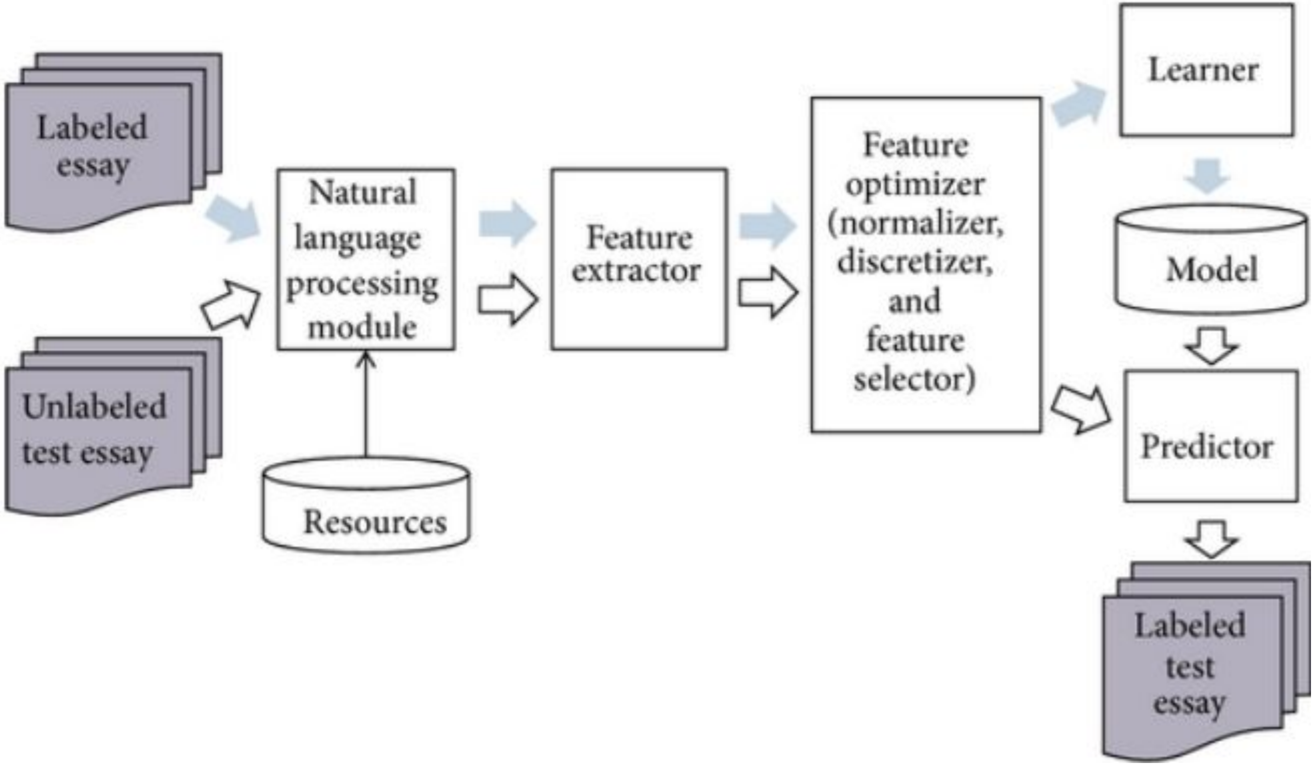
Automated Essay Scoring as Basic Regression

Ashesh Singh

Background

What is Automated Essay Scoring (AES)?


Labeled essay
Unlabeled test essay
Natural language processing module
Resources
Feature extractor
Feature optimizer (normalizer, discretizer, and feature selector)
Learner
Model
Predictor
Labeled test essay

Why AES?

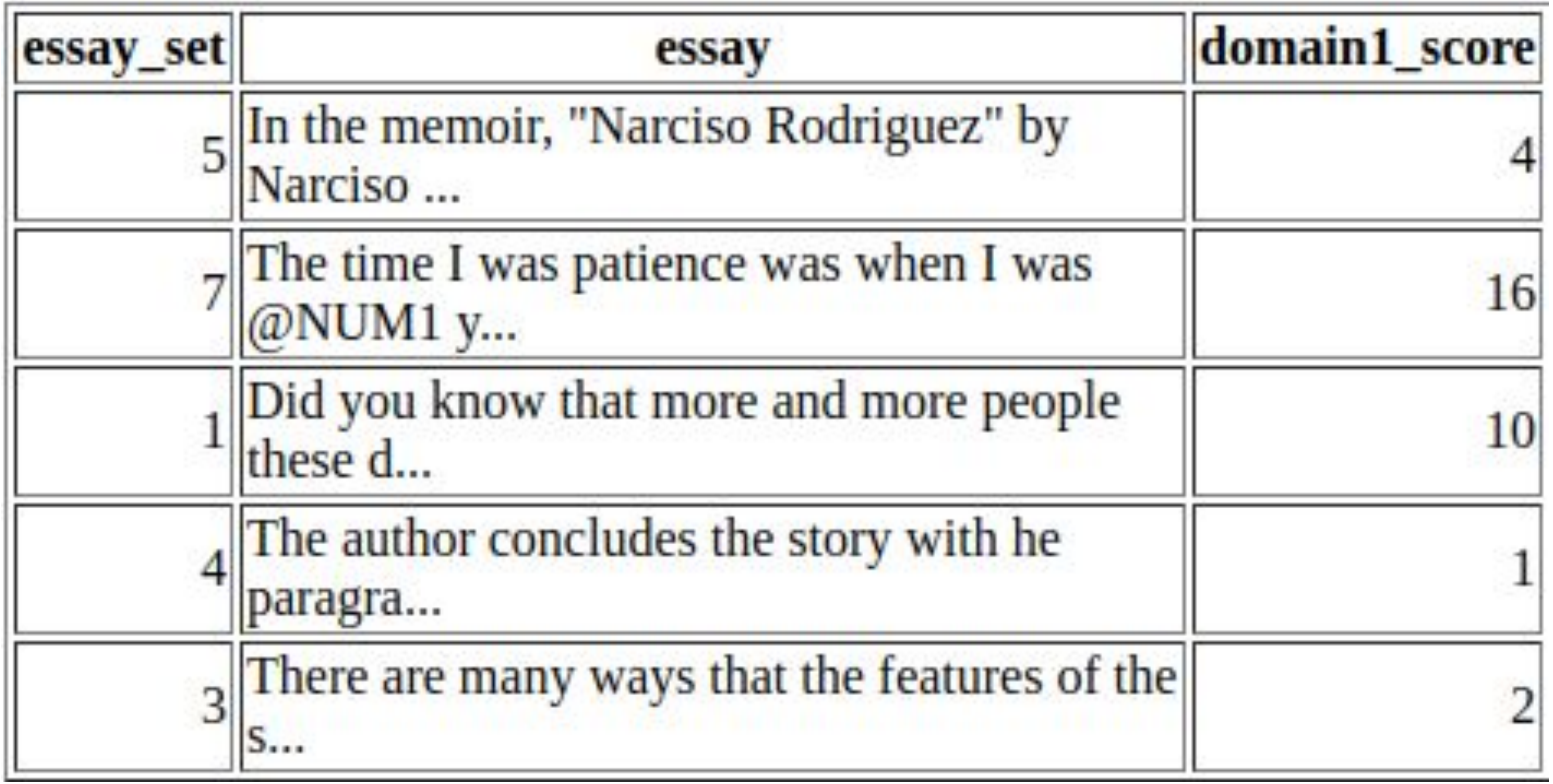
Goal

Demonstrate effect of common essay features

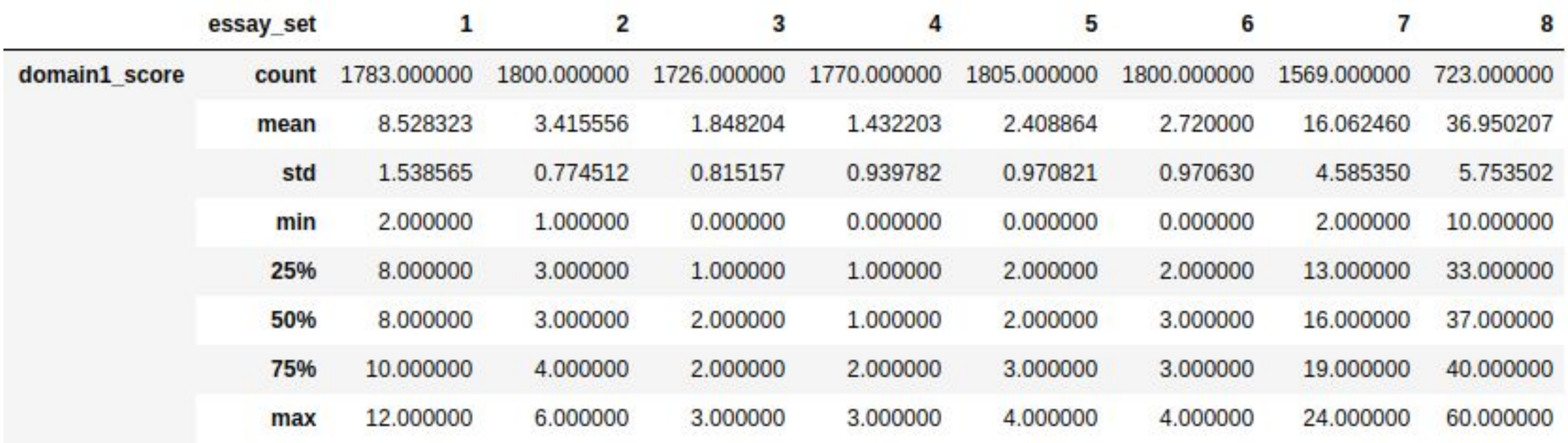
Apply techniques from this course

Hypothesis:
A large number of essay features are required to achieve a good model*

Dataset



essay_set	essay	domain1_score
5	In the memoir, "Narciso Rodriguez" by Narciso ...	4
7	The time I was patience was when I was @NUM1 y...	16
1	Did you know that more and more people these d...	10
4	The author concludes the story with he paragra...	1
3	There are many ways that the features of the s...	2



	essay_set	1	2	3	4	5	6	7	8
domain1_score	**count**	1783.000000	1800.000000	1726.000000	1770.000000	1805.000000	1800.000000	1569.000000	723.000000
	mean	8.528323	3.415556	1.848204	1.432203	2.408864	2.720000	16.062460	36.950207
	std	1.538565	0.774512	0.815157	0.939782	0.970821	0.970630	4.585350	5.753502
	min	2.000000	1.000000	0.000000	0.000000	0.000000	0.000000	2.000000	10.000000
	25%	8.000000	3.000000	1.000000	1.000000	2.000000	2.000000	13.000000	33.000000
	50%	8.000000	3.000000	2.000000	1.000000	2.000000	3.000000	16.000000	37.000000
	75%	10.000000	4.000000	2.000000	2.000000	3.000000	3.000000	19.000000	40.000000
	max	12.000000	6.000000	3.000000	3.000000	4.000000	4.000000	24.000000	60.000000

Methods

Essay Features

meta_features

'essay_length', 'avg_sentence_length', 'avg_word_length'

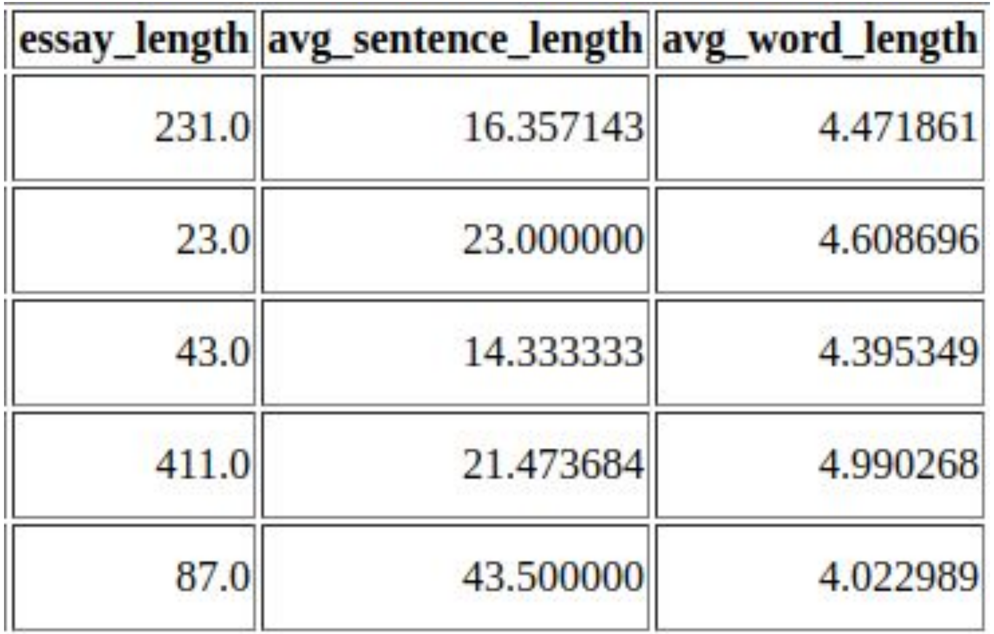
grammar_features

'sentiment', 'noun_phrases', 'syntax_errors'

redability_features

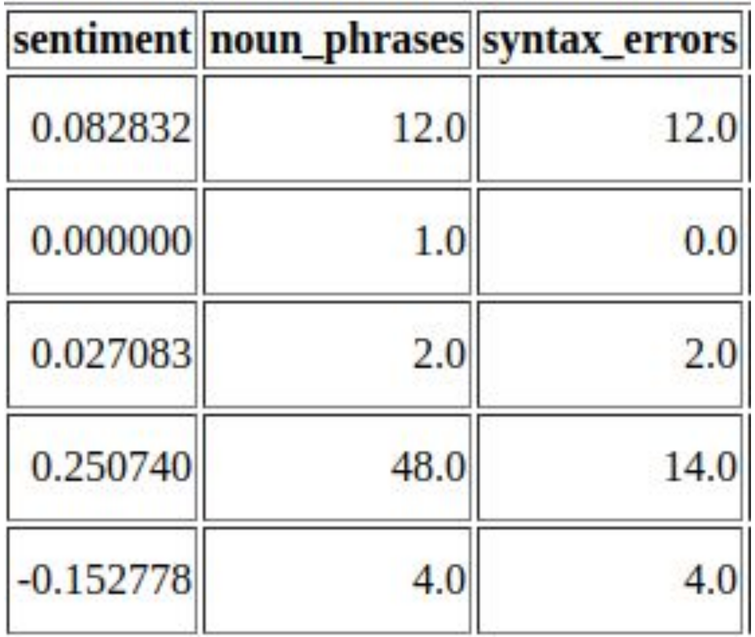
'readability_index', 'difficult_words'

Meta Features



essay_length	avg_sentence_length	avg_word_length
231.0	16.357143	4.471861
23.0	23.000000	4.608696
43.0	14.333333	4.395349
411.0	21.473684	4.990268
87.0	43.500000	4.022989

Grammar Features

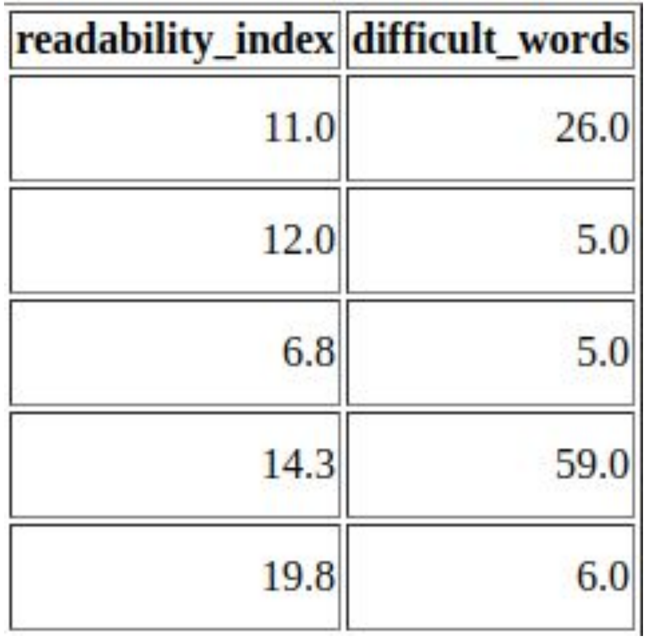


sentiment	noun_phrases	syntax_errors
0.082832	12.0	12.0
0.000000	1.0	0.0
0.027083	2.0	2.0
0.250740	48.0	14.0
-0.152778	4.0	4.0

Readability Features

Automated readability index

$$4.71\left(\frac{\text{characters}}{\text{words}}\right)+0.5\left(\frac{\text{words}}{\text{sentences}}\right)-21.43$$



readability_index	difficult_words
11.0	26.0
12.0	5.0
6.8	5.0
14.3	59.0
19.8	6.0

Model

Used a TensorFlow **Sequential** model with two densely connected hidden layers, and an output layer that returns a single, continuous value.

Training for 1000 Epochs with Callbacks for early return.

Mean Squared Error as loss function.

Results rounded to nearest integer values.

Evaluation

Quadratic Weighted Kappa (QWK)

Measures the agreement between two ratings.

In this case final predicted score and resolved human scores.

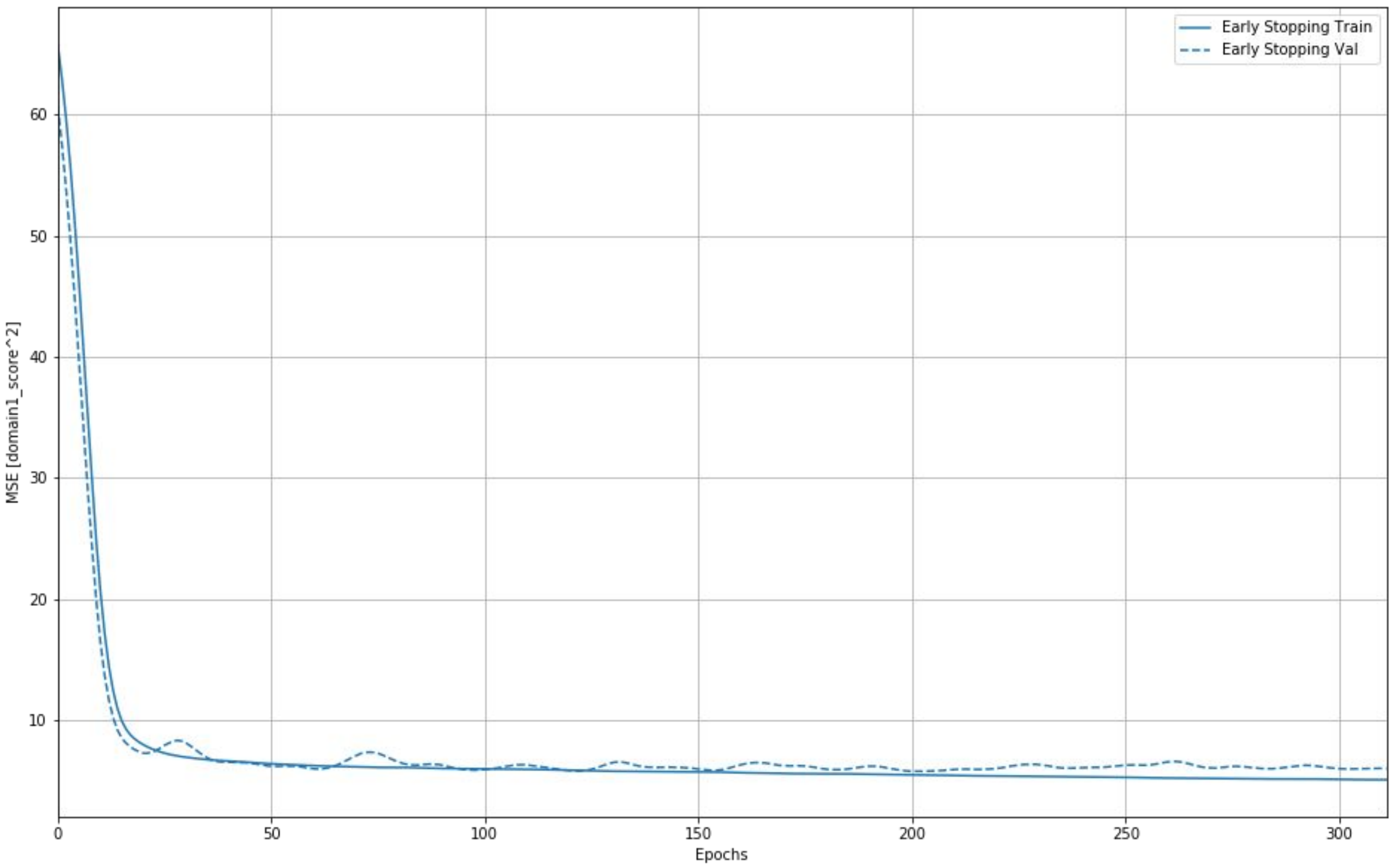
$$\kappa = (p_o - p_e)/(1 - p_e)$$

Results

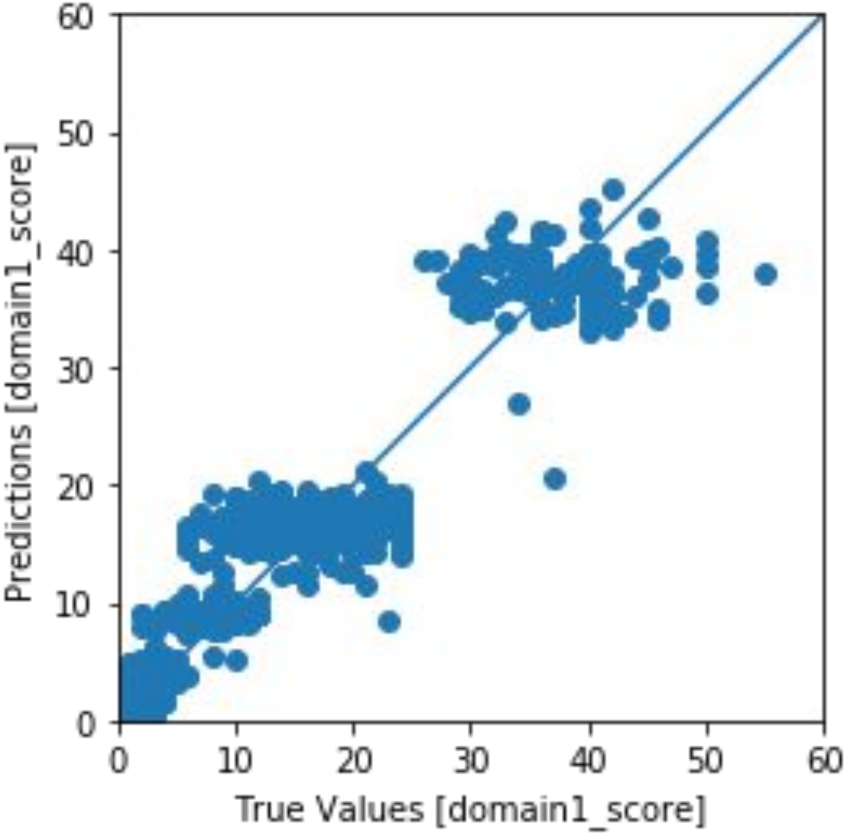
511

Obtained evaluations for 511 feature combinations.

QWK ~ 0.96*



Mean Squared Error Vs. Epoch



Predictions Vs. True Score

Inclusion of `essay_set` in training feature set always improved the results.

Observation 1

Without `essay_set`, **QWK ~ 24**

```
('essay_length',
'avg_sentence_length',
'avg_word_length',
'sentiment',
'noun_phrases',
'syntax_errors',
'readability_index',
'difficult_words')
```

Observation 2

The feature set `('sentiment',)` performed worst with **QWK ~ -0.00016**

The only feature set to have a "chance" agreement.

Expected?

Observation 3

Considering only single feature sets, `('essay_length',)` performed best with **QWK ~ 0.15**, followed by

```
('avg_sentence_length',)
('difficult_words',)
('noun_phrases',)
('syntax_errors',)
('readability_index',)
```

Expected?

Observation 4

Adding more features didn’t always give better results

Conclusion

Applied very simple ideas for feature extraction and training.

Model can do much better with prompt related feature information.

Need for more extensive data cleaning and verification of implementation logic.

References

Yi, Bong-Jun & Lee, Do-Gil & Rim, Hae-Chang. (2015). The Effects of Feature Optimization on High-Dimensional Essay Data. Mathematical Problems in Engineering. 2015. 1-12. 10.1155/2015/421642.

“Basic Regression: Predict Fuel Efficiency :   TensorFlow Core.” TensorFlow. Accessed December 3, 2019. https://www.tensorflow.org/tutorials/keras/regression#the_model.

“Automated Readability Index.” Wikipedia, Wikimedia Foundation, 23 Aug. 2018, https://en.wikipedia.org/wiki/Automated_readability_index.

“Scikit-learn.org. (2019). sklearn.metrics.cohen_kappa_score” scikit-learn 0.22 documentation. Accessed December 3, 2019.
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html