AUTOMATED ESSAY SCORING

[HIMANSHU MISHRA (IIT Kanpur)]

SEPTEMBER 19, 2021

[SHL ML INTERNSHIP]

Introduction and approach:

In the assignment, I have focused on simple structural features of the test like sentence, word length, character, adjective, counts, lemma etc. For modelling these features, I have used linear regression, and other regression like adaptive boosting regression etc.

Feature extraction:

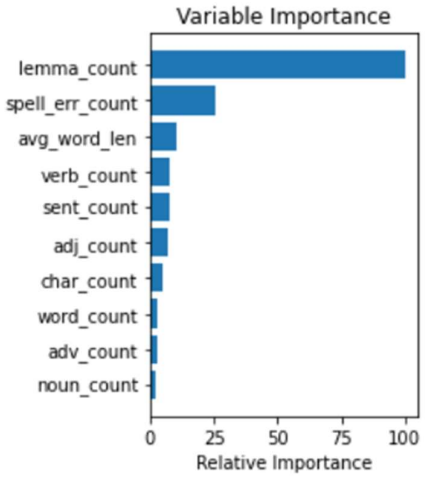
The features are extracted from big.text and English.pickle toolkits. Some if the features that are used are:

- Sentence to wordlist
- Average word length
- Word count
- Character count
- Sentence count
- Lemma count
- Bag of words (BOW)
- Number of adjectives, adverb, verbs, nouns etc

Regression:

I have used Linear regression and an ensemble of boosting regression for making predictions. I have used two metrics for evaluation of the performance namely: mean squared error and variance. MSE is a simple matric that signifies squared deviation from the ideal value whereas variance represents skewness of the fit.

The model also takes a look to the features importance related to the provided data. The models shows that score dependence on lemma count is the highest fallowed by spelling errors, average word length etc.



The results for the regressions are quite good over the validation test sets showing small MSE(0.8-0.7) value. The model implemented performs much better for narrative, persuasive, and expository essays.

The current approaches used focus on regression for predictions along with the features that best fit to regression. However, a feature that fits into validation set may be irrelevant to test set that may complicate grading algorithm which may result in overfitting.

STEPS IMPLEMENTED IN PYTHON NOTEBOOK

1. Importing required packages, libraries and data files (train.csv, test.csv, all_prompts.csv)
2. Checking the data shape, size and characteristics
3. Feature extraction is implemented which includes average word length feature, sentence to wordlist, bag of word (BOW), word count, character count, sentence count, lemma count, mis-spelled word, noun count, adjective count, verb count, adverb count etc.
4. Then splitting the training data into training data and validation set (70: 30) and training a Linear Regression model using only Bag of Words (BOW). Thus calculating MSE and variance score.
5. Again, splitting the training data into training data and validation set and training a Linear Regression model using all the features. Thus calculating MSE and variance score.
6. Similarly again training and testing the Lasso Regression and boosting regression model for the dataset and obtaining MSE and variance score for each case.
7. Boosting regression model using all features performs better than other models. Thus predicted evaluator rating for the test case is obtained using the model.
8. Finally, appending the predicted evaluator rating value to test_prediction file.