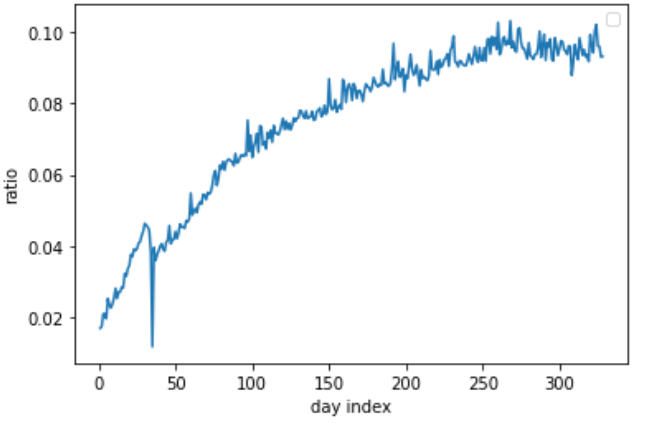
We take ratio = Number in hard mode/Number of reported results as labels for this question.



To plot the ratio over the day, we find that the ratio increases as the day increases regardless of words. This might be because when the time passes, there are more experienced players who tend to play hard mode to challenge themselves. To analyze the attribute of the word that affects such a ratio, the previous early date data might add too much noise to our analysis of the word since it is greatly influenced by time. Therefore, we only take examples when the ratio over time is relatively stable. Based on the graph, we take data after day 220.

We analyze multiple aspects of the attributes of the word. We consider mainly three kinds of the attributes: letter constitution, understanding level, and similarities between 5-words corpus.

letter constitution:

We consider the maximum letter repitition of the word. For example, the maximum letter repitition of the word ”good” is 2

Syllable count: we count the number of syllables of the word

Word\_format: we the structure of letter of the word. For example the format for “eerie” is 11231

letter of the word: we analize the specific of the word, and take several letter or letters that have some obvious correlation to the distribution of tries and ratio. The letter we choose are: 'f', 'g', 'j', 'm', 'p', 'q', 'v', 'w', 'x', 'z', 'oo', 'ir', 'ph', 'ly', 've', 'wh', 'sk', 'ch', 'ck', 'ng', 'qu', 'th','ie'

Understanding level:

Diff\_score: if syllable is bigger or equal to 2, we consider it hard; otherwise, easy

Brown\_freq: the frequency of the word in Brown Corpus

Reuters\_freq: the frequency of the word in Reuters Corpus

'Freq\_HAL', 'Log\_Freq\_HAL', 'I\_Mean\_RT', 'I\_Zscore', 'I\_SD', 'Obs', 'I\_Mean\_Accuracy'. Don’t know these, please check here:

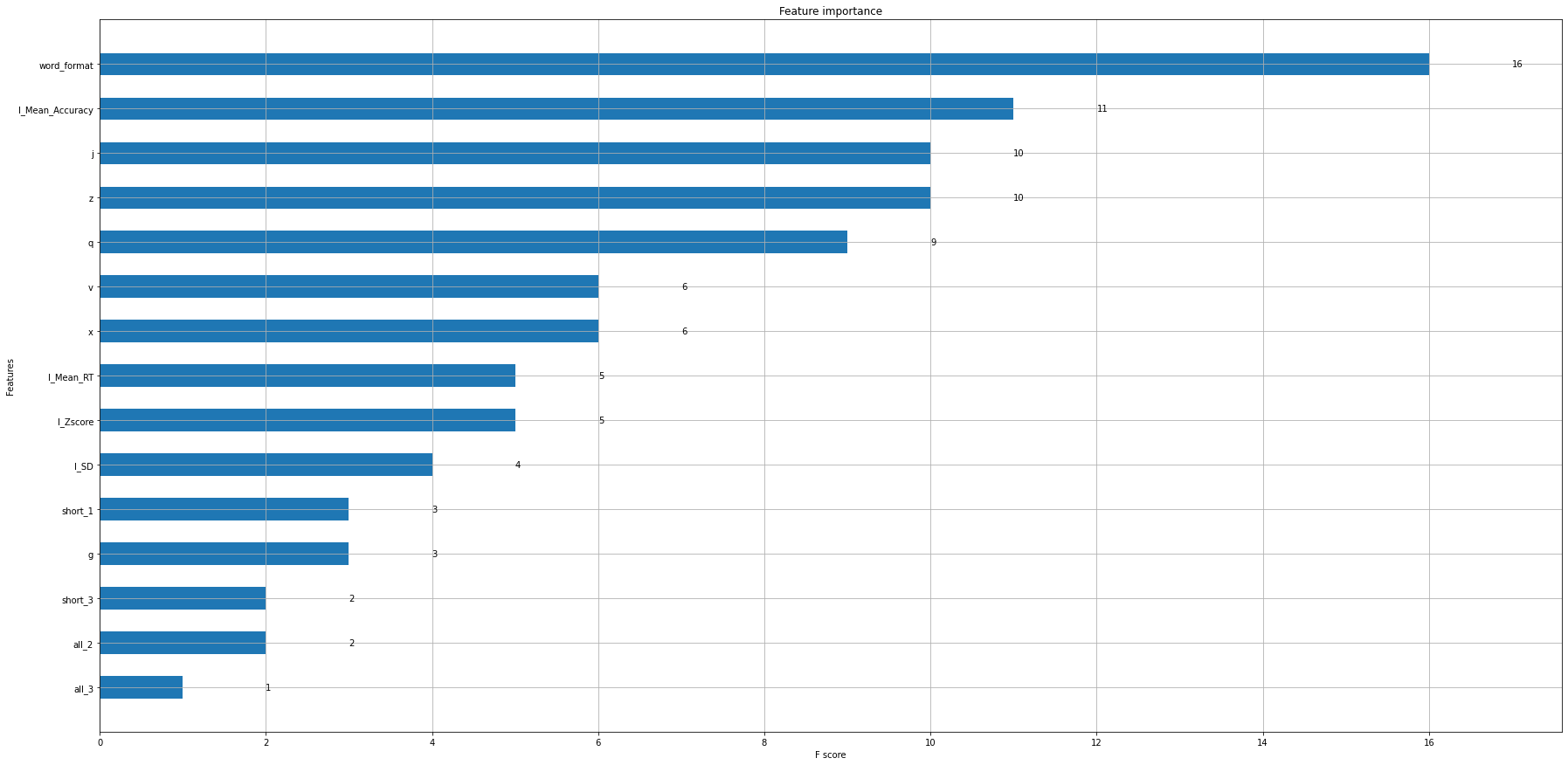
<https://www.kaggle.com/datasets/kkhandekar/word-difficulty>

similarities between 5-word corpus：

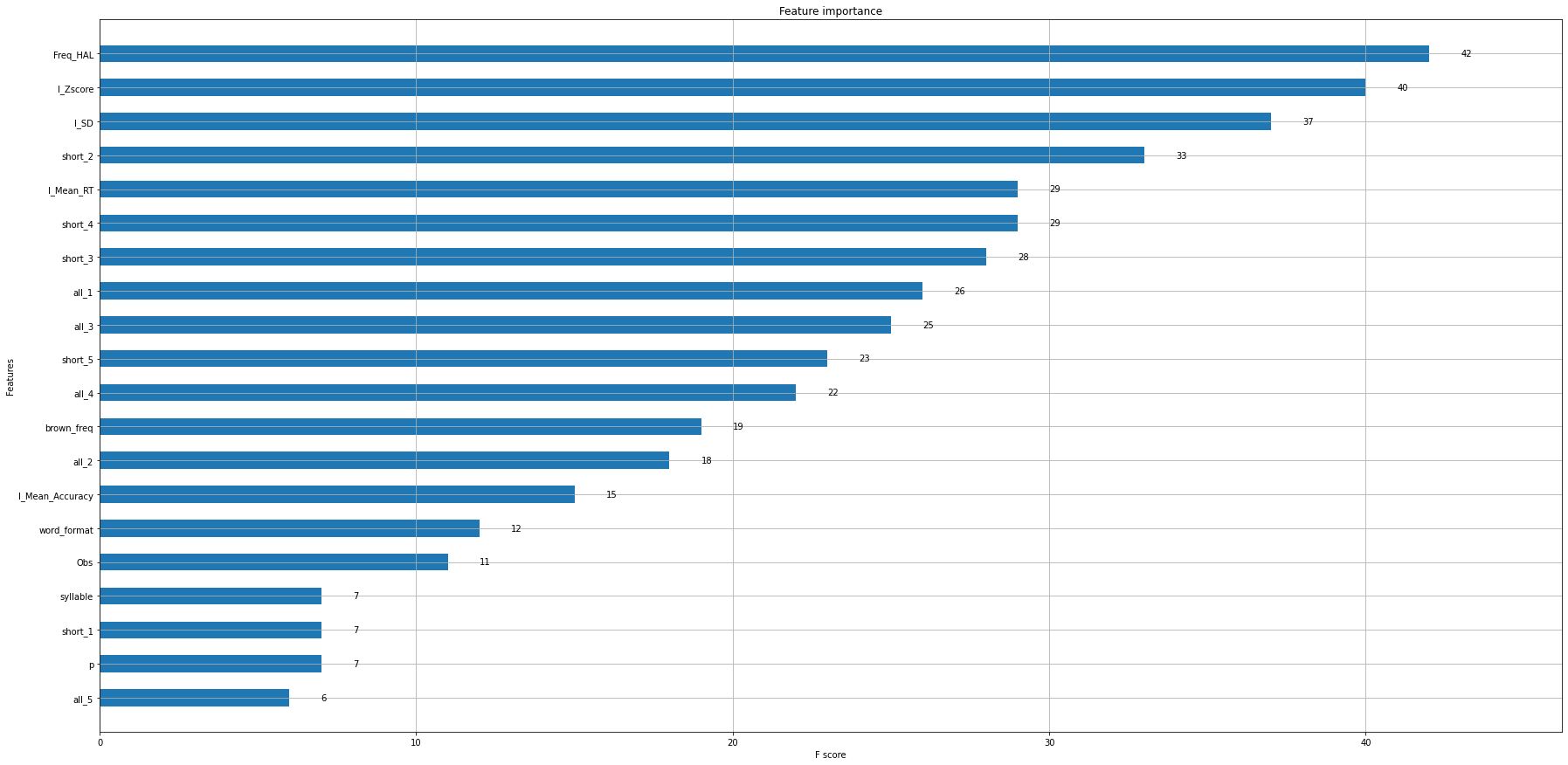
We check the rules for Wordle carefully and find that “There is a [list of 2315](https://github.com/alex1770/wordle/blob/main/wordlist_hidden) relatively common five-letter words, from which the hidden word of the day is chosen, and there is a larger [list of 12972](https://github.com/alex1770/wordle/blob/main/wordlist_all) five-letter words (which includes the smaller list) that can be used to guess with.” We analysis Levenshtein distance of the word to all the word in each of the list, and count the number of Levenshtein distance[get formula for this and write formally] as 1,2,3,4,5(since there is only 5 letters, the maximum Levenshtein distance is 5).

To better explain these attributes towards the affect on ratio, we use tree-like models. Among all the tree-like models, we use XGboost which has the best performance. We apply MSE as loss for it.

And this graph shows the top attributes based on our model.



Since this graph only uses 140 days’ data, it is relatively small and might occur the problem of overfitting. Many letters of the word is considered important if we train our model with 140 day’s data. But this might just be due to these letters accidentally occurs more during these 140 days. Thus, we also use the entire 360 days data to get a attribute importance graph to better find the importance of the attributes.



Although this model might be influenced by time series, but it still contains important information related to words. By comparing with previous graph, we find that the letter in the word attributes are less significant here, it might support our hypothesis earlier. Thus, we think the format of the word and I Mean Accuracy are important attributes to the ratio. The result for the prediction is 0.10227332. The training and test rmse is 0.002495, 0.002649. Their similar, so the result is compelling.

