Appendix A Extension to feature engineering

A.1 Deriving features from consumer consumption

This section details how the new Indicators/features have been added to complement the well-established SRL features.

As mentioned in Section 3.2, we drew some parallels between customer consumption analysis in the finance domain to those of treating students (as customers) consuming the learning materials provided in the course. Specifically, features of customer consumption to incorporate more sophisticated ways of capturing students' self-regulation behavior. Human consumption patterns are usually studied across time and space. In particular, these patterns are captured by Exploration, exploitation, and plasticity. These traits can be proxied by diversity, loyalty, and regularity metrics respectively. These metrics have been proven to play an important role in predicting the financial outcomes of individuals. In this study, we adapt these metrics for the SRL context to construct a behavioral profile of each student.

The extent to which customers spread their transactions (in our case sessions) over time (temporal diversity) is diversity in consumption. A higher diversity corresponds to spreading the transactions almost equitably over the space and time bins [1]. Originally, diversity is measured based on the formula presented in Equation 1.

$$D_i = \frac{-\sum_{i=j}^{N} P_i j \, log_2 \, P_i j}{log_2 \, M} \tag{1}$$

where $P_i j$ is the fraction of transactions in bin j for user i, N is the total number of bins with M of them being non-empty.

For the educational context, only temporal diversity is valid. We utilize temporal diversity with weekly bins. We use learning sections as transactions. The calculated diversity metric is in the range $[0, 1]$ with larger diversity corresponding to spreading studying more over the semester. For a more intuitive label, 'Diversity' is replaced with 'Uniformity' as it better represents a student uniformly spreading effort throughout the semester.

The loyalty metric represents how transactions are spread across different temporal bins. A higher value in loyalty corresponds to having most of the transactions within top-N of the bins. It is measured based on the formula displayed in Equation 2.

$$L_i = \frac{f_i}{\sum_{j=1}^{N} P_{ij}} \tag{2}$$

where f_i is the combined fraction of all transactions that occur in the top three most-frequented bins, P_{ij} is the fraction of transactions in bin j for user i.

To be more representative in the educational domain, this metric 'loyalty' is named as 'Bingeing'. It expresses concentrating the study efforts in a small number of time slots. We calculate temporal bingeing with weekly bins, with learning sessions as transactions.

Regularity measures the differences between behavioral patterns over shorter and longer periods of studying and is calculated based on the formula in Equation 3 for consumer regularity.

$$R_i = 1 - \frac{\sqrt{(D_i^1 - D_i^T)^2 + (L_i^1 - L_i^T)^2}}{\sqrt{2}} \tag{3}$$

where D_i^1 and L_i^1 are the diversity and the loyalty in a shorter period (e.g., one month), respectively, D_i^T and L_i^T is the diversity and the loyalty in the entire period, respectively.

Temporal regularity is calculated with D_i^1 and L_i^1 set to the Uniformity (diversity) and Bingeing (loyalty) in the semester weeks, respectively, and D_i^T and L_i^T set to the Uniformity and Bingeing in the entire course, respectively.

A.2 Correlation maps of features considered

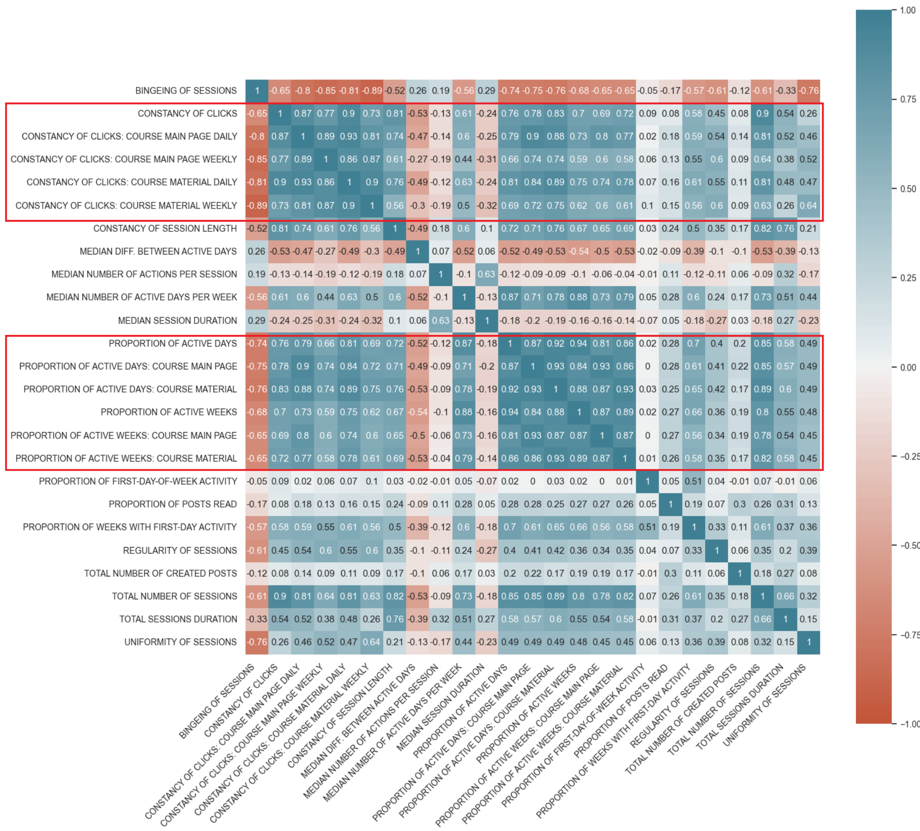


Fig. 1: Correlation map before feature selection

The feature correlation map is displayed in Figure 1. Some of the features are highly correlated and essentially represent the same information (two features groups marked by red rectangles). Including such features can mask important interactions from being discovered. Moreover, we follow Occam's razor principle and prefer a less complex model. Hence, in the aforementioned two groups, we drop all the features except of 'Constancy of clicks' and 'Proportion of active days'. The correlation matrix after feature selection is displayed in Figure 2.

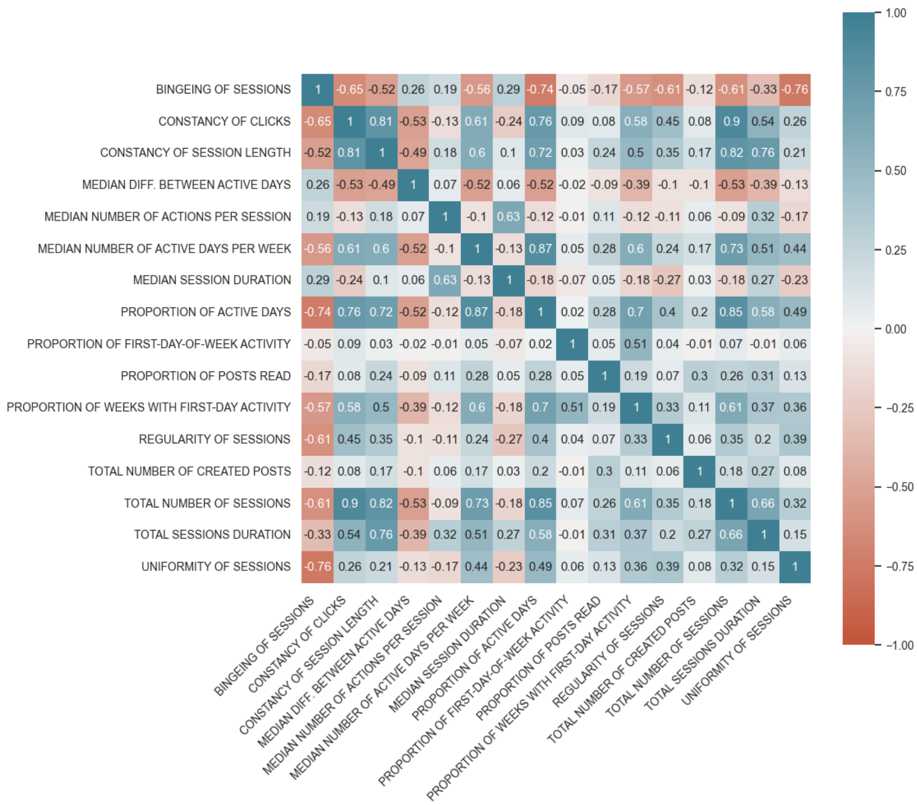
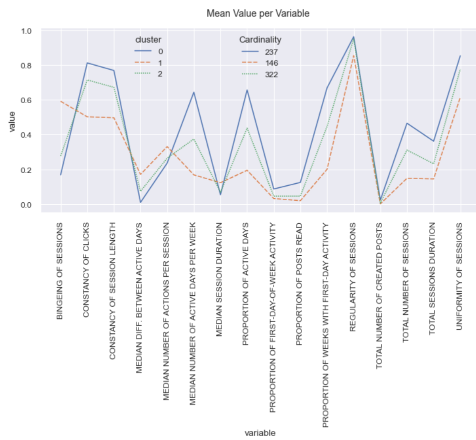


Fig. 2: Correlation map after feature selection

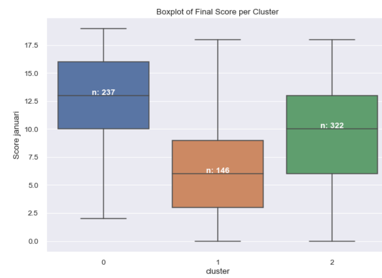
Appendix B External Validation

This section presents the external validation of the results but considering both dimensions of courses and time.

B.1 Validation across time - Accountancy AY 1920

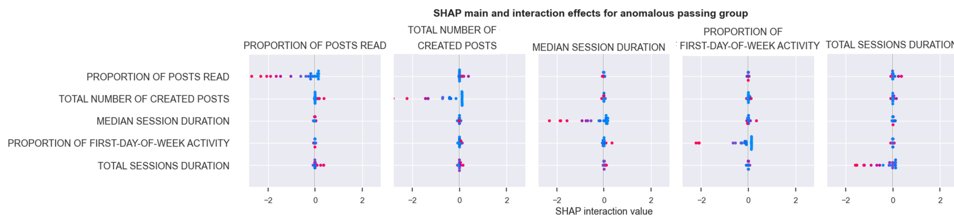


(a) Mean indicators deviations per cluster

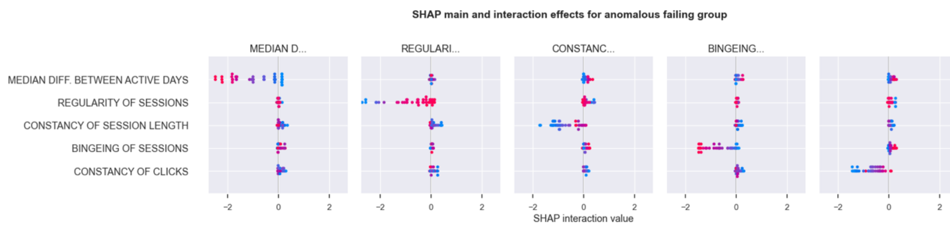


(b) Boxplot: clusters vs. Exam grades

Fig. 3: Clustering solution



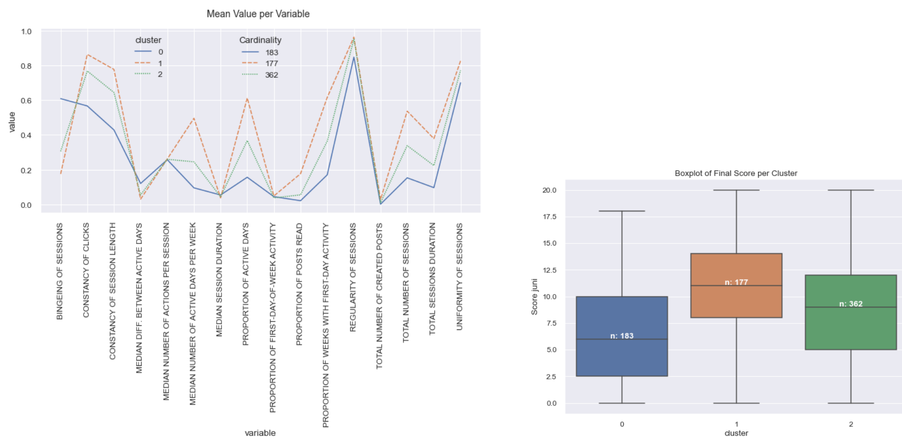
(a) Anomalous passing students



(b) Anomalous failing students

Fig. 4: SHAP summary plots. Each point represents one student and its color corresponds to the feature value (see the legend). The features are displayed in the order of importance from top to bottom.

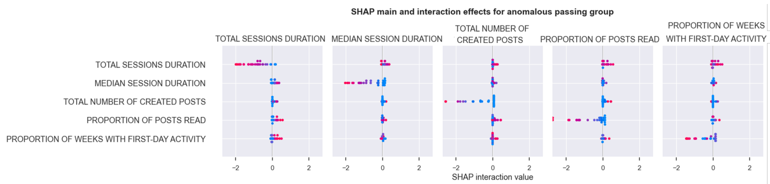
B.2 Validation across courses - Macroeconomics AY 1819



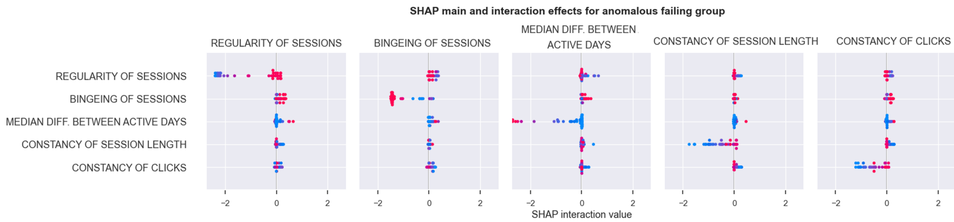
(a) Mean indicators deviations per cluster

(b) Boxplot: clusters vs. Exam grades

Fig. 5: Clustering solution



(a) Anomalous passing students



(b) Anomalous failing students

Fig. 6: SHAP summary plots. Each point represents one student and its color corresponds to the feature value (see the legend). The features are displayed in the order of importance from top to bottom.

References

1. V. K. Singh, B. Bozkaya, and A. Pentland, “Money walks: implicit mobility behavior and financial well-being,” *PloS one*, vol. 10, no. 8, p. e0136628, 2015.