EPFL

Machine Learning For Behavioural Data

Project - Milestone 4

Franziska Anna Von Albedyll
Fletcher Collis
Oscar Goudet

April 30, 2025

Contents

1 Part 1 - Inspiration and References

The inspiration for our project primarily came from the content and themes explored in the course : Machine Learning for Behavioral Data [1] and we did not have any pre-knowledge of this particular domain. We were particularily drawn to the challenge of predicting student performance using behavioural data which we believed would be an area that combines elements of machine learning, education, and behavioral science. This therefore motivated us to explore existing literature on how online learning management systems activity, like the GoGymi data at our disposal, can correlate with academic outcomes.

A key paper that informed us on the domain was ”(m)Oodles of Data: Mining Moodle to Understand Student Behaviour” by Casey and Gibson (2010) [2]. In this study, the authors examined Moodle activity logs. They found that specific types of student engagement showed positive correlations with higher academic performance. Such as daily logins, volume of material accessed, and usage over weekends. Their work provided clear evidence that student interaction patterns with course materials can be strong predictors of final outcomes. These findings helped guide our decision to model student performance based on similar behavioural variables. It is worth noting that this study also revealed that some students who performed very well academically exhibited relatively low levels of Moodle activity.

In addition to behavioural modeling, handling missing data was another important component of our work. Since we were dealing with uncleaned real data, many of our features were incomplete for certain users, particularly past performance metrics. The strategies we applied were inspired by techniques discussed in the Applied Data Analysis (ADA) course [3], including simply dropping them, mean imputation and K-nearest neighbor imputation. Understanding and evaluating the impact of these methods on model robustness was a key focus.

Feature extraction also played a central role in our project. We engineered variables to summarize recent behaviour, such as activity frequency, diversity, and timing, with the goal of capturing meaningful patterns that could predict outcomes. While we found less formal academic literature specifically about behavioral feature design in this context, our process was guided by exploratory data analysis and the goal of creating interpretable, predictive indicators. Future work could benefit from integrating more established methods like the ones explored in the paper "Comparative Analysis of Feature Selection and Extraction Methods for Student Performance Prediction across Different Machine Learning Models" (Hemdanou et al., 2024) [4]. In their study, they examined various feature extraction techniques like Principal Component Analysis (PCA) and Variational Autoencoders (VAE), which helped uncover hidden patterns in student behavior.

To incorporate user-level heterogeneity, we also tried to implement a linear regression model with mixed effects. This model incorporates user groupings as random effects. The approach allowed us to partially account for unobserved differences between users while still leveraging global patterns and our implementation was informed mainly by the course [1].

2 Part 2 - Modeling and Evaluation

The central research question guiding our work was: **Can we predict student performance based on their activity behavior?** To address this, we aimed to balance three core goals : maximizing predictive accuracy, preserving as much data as possible and maintaining model interpretability. We explored several modeling strategies, each offering different advantages and disadvantages between these priorities.

For evaluation, we also chose to look at more than one performance metric to get a fuller picture of each model's strengths and weaknesses. We used R-squared to measure how much of the variance in student performance could be explained by our features. This gives us a sense of how well the model fits the data overall. We also considered the Root Mean Squared Error (RMSE) to assess how large the prediction errors were on average. Finally, we included directional accuracy, which we defined and evaluates whether the model correctly predicts if a student's performance will be positive or negative. This metric was particularly important to us, as it reflects the model's ability to distinguish between higher- and lower-performing students, which is a core goal of our project.

Here are the ideas behind each major decision:

Baseline OLS with Drop–NaN (Method 0) We began with a straightforward linear regression on fully observed rows, dropping any exam entries missing time-in-minutes or past-exam averages. This "clean" subset preserved interpretability and set a clear performance benchmark ($R^2 \approx 0.267$, directional accuracy ≈ 0.703) and showed that the best predictor by far was the average performance on last exams (coefficient of 13.13 vs. 1.96 for the second-best coefficient). The downside of losing roughly half the data motivated us to explore ways to retain more observations.

Mixed-Effects by Clustering (Method 1) To capture unobserved heterogeneity, we clustered students on their mean engagement profiles and fitted a random-intercept model per cluster. This allowed each segment its own baseline, borrowing strength within similar learner groups. Although conceptually appealing, the added complexity and parameter overhead did not lead to any gain in predictive performance.

Median Imputation (Method 2) Next we filled missing values with the median value for the corresponding feature within each exam. This preserved all rows, avoided data loss, and kept imputation bias minimal. While it performed close to the drop-NaN baseline, it could not match the pure-case model's R^2 and blended first-exam students into a uniform median estimate.

KNN Imputation (Method 3) Seeking a more personalized fill, we used k-nearest neighbors (within the same test) to estimate each first-exam's "average past score." Though more tailored than a flat median, this method injected additional variance, neighbor scores were noisy and actually reduced R^2 relative to simpler imputations, showing that similarity in engagement did not reliably predict first-exam performance.

Gradient Boosting with Native NaN Handling (Method 4) We then shifted to a tree-based learner (HistGradientBoosting) that treats NaNs as a special branch at each split. This allowed us to feed the model raw features (missingness included) without any imputation, capturing nonlinear interactions and retaining 100% of the data. Despite its flexibility, the boosting model's R^2 did not surpass the drop-NaN OLS.

Across all models and experiments, one insight stood out clearly: the most important predictor of a student's performance was their average performance on past exams, a finding that consistently dominated behavioural features in explanatory power. Interestingly, the simplest model, a basic OLS regression dropping problematic data, outperformed more complex methods, highlighting the strength of clean data and interpretable baselines. Additionally, we observed that some behavioural features, such as average activities per day, showed a negative correlation with performance. This mirrors a surprising observation made by Casey and Gibson [2], where high activity levels in certain courses were associated with lower performance which likely reflects inefficient or unfocused engagement. These findings suggest that more activity is not always better, and that combining performance history with nuanced behavioural features offers the most promise for future modeling.

3 Part 3 - Contributions

Franziska contributed significantly to the project by working on data cleaning, preprocessing, cleaning up the main notebook, and contributing to Milestone 5.

Fletcher was responsible for data cleaning, creating the main notebook, model evaluation, and writing the report.

Oscar focused on feature extraction, model creation, model evaluation, and also contributed to the report.

Overall, the team collaborated effectively. Each member contributed meaningfully and took responsibility for different aspects of the project. We are pleased with our progress and the results achieved in this milestone.

References

[1] EPFL Machine Learning for Education Laboratory. *Machine Learning for Behavioral Data (MLBD) - CS-421 (Spring 2025)*. Accessed: 2025-04-30. École Polytechnique Fédérale de Lausanne (EPFL). 2025. URL: https://github.com/epfl-ml4ed/mlbd-2025.

[2] K. Casey **and** P. Gibson. *(m)Oodles of data: Mining moodle to understand student behaviour*. International Conference on Engaging Pedagogy, Maynooth University Research Archive. 2010. URL: https://mural.maynoothuniversity.ie/id/eprint/10180/.

[3] EPFL Data Science Lab. *Applied Data Analysis (ADA) - CS-401 (Fall 2024)*. Accessed: 2025-04-30. 2024. URL: https://epfl-ada.github.io/teaching/fall2024/cs401/.

[4] Abderrafik Laakel Hemdanou **andothers**. *Comparative analysis of feature selection and extraction methods for student performance prediction across different machine learning models*. Published in *Computers and Education: Artificial Intelligence*, October 2024. 2024. URL: https://doi.org/10.1016/j.caeai.2024.100301.

We acknowledge the use of ChatGPT for enhancing the clarity and precision of the text in this paper and for translation support. It was also used to help with code debugging. However, ChatGPT was not involved in data analysis or in deriving the project's conclusions.