Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2321082

Abstract

In this report, we propose a novel way for generating predictions for the distribution and the Number of Total Reported Results shared by Twitter users for the Wordle puzzle presented by the New York Times. The main challenges for the problem we examined are: Creating a prediction interval for the Number of Total Reported results for a given future date, predicting the associated percentages for the Reported Results, and developing a model that is capable of classifying solution words by difficulty. The predicting Wordle results problem essentially tries to use results from prior puzzle contests in order to generate analysis regarding word attributes for the given solution and predictions concerning the Total Number of Reported Results.

We assumed that the data used in our analysis is real-world data for the given puzzles from January 7th, 2022 through December 21st, 2022. However, given that data recollection is imperfect, we took into consideration performing some exploratory data analysis on the shared data set in order to spot and remove anomalies.

Considering the goal of generating a prediction interval, we performed various iterations of different types of Regression Analysis in order to achieve an accurate interval. The models we processed and tested to generate prediction estimates were Poisson Regression, followed by a generalization of the previous model that is based on the Poisson-gamma mixture distribution called Negative binomial Regression. Given the date as input for the model, we worked on creating an extra parameter using an Iterative Curve Fitting process given by a fitted series for the shared data set. After finding an accurate prediction value for the given future date, we generated the interval by getting a lower and upper bound of the prediction value in combination with an error estimate.

We performed a study of several attributes of a word that are significant for the percentage of scores reported that were played in Hard Mode. Using statistical tests to show significance, we found words containing repeated letters, words with no vowels and three vowels, and words containing the least common letters in the English language were significant attributes that contributed to the change in the percentage of hard mode reported results.

Regarding the goal of predicting the distribution of Reported Results, we realized that individual results within the distribution are codependent. Therefore, we decided to create a model that generates output for the mean number of attempts, the standard deviation of number of attempts, and the success rate altogether as a measure of difficulty of a given word; however, rather than use a multivariate regression model, three distinct models were used, one for each desired output. Moreover, three different types of regression models were explored, each using ordinary least squares regression: multiple linear regression, multiple weighted linear regression, and multiple linear regression with polynomial features. The linear regression with polynomial features showed itself to both best fit the data and extrapolate new results based upon new, unknown Wordle contests. Ultimately, these models were used to make an future prediction on the Wordle Contest on March 1, 2023 for the word EERIE.

Contents

Predicting Wordle Results

1 Exploratory Data Analysis

Before considering the idea of having a model that will predict the number of Reported Results given a date, or a model that can predict the distributions of tries of the Reported Results given a word and a date, we need to revise and tidy the data in order to remove future possible outlier implications. [7] During this process, critical processes of performing initial investigations on the data given so as to find patterns, spot anomalies, and test and check assumptions by observing summary statistics, and visual representations.

1.1 Tidying the data

The data received for this problem is a file with daily results from January 7th, 2022 through December 31st, 2022. This file 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 (indicated by X).

The goal before creating any model will be to provide a standardized way to link the structure of a data set (physical layout) with its semantics (meaning). To perform this action we will visualize general trends of data and spot anomalies. These anomalies will be removed from the data. The most predominant features to have into account for this standardized and clean structure are:

- Letter Count for words per contest
- Proportion between Hard Mode results and Total number of Reported Results
- Incorrect distribution of percentages for different word guesses

1.1.1 Letter Count

After performing analysis on the words for each contest given in the data set, these were the anomalies found:

Date	Contest Number	Word
4/29/22	314	tash
11/26/22	525	clen
12/16/22	545	rprobe

Table 1: Letter Count Irregularities

One crucial rule for Wordle puzzles is that the word given is a five-letter word. Hence, as these words shown above do not fulfill this requirement, we have decided to remove them from the given data set.

1.1.2 Proportion: Hard Mode and Total Number of Reported Results

After removing the initial anomalies, it can be observed that throughout the data set the Number of Reported Results varies as well as the Number of Hard Mode Results Reported. Visualizations below.

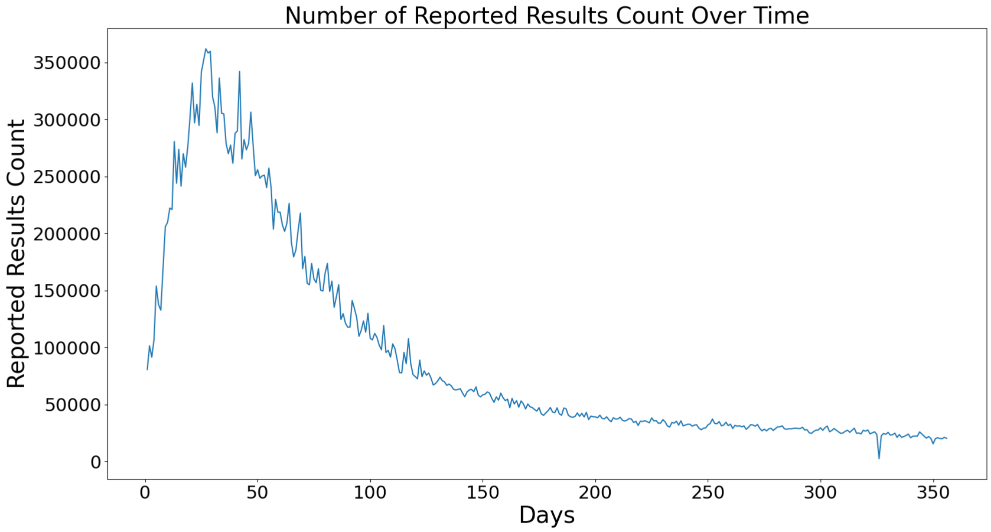


Figure 1: Reported Results Over Time

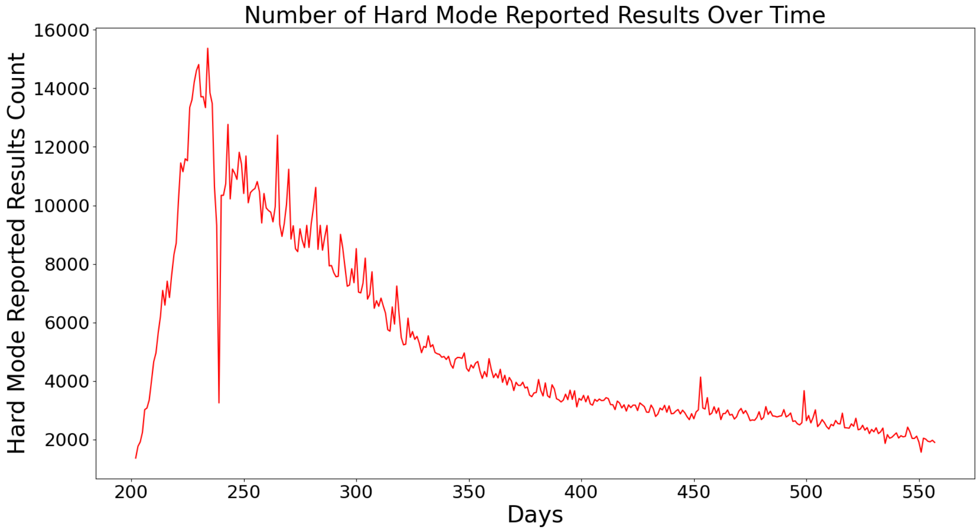


Figure 2: Hard Mode Reported Results Over Time

Displaying the range of the values for Total Reported Results and Hard Mode Reported results can

Results Mode	**Min Value**	**Max Value**
Total Reports	15,554	361,908
Hard Mode Reports	1,362	15,369

Table 2: Range of Results

be seen in the table above. However, one specific attribute that generalizes the relation between the Number of Total Reported Results and the Number of Hard Mode Results is to look at their proportion:

$$Proportion = \frac{HardModeResults}{TotalReportedResults} \tag{1}$$

This can be observed in the visualization below:

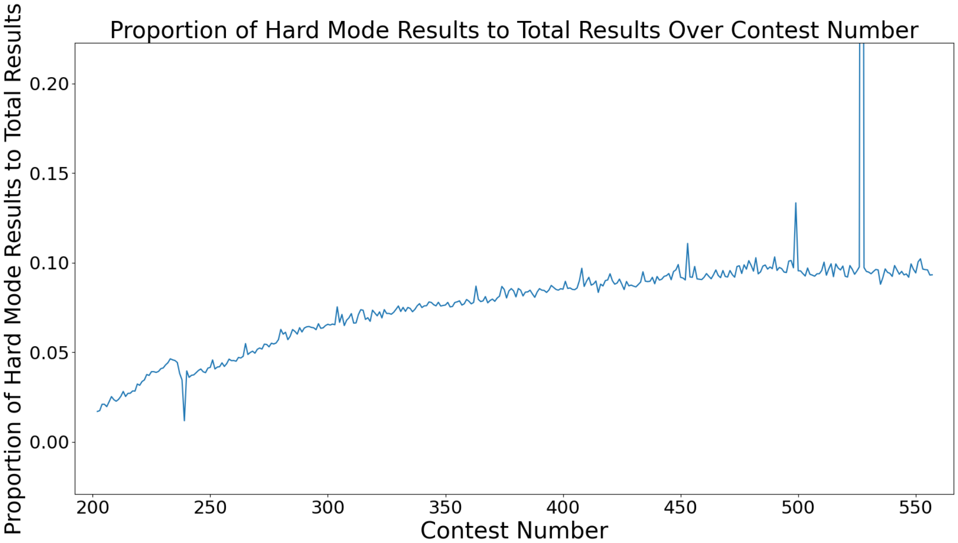


Figure 3: Hard Mode Reported Results Over Time

As can be observed in the given graph, there are three major peaks and drops that can definitely impact the study of our prediction models. Looking at the actual entries for these Contest dates can be seen in Table 3:

Contest Number	**Word**	**Total Reports**	**Hard Mode Reports**
239	robin	277471	3249
500	piney	27502	3667
529	study	2569	2405

Table 3: Outlier Entries

Considering the information in the previous table and graph, for the contest that took place in their #239 version of Wordle, as the proportion plummets, it means that there were significantly fewer players participating in Hard Mode than the previous days. Subsequently, Wordle puzzle #500 shows a significant increase in proportion. This event describes how there was a higher participation in Hard Mode than in days close to this contest. The entry that has the most dramatic proportion value occurs in contest number #529. As can be observed in the graph, the proportion of this puzzle skyrocketed. One important attribute of this entry is the Total Number of Reported Results. This value is very close to the Number of Hard Mode participants.

The entries studied, lead us to think that there can be an error within the values given within the data set. Therefore, we have decided to remove them to prevent future impacts on our predictions.

1.1.3 Incorrect Percentages Distribution

One important aspect of the daily results shared is the percentages of the number of tries it took users to guess the correct word. An important note shared with the data set was that on certain occasions the distributions would not add up to 100%. The bar graph below allows seeing the different values presented for the distributions:

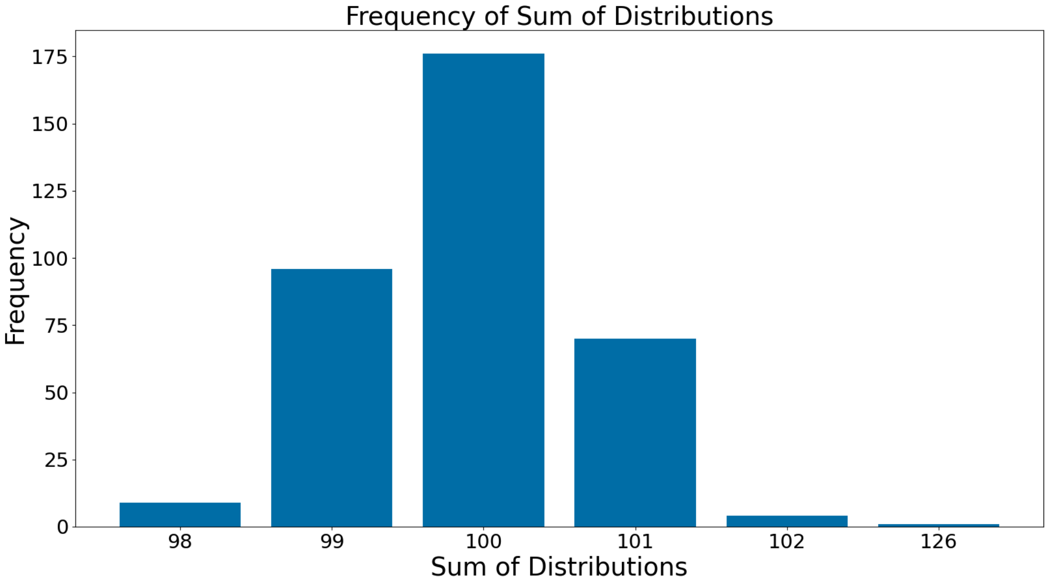


Figure 4: Hard Mode Reported Results Over Time

After reviewing the entry that contained a distribution adding up to 126%, we decided to remove the entry:

Contest #	Word	1 Try	2 Try	3 Try	4 Try	5 Try	6 Try	7+ Try
281	nymph	1	2	18	44	26	26	9

Table 4: Uncommon Distribution

1.2 Prediction Model for Number of Reported Results

One of the primary requirements for our model was to give a prediction interval for the Number of Reported Results given a future date. For more specifics, the goal is to have a prediction interval for March 1st, 2023. After tidying the data, the Number of Reported Results over time can be studied in the graph below:

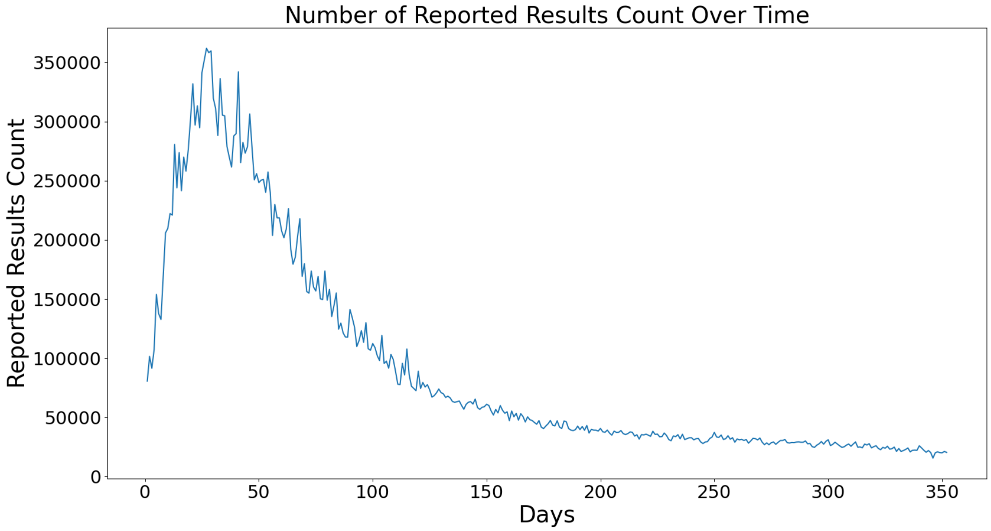


Figure 5: Number Total Reported Results with clean data set

Our initial remark concerning the information observed above is that the data for total Reported Results follows a Poisson Distribution. However, it is important to acknowledge that a Poisson distribution targets the probability of an event happening and not directly how many occurrences of the event will happen. Therefore, in order to use the data given in the study, regression is a technique that allows us to use previous data to model data in the future. Hence, our strategy aligns with implementing a Poisson regression model.

The Poisson regression model is great for the type of prediction we are attempting. However, the Poisson regression model strictly assumes that the Variance equals the Mean[**Sachin˙NegBiReg**]. This event rarely occurs as real-world data contains significant variance that can be described by a plethora of variables. Therefore, a model that aligns better with this assumption is the Negative Binomial Regression model.

The goal of this model is to obtain a prediction value that can be the center of our prediction interval. Subsequently, after obtaining this value, a good strategy to come up with a prediction interval will be to take the predicted value and have a predicted value minus the error estimate we obtain from our model. The same will be done for the higher bound of the interval

1.3 Poisson Regression

The initial observation previously mentioned considered the Poisson distribution as an explanation for the variation seen in the Number of Reported Results. The graph below shows the different Poisson distributions.

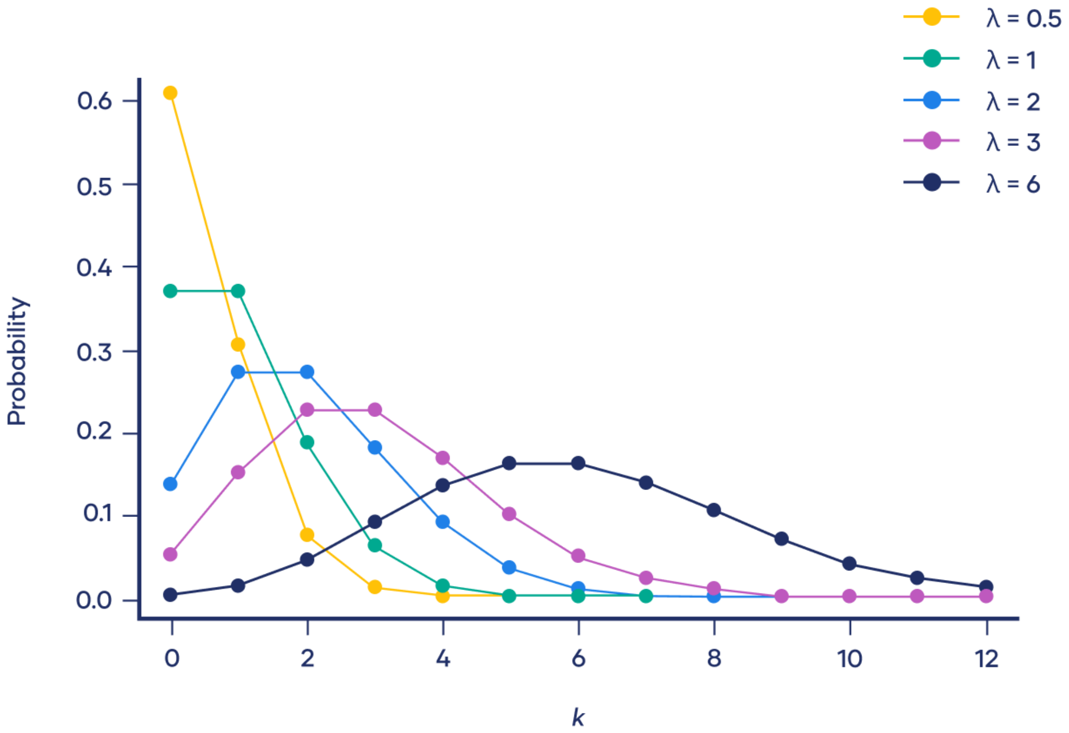


Figure 6: Poisson Distribution

As seen above, the Poisson distribution is dependent on λ, which is the mean number of events. The Poisson distribution attempts to describe the probability of an event happening. Even though this distribution follows the pattern of the data being used, it does not help with predicting the actual Number of Reported Results with a given date. Therefore, by uniting the power of the Poisson distribution with regression, we encounter a model that is called the Poisson Regression model.

The Poisson Regression model is a generalized linear model form of regression analysis[4]. The idea of implementing regression analysis for our prediction is that it allows for finding patterns and trends inside the data set given. Using the data set, we ran a Poisson Regression model using the dates as input parameters into our model. This is the case, as the goal is to have a prediction interval for a future date.

The regression training for the model was performed using the data given. Plotting the model performance with the predictions of the dates in comparison to the actual true values can be observed below:

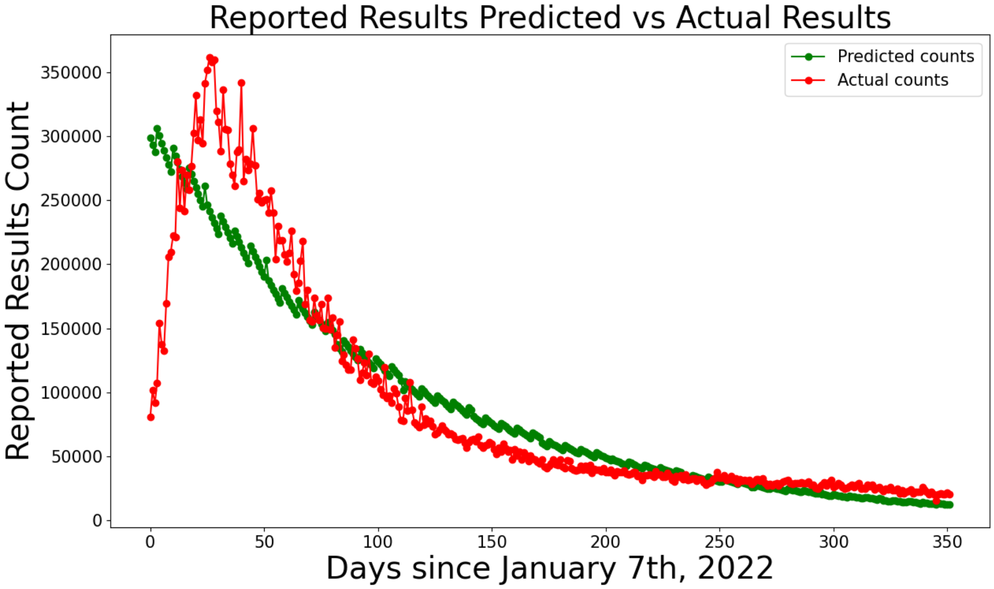
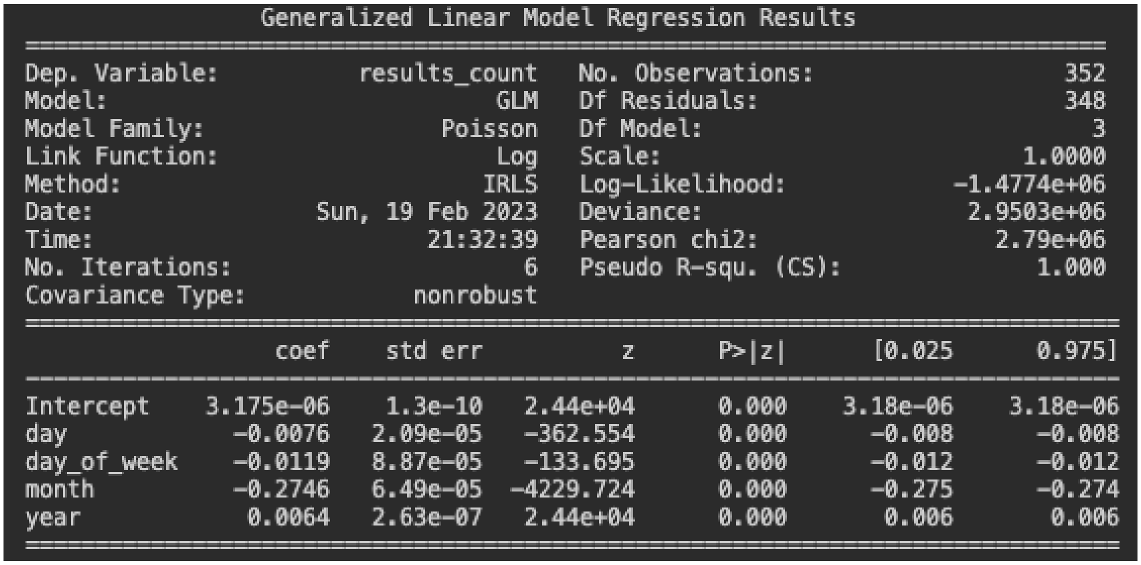


Figure 7: Reported Results VS Prediction Results Poisson Regression

Observing the result obtained from the graph, within the first days (See Figure 7, it can be seen how the actual counts for the Number of Reported Results skyrocketed for the beginning of the usage of Wordle. However, it plummets within a similar interval as it increased initially. After the significant increase and decrease in Reported Results, it steadily decreased over time. Considering the prediction obtained, we can observe how it is not performing successfully with the rapid increment and decrement seen in the first days. However, the prediction was able to slowly decrease in a similar fashion as the actual Reported Results. This prediction will lead to a high error estimate that would enlarge our prediction interval.

Our goal is to reduce the range of the interval to the fullest. After running our model, we looked at the GLM (Generalized Linear Model) Results Summary and we obtained:



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:          results_count   No. Observations:                  352
Model:                            GLM   Df Residuals:                      348
Model Family:                 Poisson   Df Model:                            3
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:            -1.4774e+06
Date:                Sun, 19 Feb 2023   Deviance:                   2.9503e+06
Time:                        21:32:39   Pearson chi2:                 2.79e+06
No. Iterations:                     6   Pseudo R-squ. (CS):              1.000
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept    3.175e-06    1.3e-10   2.44e+04      0.000    3.18e-06    3.18e-06
day            -0.0076   2.09e-05   -362.554      0.000      -0.008      -0.008
day_of_week    -0.0119   8.87e-05   -133.695      0.000      -0.012      -0.012
month          -0.2746   6.49e-05  -4229.724      0.000      -0.275      -0.274
year            0.0064   2.63e-07   2.44e+04      0.000       0.006       0.006
===============================================================================
```

Figure 8: GLM Summary Result

As can be observed in the results above, the reported values of Deviance and Pearson chi-squared are extremely large. To make a quantitative determination of the goodness-of-fit at some confidence level, say 95% (p=0.05), we need to compare the Deviance or Pearson's chi-squared value in order to see if there is significance for goodness-of-fit. Within the given summary, there are 348 Degrees of Freedom (DF Residuals = Number of observations minus Number of Degrees of Freedom in the model). Looking at a table for significant standard Chi-Squared values at p=0.05 and with approximately a number close to 350 Degrees of Freedom the value is around 280-300, which is much smaller than the reported statistic for Deviance and Pearson's chi-squared [5]. Hence as per this test, the Poisson regression model, in spite of demonstrating a decent visual fit for the test data set, has fit the training data rather poorly. This occurs because the Poisson model has an equi-dispersion assumption regarding the equality between the Variance and the Mean.

This leads us to guide our prediction model into the Negative Binomial Regression model.

1.4 Negative Binomial Regression

Negative Binomial Regression is a widespread generalization of Poisson regression because it relieves the highly restrictive assumption that was previously proven significant. The standard Negative Binomial Regression model is based on the Poisson-gamma mixture distribution[6]. This mixture is exemplifying the Poisson distribution using a Poisson parameter that in itself is a random variable [6]. Additionally, to loosen up assumptions, this model is prevalent because it models the Poisson heterogeneity (diverse in content) with a gamma distribution.

Using the Reported Results data, similarly as was done with the Poisson Regression, prediction values were obtained and plotted in comparison to actual results:

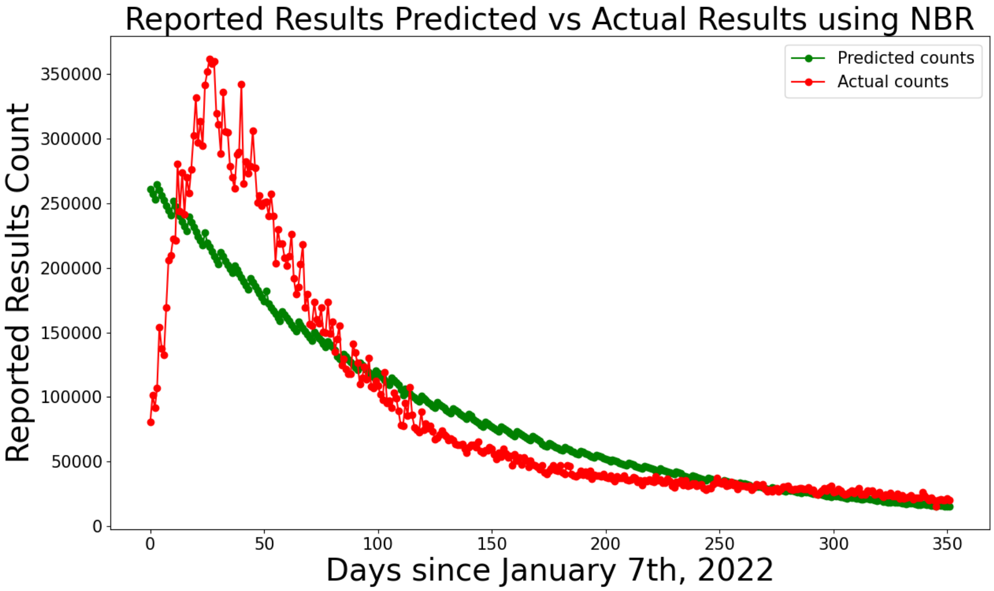


Figure 9: Reported Results VS Prediction Results Negative Binomial Regression

As can be observed in the graph, there is a similar behavior to the Poisson distribution. Within the initial changes that occur in the peak of values for Reported Results, it can be seen how the model is under-predicting the values. However, for values that occur days after the significant drop, predictions start to align better with the actual values. Further in the study, the error estimates will be presented and compared. Nonetheless, considering the lack of restriction that the Poisson model had, we can continue adding parameters to the model such that predictions improve and error estimates decrease.

One special process that we have considered using to improve our predictions is using Iterative Curve Fitting.

1.5 Iterative Curve Fitting

Iterative Curve Fitting is the continuous process of constructing a curve that attempts to best fit a series of data points. In this case, our data points are the Number of Reported Results in the data set. Curve Fitting aligns curve interpolation (obtaining value from previous ones) with smoothing so that the curve can fit the data.

Regularly, Curve Fitting is used in simple models such as Linear Regression or even Polynomial Regression [3]. However, we are encountering a data set that can be considered Non-Linear as there is no linear relationship or any sequential evidence of a relationship between the given data points.

The curve fitting function we use is defined as:

$$f(x) = \sum_{a,i=0}^{numOfCoefficients} \begin{cases} a * \left(\frac{1}{x^{\lfloor \frac{i}{2} \rfloor}}\right) & i \mod 2 = 0 \\ a * x^{\lfloor \frac{i}{2} \rfloor} & i \mod 2 \neq 0 \end{cases} \tag{2}$$

The curve obtained by fitting it with the given series above, plotted against the Number of Reported Results given in the data set, produces the following visualization:

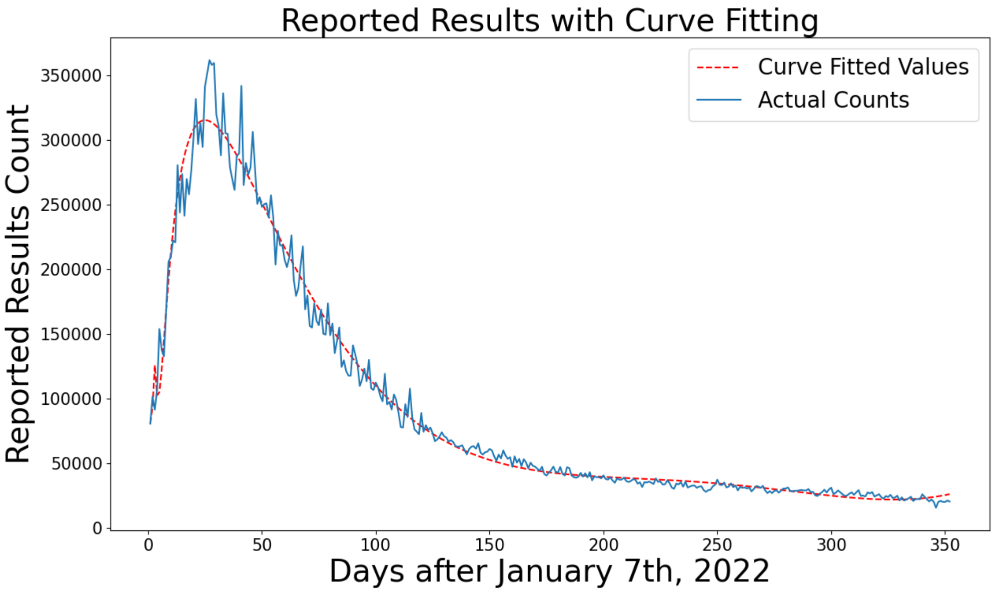


Figure 10: Reported Results VS Iterative Curve Fitting

Considering the results obtained from the graph, it can be said that the Iterative Curve Fitting process successfully fitted an accurate and smooth curve within the data points given. It is an important task to understand that the process of fitting uses the data to interpolate the curve obtained. Additionally, it can be observed how at the end of the curve there is a slight trend increment, an event that does not align with the past trend over time. Even though the curve given has great fitting capabilities, it will not satisfy an accurate prediction due to the incremental trend found in the last entries of the data set.

Given the fact that it aligns points with certain interpolation and smoothness to the curve, one positive use we can give to the curve fitting process is that we can use those points as an extra parameter to the Negative Binomial Regression model.

1.6 Negative Binomial Regression with Fitted Curve parameter

The implementation of having not only the given date as input for the Regression model but additionally having the fitted points given by the Iterative Curve Fitting process allows the model to be more precise with respect to previous values. The model has reached a point where it is not only implementing regression but also interpolation.

As done similarly with previous iterations of the model, the predictions given by the new model are plotted in comparison to the actual values for the Total Number of Reported Results:

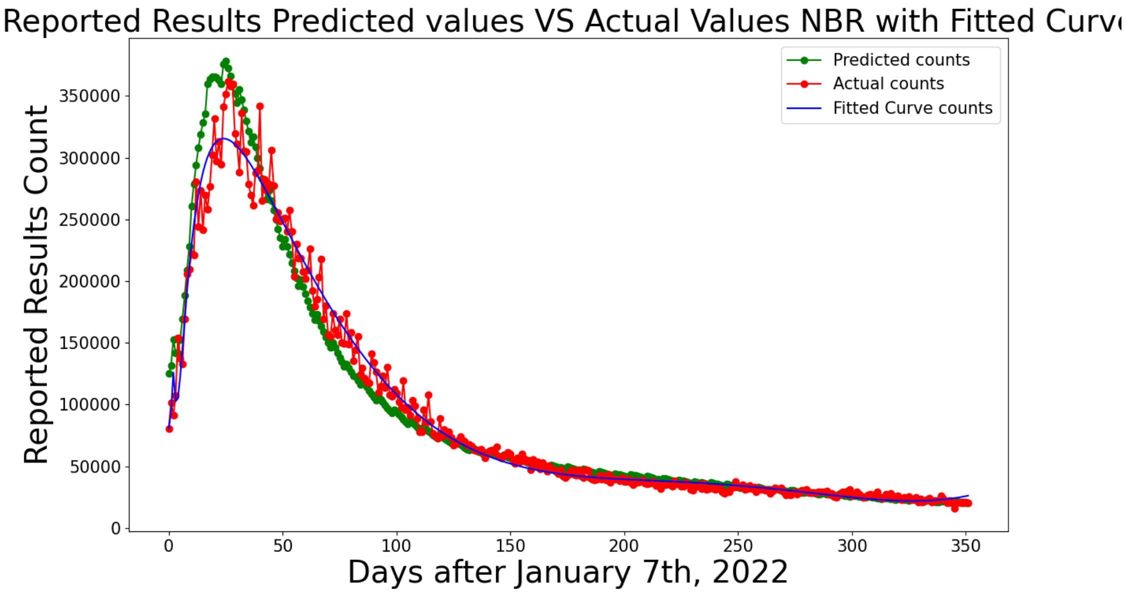


Figure 11: Reported Results VS Negative Binomial with Fitted Curve parameter

As can be observed and analyzed above, the prediction (given in green) dramatically improves the initial prediction outputs. As can be seen in Figure 7 and Figure 9, the predictions obtained for the trend that aligns with a peak in the first entries given where completely off to the actual value. However, this new iteration of the model does a better job of aligning itself with the initial dramatic incrementing and decrementing behaviors. This is the type of solution that is accurate to our goal of having a fitted prediction interval.

The next step to get our interval is to obtain an error estimate that aligns with our goal.

1.7 Error Estimation

For the given data set, the different iterations of our models have shown differences in performance and progressively become better suited for the distribution found in the Number of Total Reported Results. However, in order to obtain a numerical representation of how good the models have been, we will calculate different error measures. The errors we will be considering are:

- Mean Absolute Error
- Root Mean Squared Error
- Weighted Mean Absolute Error

1.7.1 Mean Absolute Error

Mean Absolute Error, MAE, is one of the methods of calculating error that we used. MAE takes the sum of the absolute value of the difference between each predicted and actual value. Then, it divides it by the number of values. It is defined as:

$$MAE = \frac{1}{n}\sum_{i=0}^{\infty} |y_i - \hat{y}_i| \tag{3}$$

Where y_i is the actual number of Reported Results and $\hat{y}_i$ is the predicted value.

It specifically calculates how off our predictions were on average per day. MAE contains a useful attribute: outliers are not weighted too heavily since all errors are weighted on the same linear scale [8].

1.7.2 Root Mean Squared Error

The Root Mean Squared Error, RMSE, is another error calculation method that we looked at, defined as:

$$RMSE = \sqrt{\frac{\sum_{i=0}^{\infty} |y_i - \hat{y}_i|^2}{n}} \tag{4}$$

Where y_i is the actual number of Reported Results and $\hat{y}_i$ is the predicted value.

RMSE calculates the aggregated mean of the difference of predicted values against the actual ones squared. Subsequently, it takes the square root of these errors, giving a final error estimate. The problem with RMSE is that it is a weighted measure of model accuracy that uses the same scale as the prediction target. [2]Therefore, it gives extra weight to outlier values. As we have seen through the study, at the beginning of the Number of Reported Results there are outliers that impact our data. Therefore, it will give a worse estimate than the Mean Absolute Error estimate. Early in our prediction, there were outliers that were not actually bad predicted values.

1.7.3 Weighted Mean Absolute Error

One observation we have had over the course of the analysis has been the dramatic increment and decrement of the Number of Reported Results through the initial entries of the data set. The dramatic changes occur during days 0 to 120 after January 1st, 2022 (see Figure 12). Therefore, calculating the Mean Absolute Error of the initial dramatic changes is significantly greater than the days after. This can be observed in Figure 7 and Figure 9.

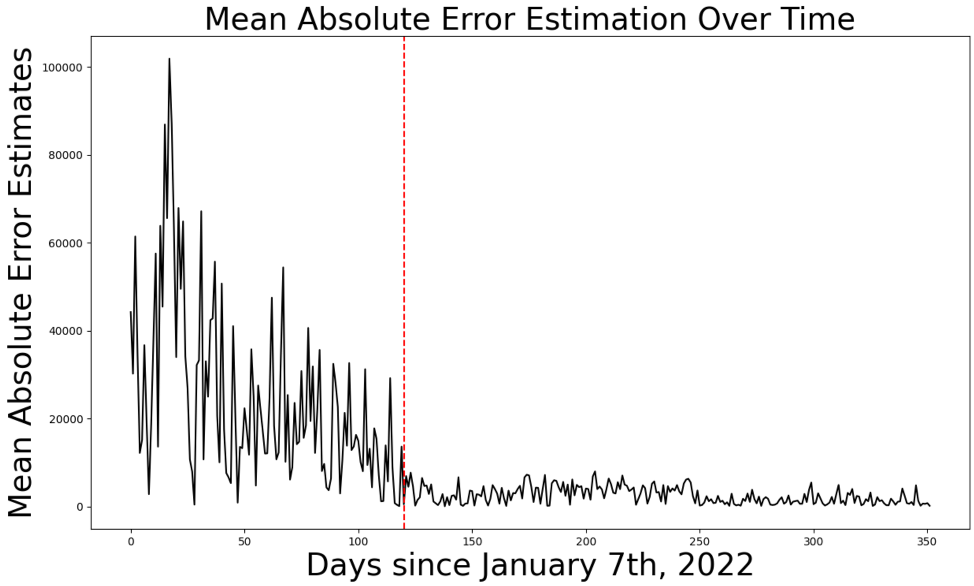


Figure 12: Mean Absolute Error Estimation Over Time

To account for this, we decided to first calculate the MAE for the first 120 days and weighted by the proportion of the days used for calculating the MAE. Henceforward, we then calculated the MAE for the rest of the days and gave it a weight proportional as well to the number of days used for the calculation presented. The equation displayed is:

$$Total\ Error = \frac{MAE_{\leq 120} * 120}{total\ days} + \frac{MAE_{>120} * (total\ days - 120)}{total\ days} \tag{5}$$

This weighs the first 120 days less than the rest. We decided to perform this weighting process as it can be observed that the trend closest to a future date tends to be more significant that the larger trend. Performing this process allows the early outliers to not impact the error value so significantly.

Below we can see a table of the Error Estimates with the given different iterations of our model:

Model Type	**MAE**	**RMSE**	**Weighted MAE**
Poisson Regression	22,341.07	39,320.72	9,737.17
Negative Binomial Regression	24,187.08	41,169.31	10,359.54
Negative Binomial Regression Fit Curve Param	10,022.86	18,674.41	3,953.56

Table 5: Error Estimates per Model Iteration

As can be observed above, there is a trend to have a worse estimate when using RMSE instead of MAE. Another observation we can get from the table is that using the Weighted MAE is significantly improving the error estimate in all model iterations. Concluding the improvement of our model in the last iteration, it can be seen how the Weighted MAE for this model type is around 4000 per day. As this is the best estimate and we have gotten to a point where we are confident enough with our model, we will predict the Number of Reported Results for March 1st, 2023, and generate the prediction interval interpolating the predicted value with the error estimate.

1.8 Final Predicted Interval

After executing the prediction for March 1st, 2023, the model predicted that there will be **1,4891.37** Total Reported Results. Creating the interval with the error estimate given in the table above (3,953.56). **The prediction interval for March 1st, 2023 is:**

$$[10,937.81, 18,844.93]$$

The interval says that we predict to have a minimum of **10,937.81** and a maximum of **18,844.93** for the Number of Total Reported Results for March 1st, 2023.

1.9 Word Attribute Analysis Regarding Hard Mode

In our data exploration, there appeared to be several attributes that affect the reported number of hard mode players. Overall, it appeared as though attributes that relate to a word's difficulty make it more likely to have been played by a larger proportion of hard mode players compared to total daily players. For example, we found that the words with repeated letters have a statically significant increase in the rate of hard mode players. Other attributes that lead to more hard mode players include the number of vowels and the occurrence of infrequent letters.

1.9.1 Demonstrating Statistical Significance

For any given attribute, the null hypothesis (H_0) is that the given attribute does not correlate with the rate of hard mode players over total players, and that any difference is due the random chance. Typically, the decision to reject H_0 or fail to reject is based on the p-value and a chosen significance level α. If the p-value is less than or equal to α, you reject H_0; if it is greater than α, you fail to reject H_0. However, this decision can also be based on the confidence interval calculated using the same α. If the reference value specified in H_0 lies outside the interval, you can reject H_0. If the reference value specified in H_0 lies within the interval, you fail to reject H_0. For our purposes, we will always use $\alpha = 0.95$.

Taking a random sampling of the means of random samplings of the rates of hard mode gives a normal distribution for which we could calculate a confidence interval (See Figure 13). With $\alpha = 0.95$, the confidence interval for hard mode rate is $(0.0729, 0.0775)$. Now to check if a specific sample grouping of the total population is statically significant, we can take the mean of the hard rate of that group and compare it to the confidence interval.

1.9.2 Specific Attributes and Possible Explanations

Some words contain repeated letters and taking the mean of the rate of hard mode for all those words gives a mean of 0.0780 which is outside the confidence interval; thus making repeated letters a statistically significant factor which determines the rate of hard mode.

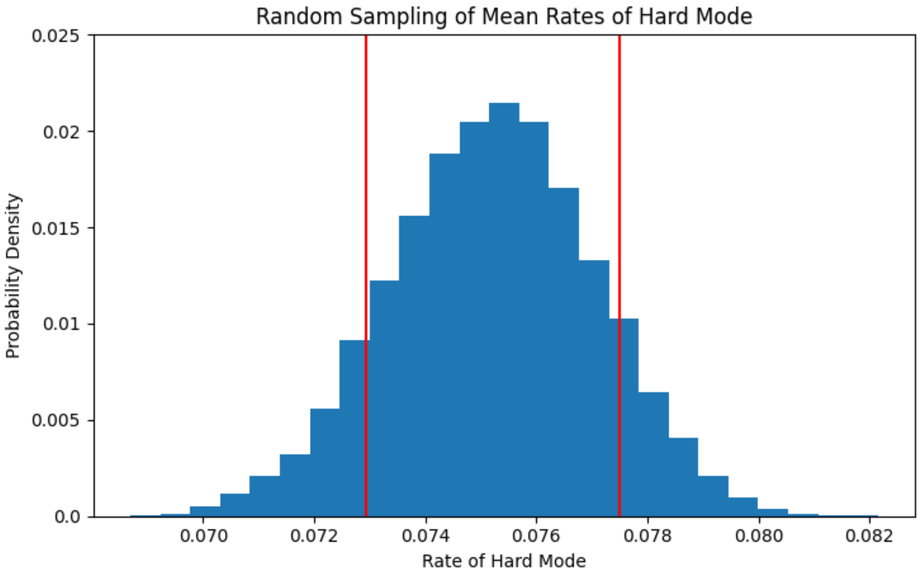


Figure 13: Random Sampling of Mean Rates of Hard Mode

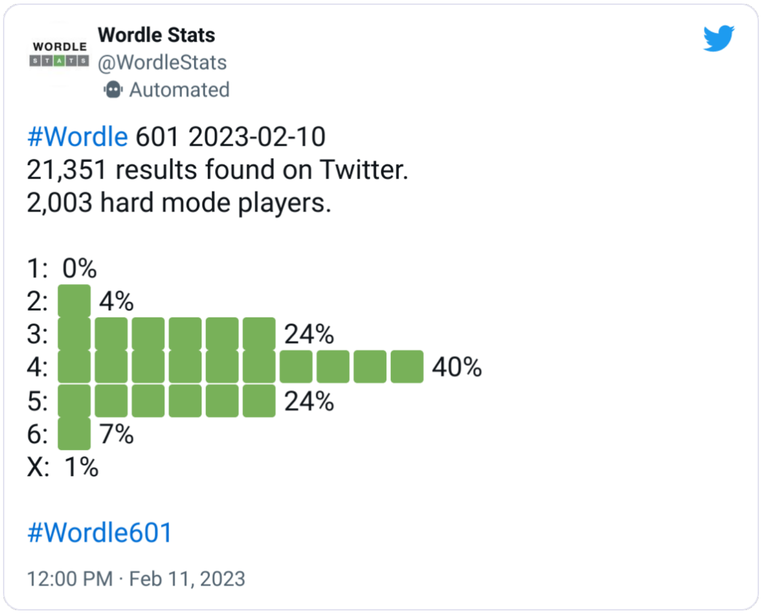
Similarly, we grouped words into four additional groups based on the vowels (A, E, I, O, U): words containing no vowels, words containing one vowel, words containing two vowels, and words containing three vowels. There were no words containing either four or five vowels so those groups are empty. For words containing no vowels, the mean rate of hard mode was 0.0921 which is far outside the confidence interval and therefore significant. Words containing one and two vowels had a mean of 0.0730 and 0.0752 respectively and being inside the confidence interval were thus insignificant. Finally, words containing three vowels were significant with a mean of 0.0811.

Additionally, the letters J, Q, X, and Z are the least common letters in the English language as well as within this dataset. Grouping words which contain one or more of these letters is significant with a mean of 0.0852.

The significant attributes are those that suggest a more difficult word. Words that have the least common letters or have comparatively many and few vowels or repeated letters are attributes of words that one would think would be more difficult. In fact, grouping the 50 most difficult words by mean number of attempts gives a statistically significant hard rate mean of 0.0780. Given the statistical correlation and significance of these attributes of the words when related to the rate of reported hard mode results, it would seem as though there may be an external aspect at play. People playing hard mode do not cause the words to have these particular attributes. It may be the case that the harder the word the more viral a Wordle contest becomes and thus those who are more experienced and more often play hard mode may be playing and reporting their scores. Additionally, it may be the case that when a word is harder less total normal mode players report their scores as those scores may be slightly worse.

2 Prediction Model for Distribution of Reported Results

In the supplied data, there are 7 individual results we are interested in predicting; however, the 7 individual results are codependent on each other. Looking over the results, the number of tries and fails falls under a normal distribution (See Figure 14). Knowing this, rather than create specific models that generate 7 outputs, one for each number of attempts and one for a fail. It makes more sense to generate 3 distinct outputs: mean number of attempts, standard deviation of number of attempts, and success rate. An argument could be made that success rate is worthless as the number of fails also appear to follow the normal distribution; however, semantically we felt as though to fail the Wordle is distinctly different than, for example, giving it a discrete value for number of attempts. Failing a Wordle does not mean the player would win the Wordle if given a 7th attempt, so it is unknown how many attempts that game would take. As such, we decided to maintain that value separately.[1]



Wordle Stats
@WordleStats
Automated

#Wordle 601 2023-02-10
21,351 results found on Twitter.
2,003 hard mode players.

1: 0%
2: 🟩 4%
3: 🟩🟩🟩🟩🟩🟩 24%
4: 🟩🟩🟩🟩🟩🟩🟩🟩🟩🟩 40%
5: 🟩🟩🟩🟩🟩🟩 24%
6: 🟩 7%
X: 1%

#Wordle601

12:00 PM · Feb 11, 2023

Figure 14: Example distribution from WordleStats on Twitter

Our strategy will involve using a multiple linear regression model with particular word attributes as input, more specifically, a linear regression with ordinary least squares. Moreover, for simplicity, rather than create a single multivariate regression model that gives three outputs, we decided to create three separate multiple regression models.

2.1 Model Features

Our model contains 5 distinct features:

- **Repeated letters**: the number of repeated letters in a given word.
- **Rare letters**: the number of rare letters in a given word where a rare letter is a letter in the set $\{J, Q, X, Z\}$.

- **Vowels**: the number of vowels in a given word.
- **Letter frequency score**: A score associate with each word representing the sum of the each letter's frequency in the dataset.
- **Word frequency score**: A score associated with each word representing how frequent that word is in the English language.

Statistical significance plays an important role for the selection of the features displayed. As analyzed and concluded previously, words with repeated letters, words with specific number of vowels, and words with uncommon letters were proven significant. Hence, there participation in the model. Additionally, one other attribute of words that can be deterministic for the level of difficulty is the study of frequency analysis of the letters within the data given and the perceived analysis for the words within the English language.

2.2 Multiple Linear Regression Model

While each model uses the same structure and features, the difference between each of the 3 models is the output data to which the model is being fit. For example, the mean number of attempts model will be fit to the mean number attempts for each word, and the success rate model will be fit to the success rate of each word. Without loss of generality, while explaining how we got to our final model, we will use the particular example of the model of the mean number of attempts.

Using our selected features, the first iteration of our model was a multiple linear regression. It was fit on the 5 features as input and a mean number of attempts calculated from the given dataset as the expected output. Figure 15 visualizes a random sampling of words and the model's prediction. This figure shows both the true means and predicted means, as well as error bars showcasing how far off the prediction was. The model generated a mean absolute error of 0.265 and a root mean square error of 0.323. At first glance this model may seems quite good; however, a closer examination shows the outliers of the easier and more difficult word (by the metric of mean number of attempts) have a much larger error than the rest of the dataset. In fact, the prediction for words with the 10 largest and 10 smallest mean attempts gives a $MAE = 0.475$, almost a half an attempt of error, and a $RMSE = 0.518$. Another metric that may be useful to look at is the coefficient of the determination, R^2, which is defined as:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \mu)^2}$$

where $\hat{y}_i$ is the predicted value μ is the mean value and y_i is the true value. This metric is commonly used with regressions to determine how well a model fits its training data where the maximum value of R^2 representing a perfectly fit regression is 1.0. Using this metric, this model has $R^2 = 0.368$.

2.3 Weighted Fit Multiple Linear Regression Model

This weighted fit model should give more importance to the outlier words based on the number of attempts in order to reduce the MAE and $RMSE$ in those specific outliers while maintaining a similar error overall. As such we weighted the training data rows by the square difference between each word's mean number of attempts and the total mean of all the dataset's mean attempts. When the model was fit with this weighted training data, the $MAE = 0.288$, $RMSE = 0.349$, and $R^2 = 0.203$ which had a larger

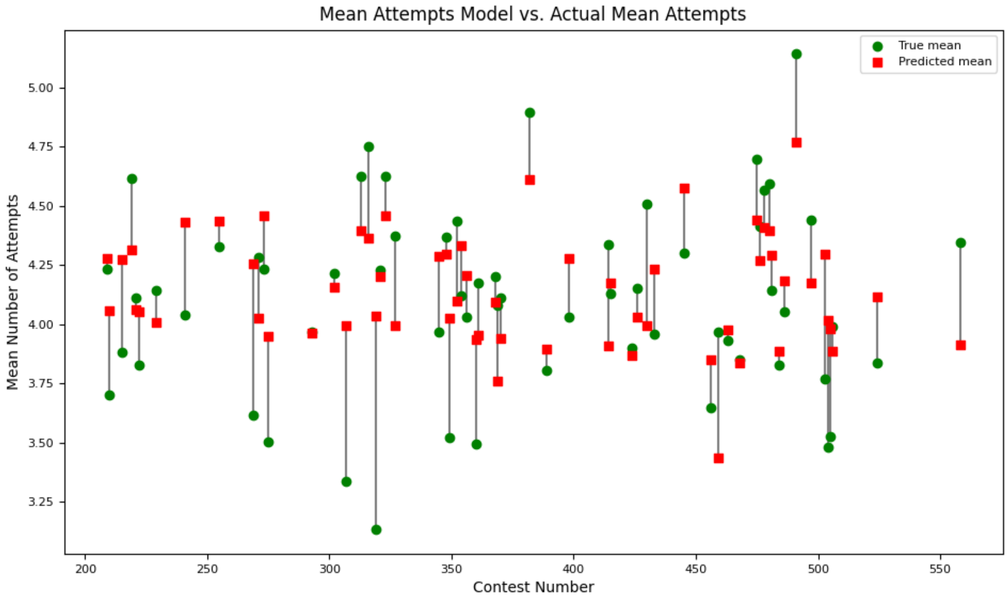


Figure 15: Prediction Sampling of the First Model

error in nearly every metric. Nevertheless, looking exclusively at the outliers, the error was improved with $MAE = 0.372$ and $RMSE = 0.433$ versus the unweighted model which gave $MAE = 0.475$ and $RMSE = 0.518$. Arguably, it is more important to correctly predict outliers rather than to predict those that are close to the mean.

2.4 Multiple Linear Regression Model with Polynomial Features

This model removes the notation of a weighted fit and attempts to improve the model through another means. Effectively, we used a polynomial regression which was generated by prepossessing the 5 linear features into 26 features: every combination of features of at most degree 3 (which includes a constant term of 1). These new 26 features are then passed through the original multiple linear regression model and fit to the appropriate dataset.

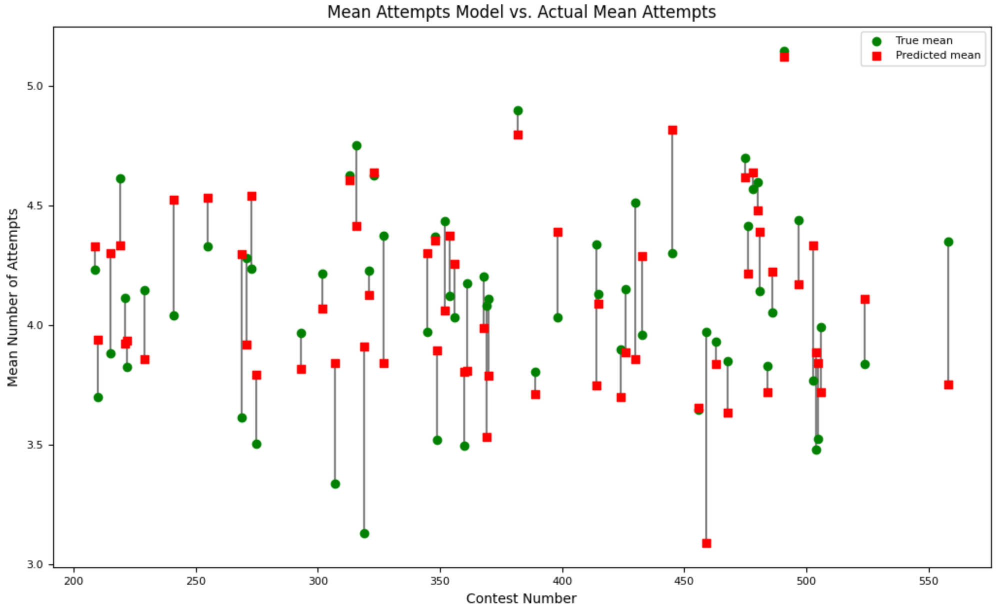


Figure 16: Prediction Sampling of the Second Model

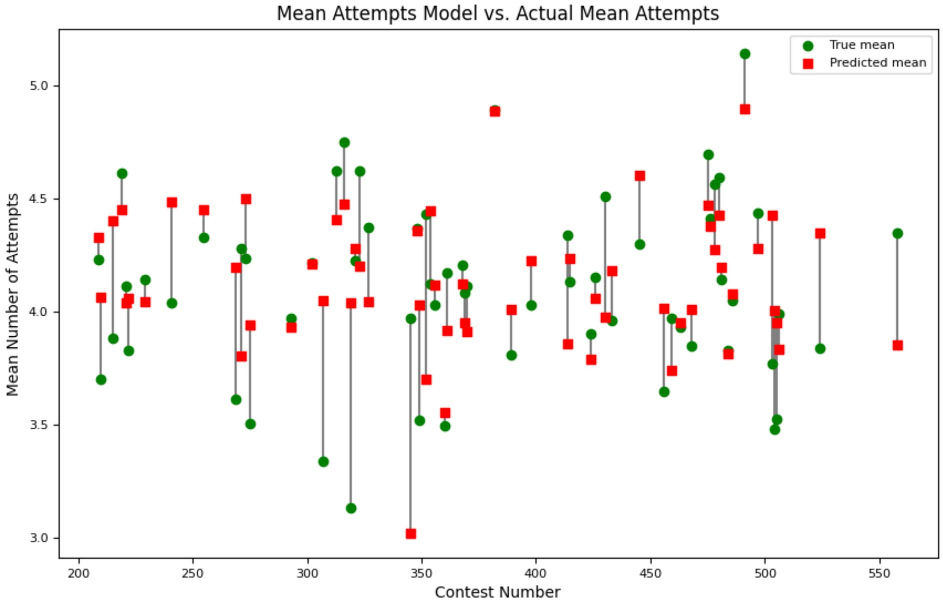


Figure 17: Prediction Sampling of the Third Model

This new model gave MAE = 0.278, $RMSE$ = 0.357, and R^2 = 0.390. This results in the best R^2 metric out out all the models; however, the other models provide better MAEs and $RMSE$s. Looking exclusively at outliers we have a MAE = 0.395 and $RMSE$ = 0.452 which a great improvement from both the first and second models.

2.5 Model Comparisons and Selection

Overall, the multiple linear regression with polynomial features was the best model, optimizing for a best-fit model with the R^2 metric. Additionally, it had improvements to the outlier words—the 10 most difficult and 10 easiest words by the metric of mean number of tries—as the MAE and RMSE were much lower than that of the first linear model. The weighted linear model while having better metrics than the other two models was not optimal as it overfit and underfit different parts of the data through the addition of the weights. By the R^2 metric, it cannot confidently predict new and unique words. As such, the polynomial features represent the best model for prediction.

Model	***MAE***	***RMSE***	R^2	**Outlier's *MAE***	**Outlier's *RMSE***
Linear	0.265	0.323	0.368	0.475	0.518
Weighted Linear	0.288	0.349	0.203	0.372	0.433
Polynomial Features	0.278	0.357	0.390	0.395	0.452

Table 6: Error Metrics for Each Model

2.6 Model Predictions for EERIE

Using our three linear regression models with polynomial features, they predict for EERIE that the mean number of attempts will be 4.800, the standard deviation of number of attempts will be 0.916, and the success rate will be 95. As such by sampling from a normal distribution with the associated values, the following graph is generated:

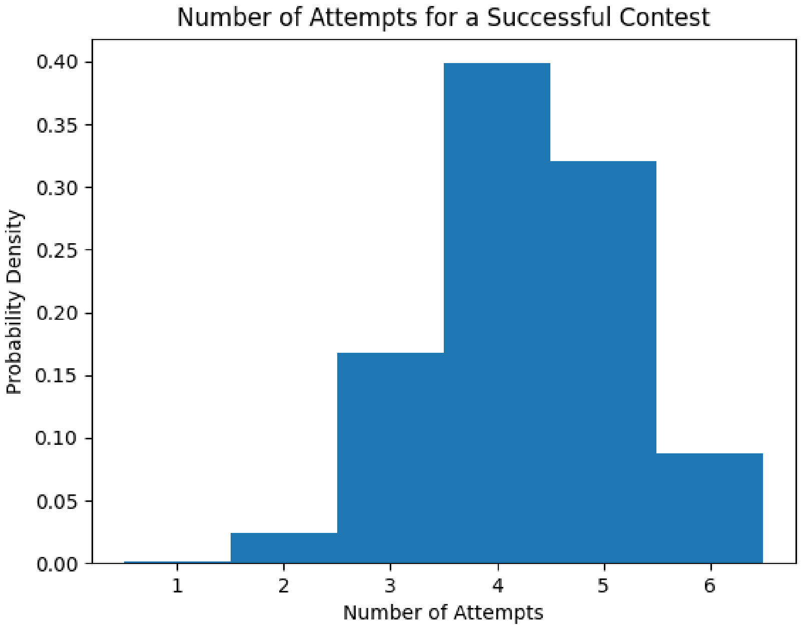


Figure 18: Prediction Sampling of the Third Model

Furthermore, Table 7 summarizes our predicted results for the word EERIE. Of note, is that due to rounding errors, the total sums to 99; however, this same result was also showcased in the supplied dataset and, as such, is not an issue.

Word	**1 Try**	**2 Try**	**3 Try**	**4 Try**	**5 Try**	**6 Try**	**7+ Try**
eerie	0	1	6	26	40	21	5

Table 7: EERIE Prediction

3 Letter to the New York Times

Dear Puzzle Editor of the New York Times,

Our team has developed models to predict the future usage and success of the popular puzzle game, Wordle, based on the data reported to the Wordle Stats Twitter page. There were interesting trends in the data. We first noticed a huge increase of over 281,000 results in the number of reported results from January 7th, 2022 to February 2nd, 2022. Then there was a significant decrease over the next few months. Since then, it has been steadily decreasing. We also noticed that the number of people playing hard mode compared to the total number of reported results has slowly increased over time, indicating that players who continue playing are getting more comfortable with the game and are willing to try it on hard mode.

Our first model involved predicting the number of reported results for specific future dates. This utilized Negative Binomial Regression as well as iterative curve fitting. We looked at a Poisson Regression model first, but it was not sufficient since it was not close enough to the original data, so we decided to use a Negative Binomial Regression model, a widespread generalization of Poisson Regression since the data fit a Poisson Distribution. This model fit the data slightly better, but there was one more necessary element, curve fitting. The goal of curve fitting is to train a curve to match a set of given data points. This model was successful in matching the data, but it was not good at predicting since the final trend was distant from previous trends. Because of this, we decided to use the data from the curve-fitting model in our Negative Binomial Regression Model, which follows the final trend of the given data, to increase the accuracy. This decreased the amount of error, calculated by taking the absolute value of the difference between the model and actual data on a given day, and summing them up. There still was a large error, because of the larger outliers in the beginning. To solve this, we weighed the days after May 7th, 2022, more than the previous days. To test the model, we predicted the number of reported results for March 1st, 2023 and got an interval from 10,937 to 18,844 results. It looks like Wordle is going to be less played in the near future. We also performed 95% confidence intervals to determine if any features of the word impacted the hard mode rate. We concluded that words that contained repeated letters, words with none and 3 vowels, and words with the least common letters; J, Q, X, and Z; did slightly impact the percentage of players playing on hard mode.

The next model we created dealt with predicting the rates given a word for 1 through 6 tries and also not getting the wordle correct. We looked into the various features of words previously mentioned and three different models: Multiple Linear Regression, Weighted Multiple Linear Regression, and Multiple Linear Regression with Polynomial Features. In each model, the input was significant word features. Overall the Multiple Linear Regression Model with Polynomial features left us with the lowest error, more accurately able to predict outcomes. Our model was tested with the word eerie: 1st try resulted in 0%, 2nd try resulted in 2%, 3rd try resulted in 16%, 4th try resulted in 38%, 5th try resulted in 30%, 6th try resulted in 8%, and 5% of people failed to guess the word. There are various limitations to the model: not enough words with 3 vowels were in the given data. This makes it difficult to predict the results for a word with 3 vowels, like eerie. There are other limitations, but our model works sufficiently well as it matched the actual data quite accurately. We hope this finds you well.

Kind Regards,
COMAP Team 2321082

References

[1] Wordle Stats Twitter Account. *Wordle Stats Feb 11th, 2023*. URL: `https://twitter.com/WordleStats/status/1624453429157576705` (cit. on p. 16).

[2] Stephen Allright. “How to interpret RMSE”. In: (2022). URL: `https://stephenallwright.com/interpret-rmse/` (cit. on p. 12).

[3] Jason Brownlee. *Curve Fitting With Python*. 2020. URL: `https://machinelearningmastery.com/curve-fitting-with-python/` (cit. on p. 9).

[4] Sachin Date. *The Poisson Regression Model*. 2022. URL: `https://timeseriesreasoning.com/contents/poisson-regression-model/` (cit. on p. 6).

[5] Sachin Date. *Values of the Chi-squared distribution*. URL: `https://www.medcalc.org/manual/chi-square-table.php` (cit. on p. 8).

[6] Dan Ma. “The Negative Binomial Distribution”. In: (2011). URL: `https://probabilityandstats.wordpress.com/tag/poisson-gamma-mixture/` (cit. on p. 8).

[7] Richard Patil. “What is Exploratory Data Analysis”. In: (2018). URL: `https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15` (cit. on p. 1).

[8] George Sief. “Understanding the 3 most common loss functions for Machine Learning Regression”. In: (2019). URL: `https://towardsdatascience.com/understanding-the-3-most-common-loss-functions-for-machine-learning-regression-23e0ef3e14d3` (cit. on p. 12).