Linking Writing Processes to Essay Quality

Calder Lenhart

Background

The "Linking Writing Processes to Writing Quality" Kaggle competition hosts a dataset comprised of keystroke logs from mock SAT essays collected from more than 5000 participants on Amazon's Mechanical Turk platform. Similar datasets have previously been published, but this one is unprecedented in size. Successfully identifying the relationships between people's writing behaviors and the quality of their work during the writing process has the potential to improve writers' metacognitive regulation.

Data

4 CSV files

- Training/test logs
- Training scores
- Test scores (hidden)

Below is a sample of the first five keystrokes in the training data

id	event_id	down_time	up_time	action_time	activity	down_event	up_event	text_change	cursor_position	word_count
001519c8	1	4526	4557	31	Nonproduction	Leftclick	Leftclick	NoChange	0	0
001519c8	2	4538	4962	404	Nonproduction	Leftclick	Leftclick	NoChange	0	0
001519c8	3	106571	106571	0	Nonproduction	Shift	Shift	NoChange	0	0
001519c8	4	106686	106777	91	Input	q	q	q	1	1
001519c8	5	107196	107323	127	Input	q	q	q	2	1

Note that all alphanumeric characters are masked with the letter "q" to ensure that the models are trained on the writing behaviors rather than content. Note that exams were human-graded on a scale of 0 to 6 in half-integer increments.

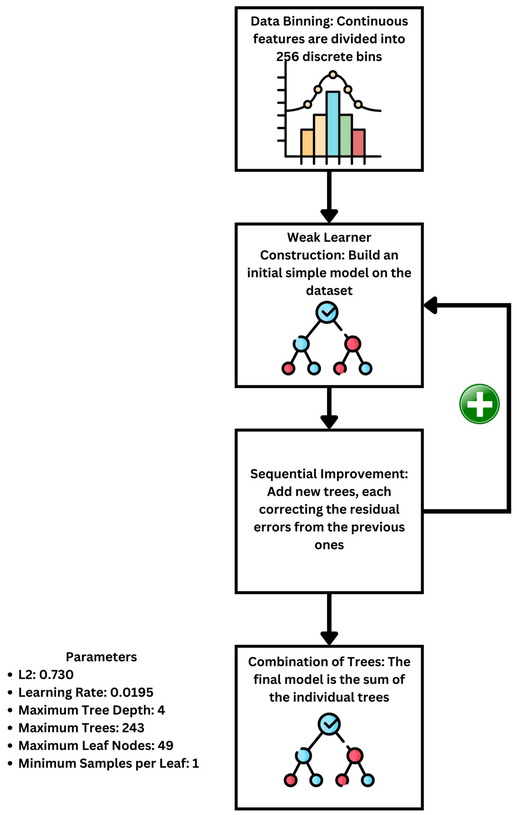
Feature Engineering

I derived 11 features from this data based on two studies of real-time prediction of writing quality using keystrokes [1][2] .

- Word count
- Words per minute
- Interkeystroke interval (IKI): The time between the downstroke of a keypress and the downstroke of the subsequent keypress
- Inter-word IKI: Interval between the end of a word and the beginning of the following word
- Intra-word IKI: Interval between the beginning and end of a word

Model Design

Selection: I chose the Histogram-based Gradient Boosting Regression Tree model available within the sklearn ecosystem. I chose this model for its ability to handle NaNs, optimization for large datasets, ability to regularize. A high-level description of the algorithm is provided below.



Parameters

- L2: 0.730
- Learning Rate: 0.0195
- Maximum Tree Depth: 4
- Maximum Trees: 243
- Maximum Leaf Nodes: 49
- Minimum Samples per Leaf: 1

Hyperparameter tuning:

Using RandomSearchCV, I ran 100 iterations on 5 folds of the training data deriving optimized hyperparameters.

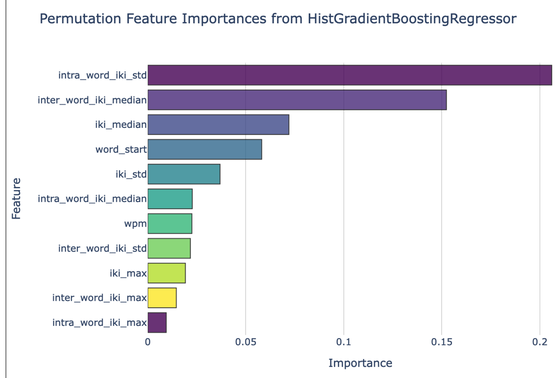
Cross Validation:

I split the training data into 10 k-folds for a more robust representation of the model's performance on the hidden testing data.

Results

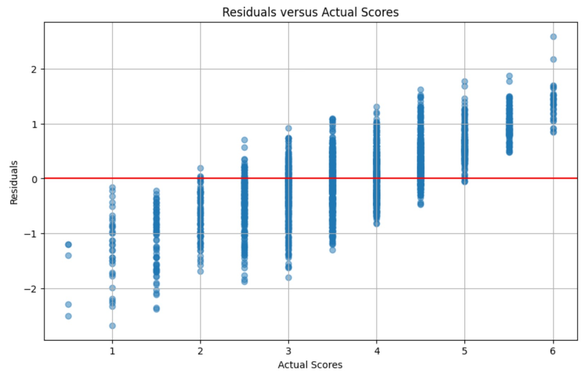
Feature Importance:

Using scikit-learn's built-in permutation feature importance model inspection technique, I found two dominant features: standard deviation of intra-word IKI and the median of the inter-word IKI.



Residuals:

The model performs acceptably when the essay quality earns a score between 3-4, but outside of range it overestimates and underestimates low- and high-scoring essays respectively.

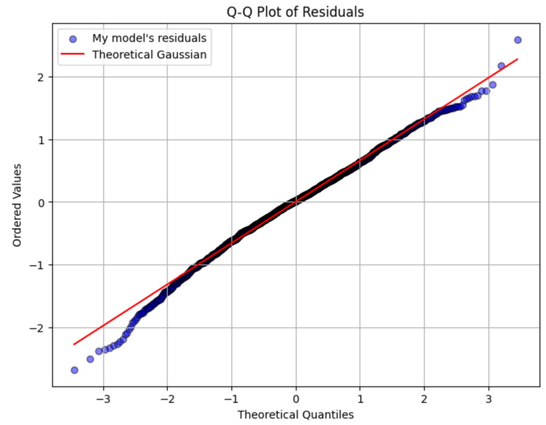


Root Mean Squared Error:

Metric	Value
Median RMSE across k-folds (training)	0.758
RMSE on hidden test set	0.729

Future Improvements

- Improve the model's handling of edge scores. As can be seen in the table below, the model's distribution of residuals at the edge cases does not follow a Gaussian.
  - Derive ~50 more features from the dataset
  - Handle edge cases in the datasets, focusing on very low and high scores



- Model efficiency
  - Resolve the hyperparameter tuning bottleneck either by optimizing the process or removing the tuning altogether
  - Refactor the feature engineering code
  - Cull high-importance features from code
- Model Construction
  - Explore other models, such as LightGBM and XGBoost
  - Try an ensemble of multiple models

References

[1] Conijn, R., Cook, C., van Zaanen, M. et al. Early prediction of writing quality using keystroke logging. Int J Artif Intell Educ 32, 835–866 (2022). https://doi.org/10.1007/s40593-021-00268-w

[2] Malekian, Donia, et al. "Characterising Students' Writing Processes Using Temporal Keystroke Analysis." International Educational Data Mining Society (2019).