is, is not exactly what we were looking to produce.

Model 2

The accuracy of Model 2 suffered due to lack of previous years data as well as small sample size. The more granular we defined 'difficulty' the better the outcome of this model would have been; however this effort would have been of no use given we only had 14 values for 'medium difficulty' during March 2022. Our results are as expected, predicting a medium difficulty word increased tries 4 and 5's percentages while every other percentage dropped. Given the amount of tries 6 and 7 for more difficult '2' rated words, this would explain the near outlier status of those tries, and how it affected the word EERIE's predicted percentages.

Model 3

The accuracy of our model sits at 64.21% which is too low to consistently correctly rank new words. This is its largest weakness and what primarily prevents it from use. To fix this issue we suggest that moving forward more features are derived or found from the solution words. Some other features that we noticed were, does the word begin or end with a vowel or consonant. However, the model's strength is that understanding where a word falls in the

classifier. Since the ranking is based off of the expected value of tries, we know an easy word is 3-4 tries, a medium word is 4-5 tries and a hard word is 5-6 tries. Another weakness of the model is it assumes that every player has a perfect knowledge of the human language which can be a large stretch. Another weakness is that the model becomes no better than random assignment if the distribution of tries it takes to get a solution were skewed, this is because the targets rely exclusively on the expected value of tries.

Dear New York Times Puzzle Editor,

Wordle is a popular game amongst all age groups. However, the number of reported results would suggest that the game is losing attraction. The projected number of reported results for March 1st, 2023, is 14,650. All we need to do is compare it to the 240,137 reported results from March 1st, 2022, to see that the game is trending downward. We do take into consideration that our model may be underestimating the true value of the reported results for March 1st, 2023.

The correlation coefficient of 0.06 shows there is a lack of relationship between the attributes of a word and the number of results reported on Hard Mode. It would be suggested by this group that further research be done into the promotion of the game. If the number of results is dropping and it is not based on the word choice, then more research must be done and other models must be built.

We were able to predict the values for each try for the word EERIE for a given date, and used March 1st as an example. However, the performance of the prediction model was hampered by many factors including small sample size and lack of year over year data.

While we were able to classify the words and scale them on a 0 through 2 scale, we would not suggest using our model for the prediction of how difficult a word is. Its likelihood to produce false results is too high to suggest its use. However, I do recommend classifying the difficulty of results on an easy, medium, and hard scale as that is approachable and easy to understand for just about everyone. I would suggest taking a deeper dive into how unique a word is in terms of its use in the English language not just the frequency of its letters.

References:

Wordle logo from The New York Times website. Accessed on February 20, 2023 at

https://nytco-assets.nytimes.com/2022/08/cropped-Screen-Shot-2022-08-24-at-8.49.39-AM.png.

"Wordle Stats." Twitter, July 20, 2022.

"wordle-allowed-guesses.txt" Accessed on February 18, 2023
https://gist.github.com/cfreshman/cdcdf777450c5b5301e439061d29694c