Predicting Student Dropout and Academic Performance in Online Learning using Temporal and Survival Models

Sahanaa Dinesh

Supervised by: Dr. Felipe Campelo

June 20, 2025

Abstract

Online education platforms continue to grow, yet student retention remains a key challenge. This project aims to build a predictive system using the Open University Learning Analytics Dataset (OULAD) to identify students at risk of dropping out early during a course. Specifically, it focuses on modeling time-to-dropout using survival analysis techniques, while optionally extending to predict final academic outcomes.

The ultimate goal is to deliver a model that acts as an early warning system using weekly engagement data, demographic indicators, and assessment performance. The model will not only predict whether a student is likely to drop out, but also estimate when that is most likely to happen. If successful, this system could guide proactive interventions and resource allocation in online education.

Ethics statement: This project fits within the scope of the blanket ethics application, as reviewed by my supervisor Dr. Felipe Campelo. I have completed the ethics test on Blackboard. My score is 15/15.

1 Project Plan

The rapid growth of online education has transformed access to learning, but it has also amplified challenges with student engagement and retention. Platforms such as MOOCs and virtual learning environments (VLEs) report dropout rates exceeding 80%, raising questions about the underlying causes and how predictive analytics can be used to address them. Identifying at-risk students early and estimating when they are likely to withdraw can help institutions intervene effectively, improving academic outcomes and student well-being.

This project proposes a hybrid predictive system based on the Open University Learning Analytics Dataset (OULAD), a comprehensive and anonymized collection of student-level data including demographics, VLE activity, assessment scores, and registration history [1]. The system will combine two predictive objectives: (1) forecasting whether and when a student will drop out using survival analysis, and (2) estimating final course outcomes using classification. The proposed approach integrates interpretable survival models (Cox and Bayesian variants) with sequence-based deep learning (LSTM) and multi-task learning to enable robust, early, and explainable predictions.

The motivation stems from several gaps in existing research. Traditional dropout prediction models, such as decision trees and ensemble classifiers [2], focus on binary classification without accounting for the timing of events. While some recent work has applied Cox models to predict dropout over time [3, 4], these often overlook the rich temporal patterns found in weekly VLE engagement logs. Conversely, LSTM models have been used to model student clickstream sequences [5, 6], but often lack interpretability and do not handle censored data—a key requirement for time-to-event modeling.

To bridge this gap, this project proposes a tiered architecture. The core component is a Cox Proportional Hazards model with stratification by course module and presentation. This baseline will be extended using shared frailty Cox models and Bayesian survival regression [4]. These models will ingest features such as age, education level, previous attempts, weekly click counts, and cumulative assessment scores. Survival targets will be derived from the `date_unregistration` column in `studentRegistration.csv`, with censoring applied for retained students.

In parallel, the project will experiment with an LSTM-based survival model that encodes week-by-week VLE engagement. Inspired by He et al. [7], this model will learn temporal dropout risk profiles and may include explainability layers such as SHAP for model transparency. To mitigate class imbalance (as dropouts are often the minority class), GAN-based oversampling techniques [8] will be applied to generate synthetic dropout trajectories. Features such as inactivity duration and assessment delays [9] will be incorporated to enrich the time series.

Another key innovation is multi-task learning. Since dropout and final grade prediction share behavioral and demographic predictors, a shared neural encoder will support two heads: one for survival analysis and another for classification. This joint learning framework is expected to improve sample efficiency and provide consistent early-warning outputs [10]. Optional extensions include using unsupervised clustering to label low-engagement profiles [11], and feeding those labels as latent risk priors into the survival models.

Evaluation will follow standard survival metrics, including the Concordance Index (C-index), Integrated Brier Score, and Kaplan-Meier plots to visualise risk stratification. For classification, accuracy, macro F1-score, and area under the precision-recall curve (AUPRC) will be used. Benchmarking will compare the performance of static Cox models, Bayesian extensions, and LSTM-based survival architectures.

In summary, this project aims to design a hybrid system that delivers both accurate and interpretable dropout predictions, while offering insight into final academic outcomes. By leveraging temporal modeling, uncertainty estimation, and multi-objective learning, the system can serve as a data-driven early intervention framework for online education providers.

References

[1] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," *Scientific Data*, vol. 4, p. 170171, 2017.

[2] M. Hussain, T. Dahanayake, and G. H. Abeywardena, "Student performance prediction using decision tree technique," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 10, 2018.

[3] C. Martínez-Carrascal, V. Esteve, and J. García-García, "An explainable survival analysis model to predict dropout risk in online education," *Educational Technology & Society*, vol. 26, no. 1, pp. 45–60, 2023.

[4] C. Masci, N. Menachemi, and F. Paolucci, "Predicting student dropout in online courses using survival analysis," *Computers & Education*, vol. 194, p. 104693, 2023.

[5] M. Fei and Q. Ye, "Temporal models for predicting student dropout in massive open online courses," in *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics*, 2015, pp. 256–264.

[6] R. Liu, Q. Xu, and X. Zhang, "Predicting student dropout in moocs using lstm networks," *International Journal of Emerging Technologies in Learning*, vol. 13, no. 10, pp. 90–103, 2018.

[7] J. He, J. Bailey, B. I. P. Rubinstein, and R. Zhang, "Identifying at-risk students in online learning environments: A machine learning perspective," *IEEE Transactions on Learning Technologies*, vol. 13, no. 3, pp. 549–561, 2020.

[8] M. M. Ahsan, R. Hossain, and P. S. Jenkins, "Using gans to balance student datasets for dropout prediction," *Proceedings of the Educational Data Mining Conference (EDM)*, 2021.

[9] N. Hlioui, K. Drira, and M. S. Gouider, "A withdrawal prediction model of at-risk learners based on behavioural indicators," *International Journal of Information and Communication Technology Education*, vol. 17, no. 1, pp. 70–88, 2021.

[10] H. Almaazmi, F. Alkaabi, and M. Khalil, "A multimodal deep learning approach to predict dropout in moocs," *Electronics*, vol. 12, no. 3, p. 731, 2023.

[11] S. Palani and C. Venugopal, “Clustering student engagement profiles using temporal interaction data,” in *Springer International Conference on Educational Technology and Learning*, 2021, pp. 123–135.

A Project Timeline

- Week 1–2 (2–13 June): Dataset familiarisation, literature review, problem definition
- Week 3–4 (16–27 June): Write project plan, set up GitHub repo, conduct EDA, merge and clean data
- Week 5–6 (30 June – 11 July): Implement Cox survival models, feature engineering
- Week 7–8 (14–25 July): Evaluate alternative models (e.g., LSTM, classification fallback)
- Week 9–10 (28 July – 8 August): Finalise model, run evaluations, generate plots
- Week 11–13 (11–31 August): Write dissertation, polish code and documentation, submit project

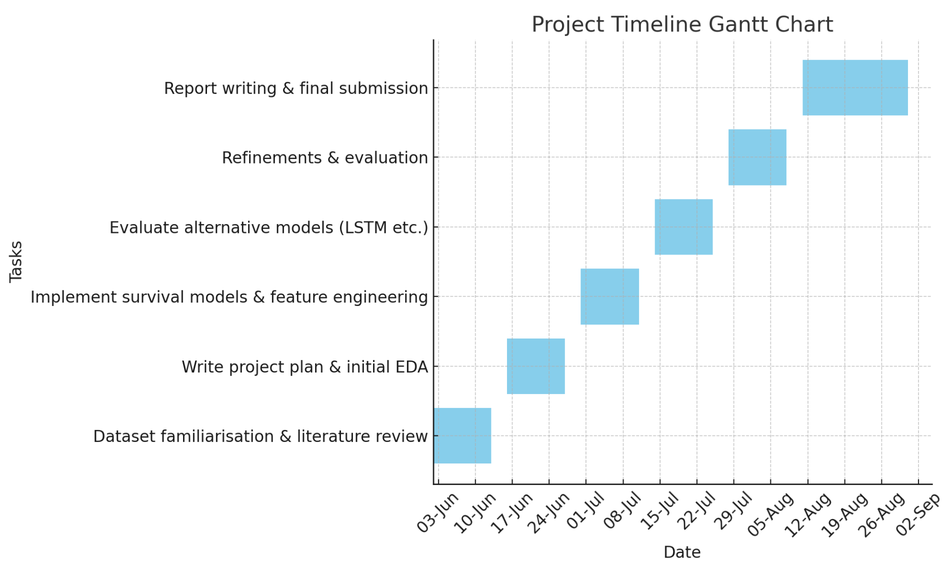

Figure 1: Project Timeline Gantt Chart: Showing planned activities across June–August 2025.

B Risk Assessment

Managing risk is an important aspect of this project, especially due to the sequential dependencies between data preprocessing, model implementation, and analysis. Table 1 summarises the key anticipated risks, their likelihood and impact, and the mitigation strategies that will be followed to minimise their effect on the project timeline and quality.

Table 1: Summary of Risk Mitigation Strategies

Risk	**Likelihood**	**Impact**	**Mitigation**
Survival models too complex or perform poorly	Medium	High	Build a backup classification model (e.g., XGBoost) as a fallback
Sparse or inconsistent VLE data	Medium	Medium	Use weekly feature aggregation and imputation for missing data
Time limitations for deep models (e.g., LSTM)	Medium	Medium	Prioritise Cox models first; only explore LSTM if ahead of schedule
Report writing delayed due to technical work	Low	High	Start drafting sections (literature, methodology) in parallel from July