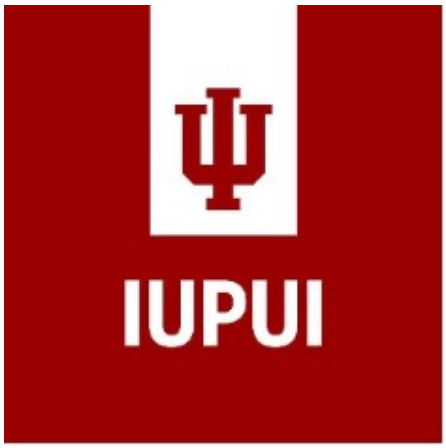
Linking Writing Processes to Writing Quality

CSCI 57300 Data Mining

By

Ayush Manojkumar Lodha
Dhiraj Pimparkar
Neeraj Gupta

Instructor
Prof. Mohammad Al Hasan



Indiana University - Purdue University Indianapolis
USA
Nov, 2023

Contents

1 Problem Statement

Does typing behavior affect the outcome of an essay?[1]

The objective of this data mining project is to investigate the impact of typing behavior on the overall quality of written essays. This research aims to develop predictive models based on a dataset of keystroke logs, capturing various writing process features. The challenge lies in understanding how diverse writing behaviors, such as planning, revisions, and pause patterns, affect the final quality of written content. Although previous studies have explored some process features, they often relied on limited datasets, and only a fraction of the potential process features have been examined. Therefore, the main problem to address is the need for a data-driven exploration of the relationship between writers' behaviors during the writing process and the resulting writing quality. This investigation could yield valuable insights for enhancing writing instruction, developing automated writing evaluation techniques.

2 Abstract

It's difficult to summarize the complex set of behavioral actions and cognitive activities in the writing process. Writers may use different techniques to plan and revise their work, demonstrate distinct pause patterns, or allocate time strategically throughout the writing process. Many of these small actions may influence writing quality. Even so, most writing assessments focus on only the final product. Data science may be able to uncover key aspects of the writing process. We aim to use process features from keystroke log data to predict overall writing quality. These efforts may identify relationships between learners' writing behaviors and writing performance. Additionally, given that most current writing assessment tools mainly focus on the final written products, this may help direct learners' attention to their text production process and boost their autonomy, metacognitive awareness, and self-regulation in writing.

3 Dataset

The dataset that we will be working with consists of two tables: sequential logs data and scores for each essay (train_logs and train_score). The first table, train_logs, has the data on the actions that were performed while the essay was being written. For example, when the user pressed a key, whether he deleted or inserted any character, etc. The second table, train_score, has scores awarded to that essay.

Number of samples	8,405,898
Total number of columns	11

Table 1: Train_logs

Number of samples	2,471
Total number of columns	2

Table 2: train_score

The description of the features present in the dataset is below-

Feature Name	**Feature Type (Categorical/Numerical)**	**Description**
id	Categorical	The unique ID of the essay
event_id	Categorical	The index of the event for the user, ordered chronologically
down_time	Numerical	The time of the down event in milliseconds
up_time	Numerical	The time of the up event in milliseconds
action_time	Numerical	The duration of the event (difference between down_time and up_time)
down_event	Categorical	The name of the event when the key/mouse is pressed
up_event	Categorical	The name of the event when the key/mouse is released

Feature Name	Feature Type (Categorical/Numerical)	Description
activity	Categorical	The category of activity to which the event belongs: • Nonproduction - The event does not alter the text in any way • Input - The event adds text to the essay • Remove/Cut - The event removes text from the essay • Paste - The event changes the text through a paste input • Replace - The event replaces a section of text with another string • Move From [x1, y1] To [x2, y2] - The event moves a section of text spanning character index x1, y1, to a new location x2, y2
text_change	Categorical	The text that changed as a result of the event (if any)
cursor_position	Numerical	The character index of the text cursor after the event
word_count	Numerical	The word count of the essay after the event
score	Target Feature	Can be considered as numerical and categorical based on the dataset's choice: Score of the essay from 0 to 6

4 Exploratory Data Analysis

We have tried to explore the dataset from various angles and have a few starting points that we can pursue in the final report.

4.1 Distribution of score vs features

We found the following insights while exploring the relationship of scores with different features,

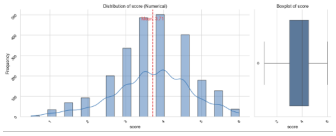
1. The data analysis reveals that the majority of essays receive scores ranging from 2 to 5, as indicated by the distribution graph of scores. (Refer to plots a, b & c)
2. Furthermore, a robust positive correlation is observed between the number of events associated with each essay and the assigned score. (Refer to plots a, b c)
3. Essays that extensively employ mouse actions tend to be associated with proficient users, and they generally receive scores ranging between 3.0 and 4.0. This suggests a positive correlation between the use of mouse actions and higher essay scores. (Refer to plots d & e)
4. The distribution of scores spans a normal range from 0 to 6, indicating a relatively balanced spread of scores across the dataset. (Refer to plots d e)

5. Users who frequently engage in media actions, and thus may be considered distracted, tend to score within the range of 3.0. This finding highlights a potential correlation between the use of media actions and the resulting essay scores, specifically in the context of distracted users. (Refer to plots d & e)

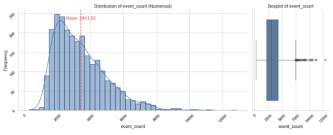
6. We can see that action time categories v.s scores are normally distributed for all instances. (Refer to plot f)

7. We can see that the ratio of Non-Production vs Remove/Cut is decreasing as the score increases, implying that the essay scoring higher is less non-productive. (Refer to plots g & h)

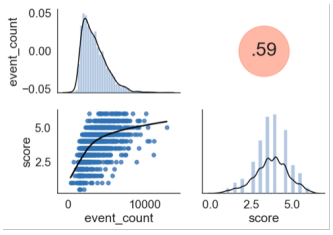
8. Replace, Paste, and Move are not used much by users during writing the essay when compared to the Input and Non-Production. (Refer to plot g & h)



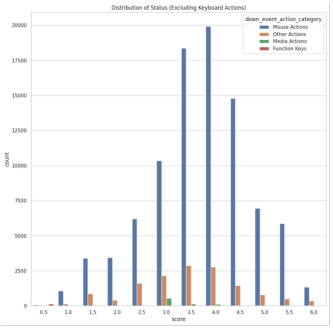
(a) Plot 1



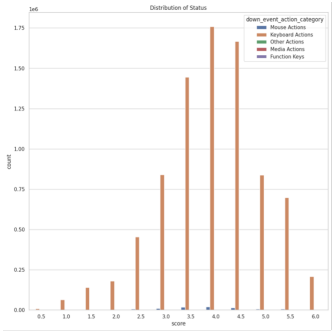
(b) Plot 2



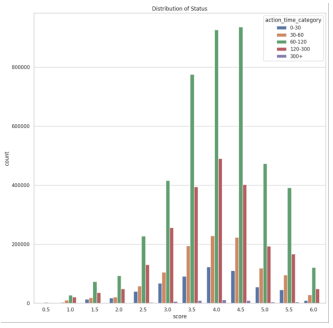
(c) Plot 3



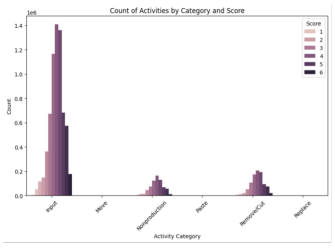
(d) Plot 4



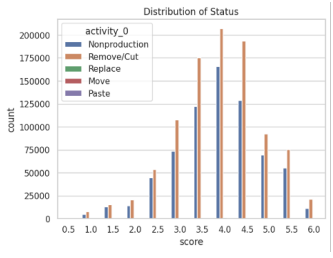
(e) Plot 5



(f) Plot 6



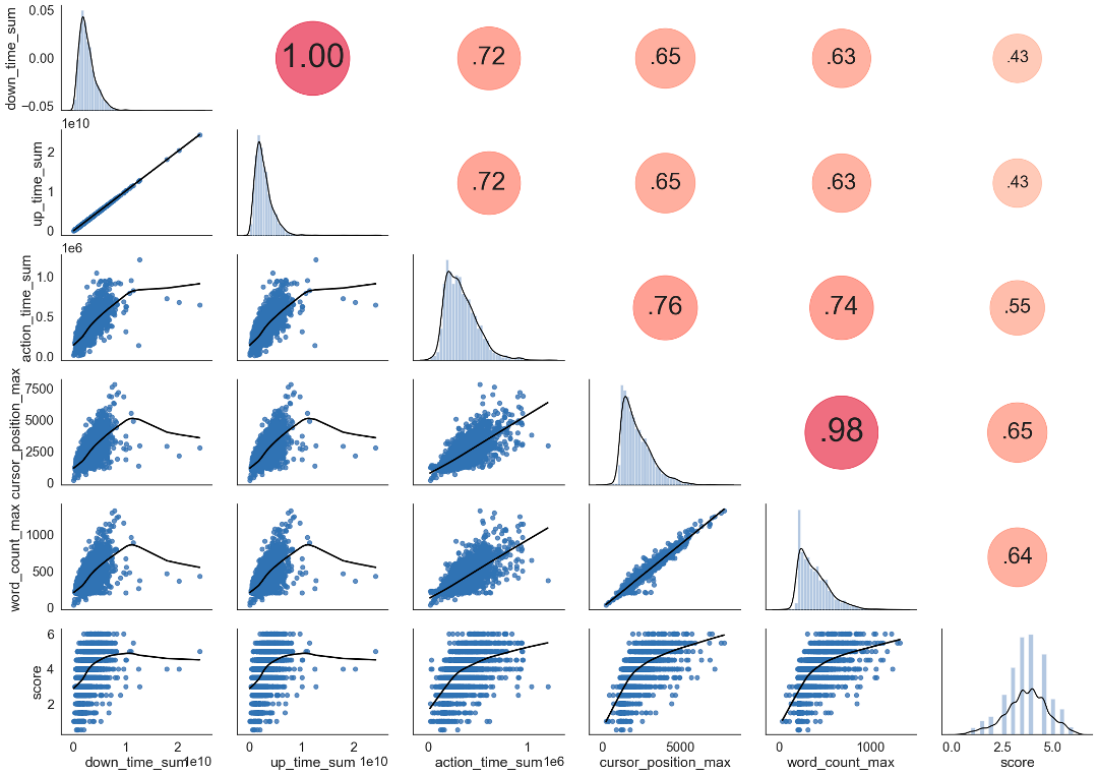
(g) Plot 7



(h) Plot 8

Figure 1: Exploratory plots

4.2 Distribution of Numerical Features (Aggregated for each essay) and Score



1. A robust relationship is observed between aggregated down_time and up_time, signifying a high degree of correlation between the two variables. To mitigate multicollinearity, during the feature engineering phase, it may be advisable to eliminate one of them.

2. The data exhibits substantial skewness, evident from the distribution of up_time and down_time at various data points.

3. Additionally, the presence of potential outliers becomes apparent when inspecting scatter plots. These outliers are individual data points that significantly deviate from the main cluster, suggesting the need for further domain-specific investigation.

4. Given the strong correlation and the presence of outliers, it becomes apparent that data pre-processing steps, such as log transformation or normalization of features, should be considered valuable techniques to enhance the modeling process. These pre-processing hints are essential for improving the reliability of our predictive models.

5 Proposed Approach/ Methodology

An important discovery emerges from the above analysis: the essay score can be treated as both a categorical and numerical feature, depending on the choice of the learning method or the model employed. This versatility offers valuable insights for selecting the most appropriate approach in our modeling and analysis.

5.1 Data Preprocessing and Feature Engineering

1. Initial data consists of 'train_logs' and 'train_scores'. We will process this data to create individual samples, where each sample represents the typing behavior of a user, with the corresponding label being the score of the essay they produced.

2. Sequential data will be converted into tabular format, ensuring each user's typing behavior features are represented separately

3. Feature extraction from the typing behavior data, which includes parameters such as pause patterns, frequency of additions and deletions, revision history, typing speed, and other relevant metrics.

5.1.1 Data Splitting

Split the dataset into training and validation sets to evaluate model performance effectively.

5.2 Model Selection

Depending on whether we approach this as a regression task (predicting essay score) or a classification task (e.g., classifying essays into quality categories), we will consider suitable machine learning models. We can use the following regression or classification model. We will implement cross-validation to ensure the model's robustness and to mitigate overfitting issues.

5.2.1 Regression Model

Regressors for Supervised Learning	**Examples**
Multilinear Regressors	Linear
Tree-based Regressors	Decision Trees
Ensemble-based Regressors	Random Forest Regressors
Boosting-based Regressors	GBM, LightGBM

5.2.2 Classification Model

Classifiers for Supervised Learning	**Examples**
Logistic Regression	Binary and Multiclass Classification
Decision Trees	Decision Trees for Classification
Random Forest	Ensemble of Decision Trees
Support Vector Machines (SVM)	Binary and Multiclass Classification
Gradient Boosting (e.g., XGBoost)	Ensemble Learning with Boosting
Neural Networks	Deep Learning for Classification

5.2.3 Hyperparameter Tuning

The model's hyperparameters will be fine-tuned to optimize its performance. Again, this will be based on the nature of the model we decide to use.

5.3 Model Evaluation

The evaluation metric will be based on the nature of the problem, for example, if we analyze the problem as a classification problem, a confusion matrix can be used. However, if we treat this problem as a regression problem, then Root Mean Square Error (RMSE) will be used to measure the model's accuracy in predicting essay scores.

References

[1] The Learning Agency Lab. Linking writing processes to writing quality. URL: https://www.kaggle.com/competitions/linking-writing-processes-to-writing-quality, 2023.