Vikram reddy
Essay scoring model

Aim:

The aim of this model is to score the essay by extracting the features from the model

Features used:

1. Essay_set
2. Essay
3. Score
4. Unique_words,
5. Sentiment
6. Sent length
7. Word count
8. Mistake count
9. Noun count
10. Grammatical mistake
11. Avg word length in each essay
12. Average sent length in each essay
13. Long word count
14. Vocab richness
15. Proper noun count
16. Adj count

Sentiment:The sentiment property returns a namedtuple of the form Sentiment(polarity, subjectivity). The polarity score is a float within the range [-1.0, 1.0].

Word count and Sentence count: These are very basic features of any text document and do influence the scoring of the document as well. So, to extract these features, we use the textmining library in python. This library also provides a list of 276 common stop words in English language. Now, since these stop words are of not much importance, we skipped them while calculating the word count of each document from the termdocument matrix. To get the sentence count, we simply split the document using '.' and thus, count the number of segments obtained

POS Tags: Another crucial set of features for evaluating any piece of writing is the number of words in various syntactic classes like nouns, adverbs, verbs, adjectives etc. These features are crucial for evaluating the quality of content in the essay. To get the counts of words in each POS (part-of-speech) class, we use the NLTK library in python. This library gives us the POS tag for each word in an essay, and thus, we extract the number of nouns, adverbs, adjectives and verbs.

Spelling Mistakes: An important parameter while scoring an essay is the spelling mistakes. So, number of spelling mistakes in an essay is also a feature for our model.

Vocab riches:

The usage of longer words often indicates a greater depth of language in the student. A method to assess the vocabulary difficulty in an essay is to measure lexical diversity and vocabulary richness. Yule's I characteristic scans through words in an essay and stems each one in order to find the total 4 number of unique words. This number is then normalized and adjusted to prefer non-repeating words to create a quantitative measurement for lexical diversity and vocabulary richness.

Long word:If a word's length was greater than seven characters, it was added to the long word count feature. The usage of longer words often indicates a greater depth of language in the student

NOTE: I have used Regression not a Classification model

Models used:

1. Random forest regression
2. Gradient Boosting regression

Random Forest Regression

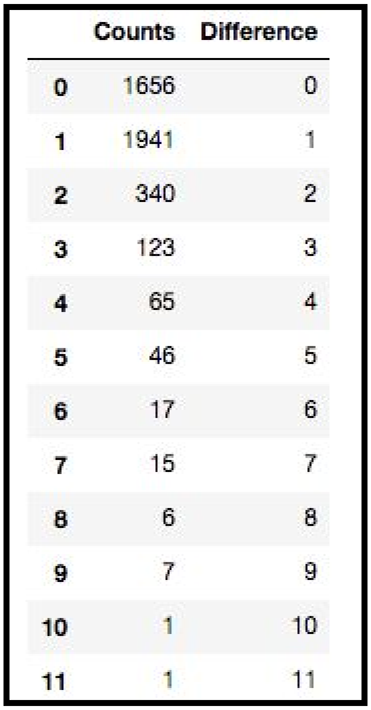
The Random Forest is one of the most effective machine learning models for predictive analytics, making it an industrial workhorse for machine learning.

The **random forest** model is a type of additive model that makes predictions by combining decisions from a sequence of base models. More formally we can write this class of models as:

$$g(x)=f0(x)+f1(x)+f2(x)+...$$

where the final model g is the sum of simple base models fi. Here, each base classifier is a simple decision tree. This broad technique of using multiple models to obtain better predictive performance is called **model ensembling**. In random forests, all the base models are constructed independently using a **different subsample** of the data.

Difference between original and predicted values and count of essays having that difference



	Counts	Difference
0	1656	0
1	1941	1
2	340	2
3	123	3
4	65	4
5	46	5
6	17	6
7	15	7
8	6	8
9	7	9
10	1	10
11	1	11

So out of 4218 Essays in Testing data 1656 predicted exactly and 1941 with difference of 1

So in total 1656+1941= 3597 out of 4218 have predicted very close with +-1 difference

Gradient Boosting:

The Boosted Trees Model is a type of additive model that makes predictions by combining decisions from a sequence of base models. More formally we can write this class of models as:

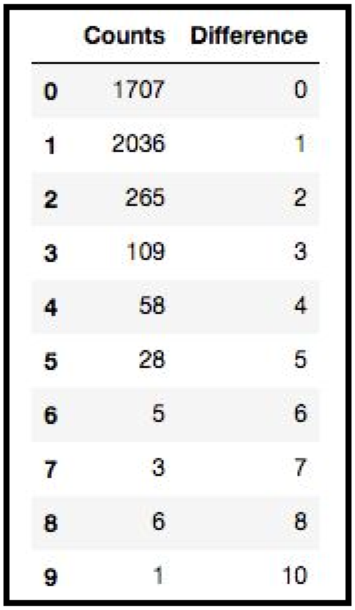
where the final classifier is the sum of simple base classifiers . For boosted trees model, each base classifier is a simple decision tree. This broad technique of using multiple models to obtain better predictive performance is called model ensembling.

Unlike Random Forest which constructs all the base classifier independently, each using a subsample of data, GBRT uses a particular model ensembling technique called gradient boosting.

The name of Gradient Boosting comes from its connection to the Gradient Descent in numerical optimization. Suppose you want to optimize a function , assuming is differentiable, gradient descent works by iteratively find

For regression problems with squared loss function, is simply . The algorithm simply fit a new decision tree to the residual at each iteration.

Difference between original and predicted values and count of essays having that difference



	Counts	Difference
0	1707	0
1	2036	1
2	265	2
3	109	3
4	58	4
5	28	5
6	5	6
7	3	7
8	6	8
9	1	10

So out of 4218 essays 1707 essays have been exactly scored and 2036 with a difference of 1 from original scores.

So 1707+2036= 3743 essays out of 4218 have been scored almost exactly with a difference of -+1

As we have built the regression model and compared to both the models,Gradient Boosting has achieved the good accuracy rate

Of score values ranging from 0-60 and max difference of 10 for only one value we can consider that this model has achieved good results and it is reliable