Predicting submission behavior in Kaggle competitions

Kim, Minjun
20195024

In this essay, I wrote about a methodology for predicting the submission behavior of participants in Kaggle competitions. Through data analysis and feature engineering, I tried to find the factors that influence participants' submission patterns. In addition to this, I wanted to use the Hawkes Process and survival analysis to understand how participants' past behaviors and personal characteristics influence their future submission behaviors.

1 Data Distribution in *UserProfile*

Status : Graduate (Grad) and undergraduate (Undergrad) students are roughly equally distributed, encompassing students at various academic stages **(Year)**.
 # Credits : A significant number of credits are earned by most students, typically ranging between 15 to 18 credits.
GPA : A high distribution of GPAs is observed, with the majority of students maintaining a GPA of 3.0 or above.
IQ : IQ scores exhibit a wide distribution, spanning between 110 and 130.
Study Hour : Study hours are distributed between 17 and 28 hours per week.
MBTI : A roughly equal distribution is observed across different MBTI types.

2 Feature Engineering

submission_count (Number of Submissions): This feature indicates the number of submissions for each participant. It reflects the level of activity of the participants, with more active participants potentially having a higher likelihood of resubmitting.
average_event_interval (Average Time Interval Between Events): This feature calculates the average time interval between submission events for each participant. It indicates how frequently a participant submits and provides important information for the analysis of the Hawkes Process.
time_since_last_submission (Time Difference from the Last Submission to the Current Time): This feature reflects the frequency of a participant's activity up to the current moment. From the perspective of the Hawkes Process, a short time interval signifies recent activity, suggesting that the intensity of the stimulus for future activities (such as resubmissions) might be high.
time_until_deadline (Time Difference from the Last Submission to the Deadline): This feature indicates the remaining time a participant has to make additional submissions. As the deadline approaches, the submission activity of the participants may increase. This suggests that in the Hawkes Process, the temporal pressure as the deadline nears could provide a stronger stimulus for future submissions.

last_accuracy_score: This feature indicates the accuracy score of the participant's most recent submission as of today's date. Participants who have lower score might get impetus to make additional attempts of submissions in order to improve their scores. In a Hawkes Process analysis, this feature might help us understand how the score from the latest submission, up to today, could affect the likelihood and intensity of participants' future submissions.
Other features: I would like to deal with qualitative variables, for example, **Status, Department**, and **MBTI** by using one-hot encoding for the survival analysis. This will convert them to numerical form allowing the model to use them effectively.

3 Model Selection

I believe the Cox Proportional Hazards Model is well-suited for this analysis. The 'survival time' should be defined as the interval from the current moment, 18:00 on the 8th of November, until the end of the competition. This model will be used to determine the hazard ratio, predicting the timing of potential future submissions by participants. By integrating the features I've developed, this model aims to offer a detailed understanding of the various elements that might affect a participant's decision to make additional submissions before the competition's deadline.

I have considered the Hawkes Process in the feature engineering, if I can develop a Hawkes Process model, it would enable us to understand the impact of each submission on the likelihood of subsequent ones. Therefore, I believe that combining the Hawkes Process Model with the Cox Proportional Hazards Model could potentially improve the accuracies of the model as well as the effectiveness of survival analysis.

4 Model Evaluation

The Concordance Index (c-index) would be a vital metric in evaluating my model, particularly the Cox Proportional Hazards Model. The c-index assesses how well the model differentiates between participants with a higher likelihood of resubmitting and those with a lower likelihood in terms of the timing of their resubmissions.

Finally, I plan to analyze the importance of the features by identifying which ones most significantly impact the survival probabilities, which in this context refer to the probabilities of resubmission.