



INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact Factor: 6.078

(Volume 12, Issue 5 - V12I5-1256)

Available online at: <https://www.ijariit.com>

Uncertainty-Calibrated and Explainable Machine Learning for Early Prediction of Academic Performance and Identification of At-Risk University Students

Mohammad Zaid

mohammad.zaid2025@vitstudent.ac.in

Vellore Institute of Technology, Vellore,
Tamil Nadu

Steave Saji

steave.saji2025@vitstudent.ac.in

VIT Vellore

Rayhan Abdulla Thalsim

rayhan.abdulla2025@vitstudent.ac.in

Vellore Institute of Technology, Vellore,
Tamil Nadu

ABSTRACT

Early identification of students who may experience unsuccessful academic outcomes can support timely academic guidance, provided that predictions are evaluated for both discrimination and reliability. This study evaluates a Logistic Regression workflow using the Open University Learning Analytics Dataset (OULAD). Student outcomes were converted to a binary target, with Pass and Distinction treated as successful outcomes and Fail and Withdrawn as unsuccessful outcomes. Models were evaluated at weeks 2, 4, 6, and 8 using student characteristics, enrolment information, and virtual learning environment (VLE) activity. A second configuration added assessment-related features. The best single-split result occurred at week 8 for the activity-plus-assessment configuration (accuracy = 0.7789, F1 = 0.7796, ROC-AUC = 0.8654). Its Brier score and log loss were 0.1463 and 0.4394, respectively. Across confidence thresholds, selective accuracy increased as coverage decreased; at week 8 and a threshold of 0.90, the activity-plus-assessment configuration retained 29.7% of cases with 97.1% accuracy. Across three random seeds, its week-8 mean ROC-AUC was 0.8629 (SD = 0.0029). These findings suggest that later cutoffs and assessment-related information can improve predictive performance in this dataset. However, the availability time of assessment scores must be verified to rule out temporal leakage, and results require external validation before practical deployment.

Keywords: Educational Data Mining, Early Prediction, Student Performance, Logistic Regression, Calibration, Selective Prediction, Explainability, Oulad

1. INTRODUCTION

Universities increasingly use learning-management systems and student information systems to understand engagement and identify students who may benefit from additional support. Educational data mining uses these records to model academic outcomes, but a useful early-warning system must do more than assign a class label. It should provide meaningful probability estimates, indicate when its predictions are uncertain, and allow educators to inspect which features contribute to a prediction.

Prior work has demonstrated the value of educational data mining for student-performance prediction, while also highlighting the importance of choosing features and prediction times that reflect realistic intervention opportunities. OULAD provides anonymized information on student characteristics, assessments, registrations, and interactions with a virtual learning environment, making it suitable for studying prediction over time [1,2]. Logistic Regression is used here as a transparent baseline that produces probabilities and coefficients that can be inspected.

This study evaluates a reproducible workflow across four early-semester cutoffs. It compares a feature set based on student and VLE activity information with a second set that adds assessment-related variables. Beyond classification metrics, it examines Brier score, log loss, confidence-based selective prediction, coefficient-based interpretation, and robustness to alternative random splits.

2. RELATED WORK

Educational data mining applies statistical and machine-learning methods to educational records to identify patterns and predict outcomes. The OULAD dataset introduced by Kuzilek, Hlosta, and Zdrahal provides a documented benchmark containing demographic information, assessment records, registration information, and VLE interactions [1]. The dataset portal provides the underlying data and documentation [2].

Machine-learning models can achieve useful discrimination while still producing probabilities that do not correspond closely to observed event frequencies. Brier score and log loss therefore complement classification metrics, and calibration curves can reveal whether probability estimates align with observed outcomes. Niculescu-Mizil and Caruana compared probability-estimation behaviour across supervised learning methods and motivated explicit evaluation of probabilistic predictions [4].

This work contributes an integrated experimental comparison of early cutoffs, activity-only versus activity-plus-assessment features, confidence-based selective prediction, and split robustness within a Logistic Regression workflow. It should be viewed as a dataset-specific empirical study rather than evidence that a deployed student-risk system would generalize to other institutions.

3. DATASET AND OUTCOME DEFINITION

The analysis uses the Open University Learning Analytics Dataset (OULAD), which includes student-course records and linked tables for assessments, student assessment submissions, registration, and VLE interactions [1,2]. The project records 32,593 student-course records. The observed final outcome counts used in the project were 12,361 Pass, 10,156 Withdrawn, 7,052 Fail, and 3,024 Distinction.

For binary prediction, Pass and Distinction were assigned to the positive (successful) class, while Fail and Withdrawn were assigned to the negative (unsuccessful) class. This target supports a simple early-warning framing but combines outcomes that may have different meanings; withdrawal is not necessarily equivalent to failing an assessment.

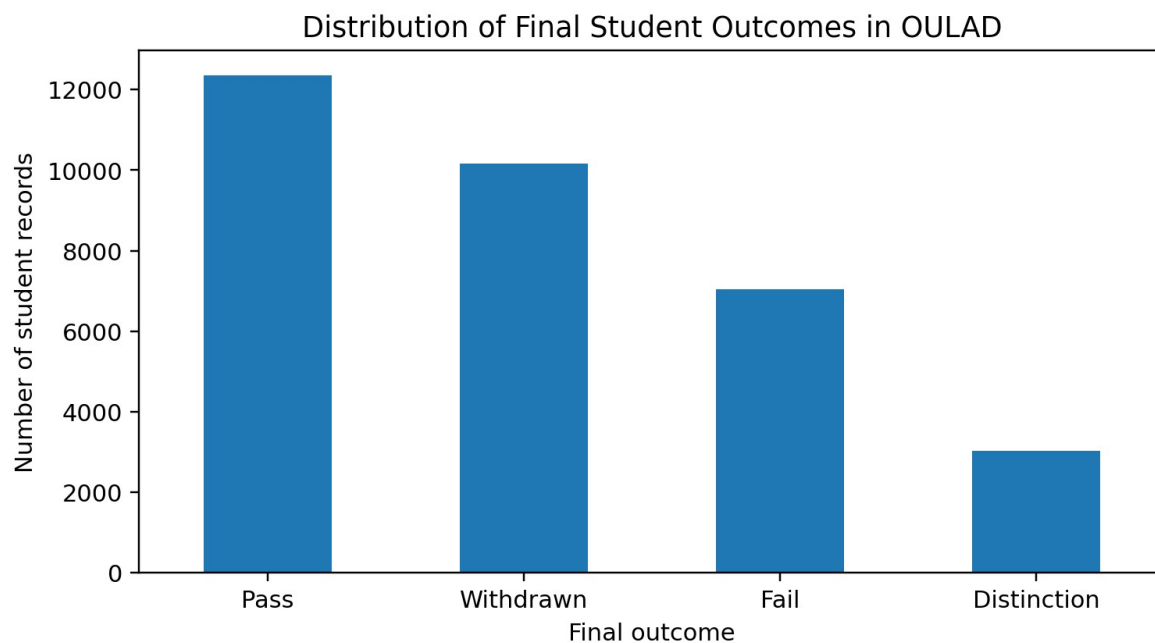


Figure -1

4. METHODOLOGY

4.1 Prediction cutoffs and feature configurations

Four prediction cutoffs were evaluated: week 2, week 4, week 6, and week 8. VLE activity was summarized up to each cutoff, including total clicks, active days, and distinct resources. Student/course characteristics and enrolment-related variables were also used. The activity-only configuration used these predictors without assessment-derived features. The activity-plus-assessment configuration added assessment-related summaries, including mean score, assessment count, and an indicator for whether an assessment was represented.

A critical temporal-validity issue remains: scheduled assessment dates and submission dates do not by themselves establish when a score was graded and available to a system. Therefore, assessment-related results should be interpreted as provisional until score-availability timing is verified against the prediction cutoff.

4.2 Model and preprocessing

Logistic Regression was trained using scikit-learn; no externally pretrained model was used. Numeric variables were imputed using the median, with missingness indicators enabled where configured. Categorical missing values were imputed using the most frequent category and categorical variables were one-hot encoded. Numeric features were standardized. Preprocessing and classification were implemented in a pipeline to apply transformations consistently.

4.3 Train-test separation and robustness

The main evaluation used a group-based split with student identifier as the grouping variable, a 20% test size, and random seed 42, reducing the chance that records for the same student occur in both train and test sets. Robustness experiments used seeds 21, 42, and 84. These repeated splits assess sensitivity to partitioning; they are not a substitute for external validation on an independent dataset.

4.4 Evaluation metrics

Accuracy measures the fraction of correctly classified test examples. F1-score is the harmonic mean of precision and recall for the positive class. ROC-AUC measures ranking discrimination across classification thresholds. Brier score is the mean squared error of predicted probabilities, while log loss penalizes inaccurate probability estimates, particularly confident errors; lower values are generally preferred for both. Calibration curves compare mean predicted probability with observed positive-outcome frequency.

4.5 Confidence-based selective prediction

For a predicted positive-class probability p , confidence was defined as $\max(p, 1-p)$. Predictions were retained only when confidence met thresholds of 0.55, 0.60, 0.65, 0.70, 0.80, or 0.90. Coverage is the proportion of test cases retained, and selective accuracy is accuracy among retained cases. Both must be reported because higher accuracy may be achieved by withholding a larger fraction of cases.

4.6 Explainability

The fitted Logistic Regression coefficients were inspected as a global, model-specific explanation. Positive coefficients increase the model's log-odds for the encoded successful class, conditional on the other predictors and category reference levels; negative coefficients decrease them. Coefficients are not causal effects, and categorical coefficients depend on the one-hot encoding and reference categories.

5. RESULTS

5.1 Classification performance across cutoffs

Table -1

Cutoff	Feature configuration	Accuracy	F1-score	ROC-AUC
Week 2	Activity only	0.7089	0.6868	0.7817
Week 2	Activity + assessments	0.7102	0.6881	0.7838
Week 4	Activity only	0.7246	0.7066	0.8035
Week 4	Activity + assessments	0.7493	0.7484	0.8311
Week 6	Activity only	0.7376	0.7209	0.8165
Week 6	Activity + assessments	0.7633	0.7645	0.8481
Week 8	Activity only	0.7480	0.7325	0.8292
Week 8	Activity + assessments	0.7789	0.7796	0.8654

Table 1 shows that all three classification metrics improved across the tested cutoffs for the activity-only configuration. Its ROC-AUC increased from 0.7817 at week 2 to 0.8292 at week 8. The activity-plus-assessment configuration reached 0.8654 ROC-AUC at week 8 and generally exceeded activity-only performance from week 4 onward. At week 8, adding assessment-related features increased accuracy from 0.7480 to 0.7789 and F1 from 0.7325 to 0.7796. The week-2 difference was small, so the apparent benefit of assessment features was more substantial at later cutoffs. These comparisons remain subject to the temporal-availability caveat described above.

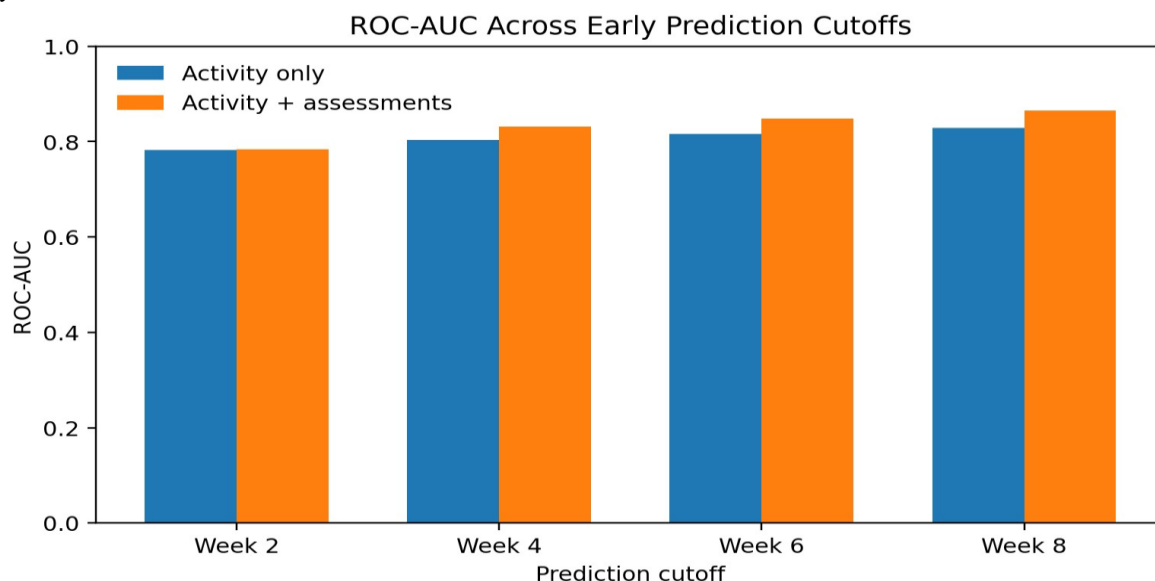


Figure -2: ROC-AUC by prediction cutoff and feature configuration

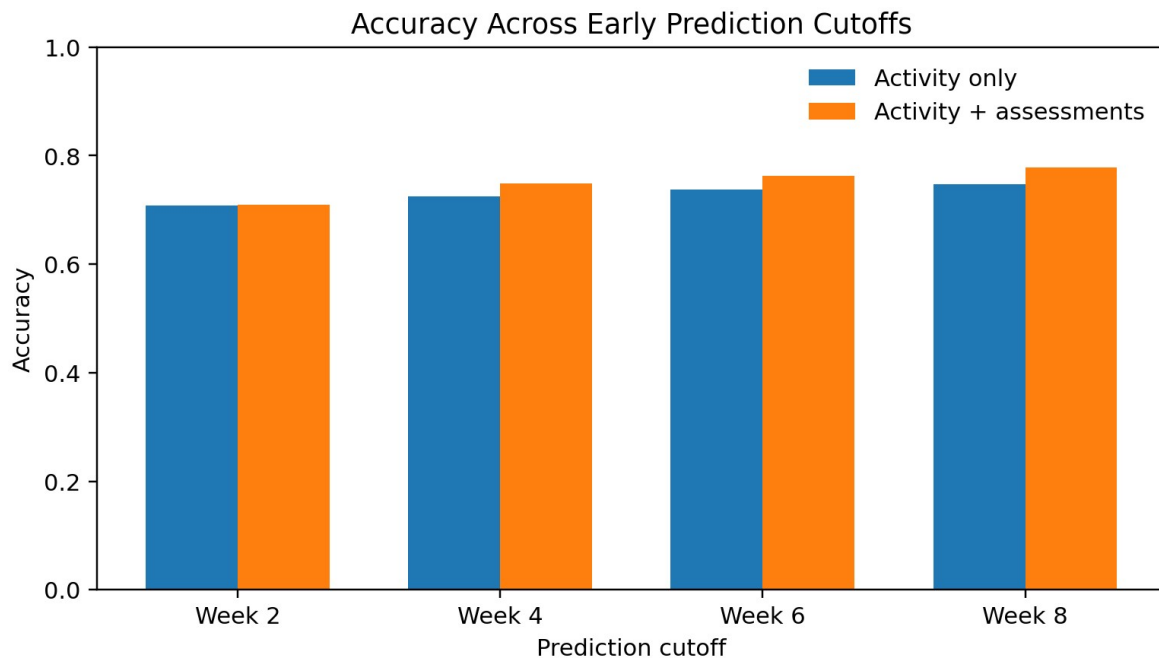


Figure -3: Accuracy by prediction cutoff and feature configuration

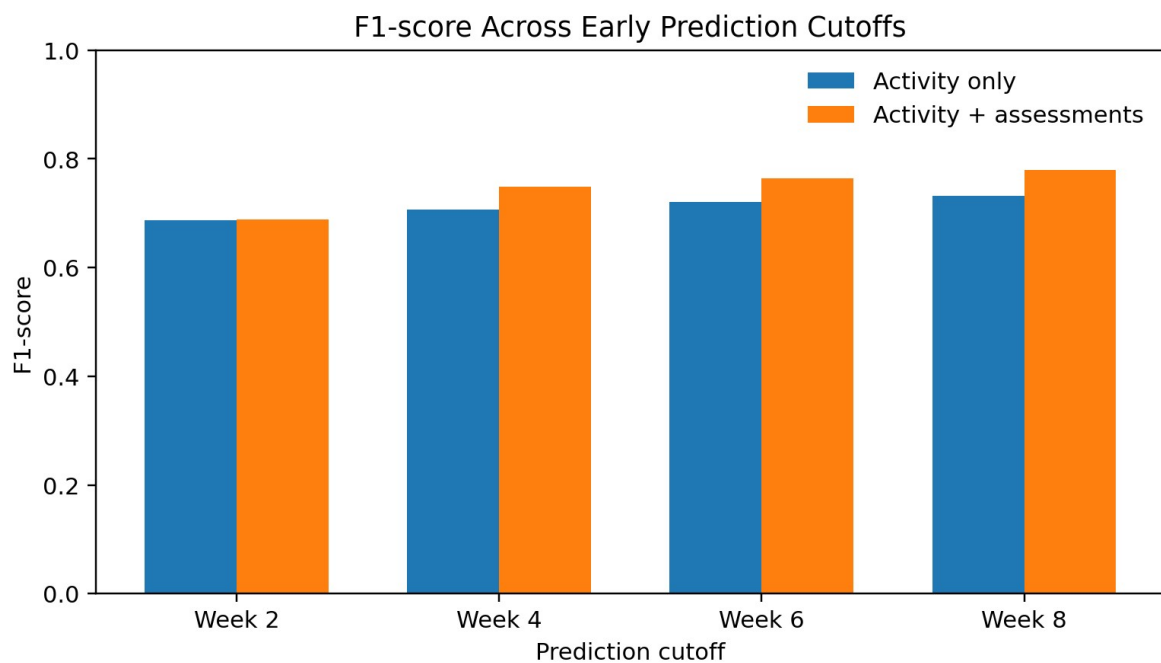


Figure -4: F1-score by prediction cutoff and feature configuration

5.2 Probabilistic performance

Table -2: Probabilistic performance across prediction cutoffs

Cutoff	Feature configuration	Brier score	Log loss	ROC-AUC
Week 2	Activity only	0.1907	0.5638	0.7817
Week 2	Activity + assessments	0.1899	0.5622	0.7838
Week 4	Activity only	0.1810	0.5391	0.8035
Week 4	Activity + assessments	0.1653	0.4932	0.8311
Week 6	Activity only	0.1750	0.5243	0.8165
Week 6	Activity + assessments	0.1562	0.4660	0.8481
Week 8	Activity only	0.1688	0.5086	0.8292
Week 8	Activity + assessments	0.1463	0.4394	0.8654

Table 2 reports the Brier score and log loss for each configuration and cutoff. Both metrics generally decreased with later cutoffs. The activity-plus-assessment model achieved a week-8 Brier score of 0.1463 and log loss of 0.4394, compared with 0.1688 and 0.5086 for activity-only features. At week 2, the difference was small. Lower Brier score and log loss indicate better probabilistic predictions on the evaluated test split, but do not by themselves establish that probabilities are calibrated for every subgroup or in a new institution.

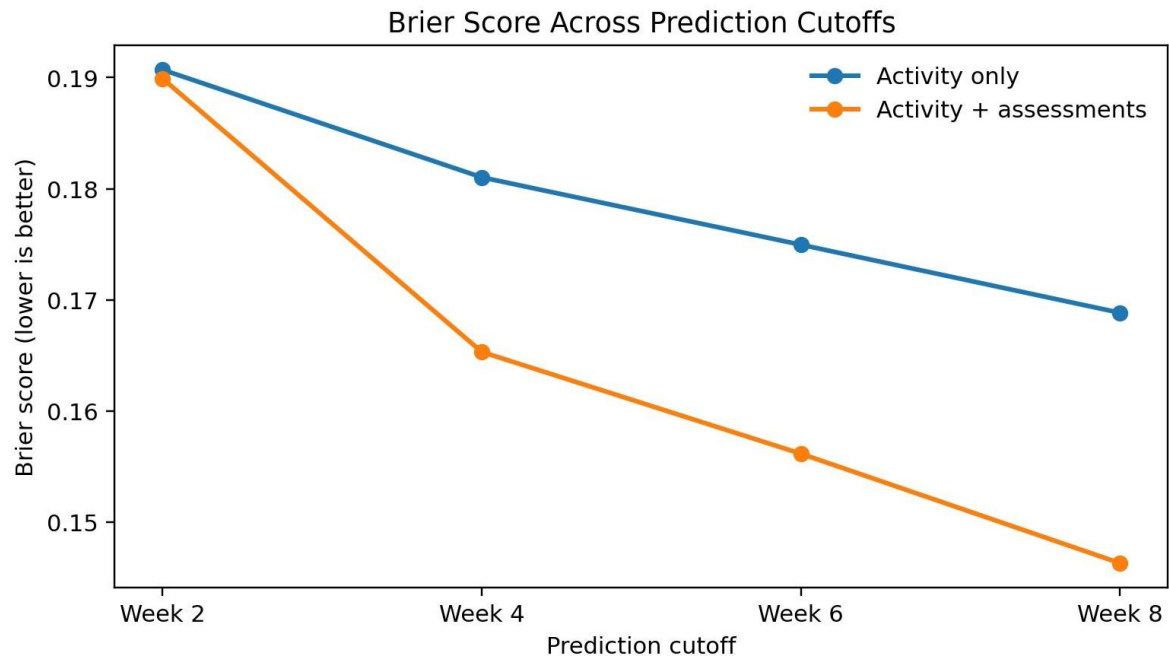


Figure -5: Brier score by cutoff; lower values indicate lower squared probability error

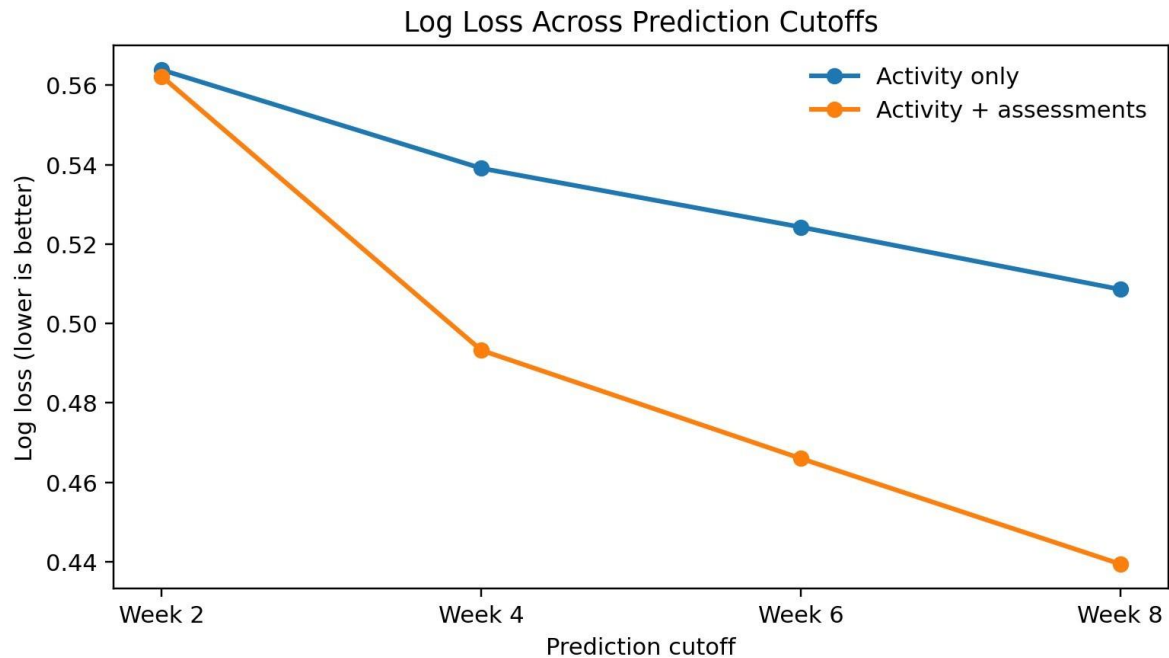


Figure -6: Log loss by cutoff; lower values indicate better probabilistic performance

5.3 Selective prediction and uncertainty

Table -3: Selective prediction at different confidence thresholds

Feature configuration	Confidence threshold	Coverage	Selected test cases	Accuracy among retained cases
Activity only	0.55	90.6%	5860	76.9%
Activity only	0.60	81.3%	5258	79.4%
Activity only	0.65	72.0%	4660	81.8%
Activity only	0.70	62.6%	4053	83.9%
Activity only	0.80	42.0%	2715	88.8%
Activity only	0.90	18.8%	1216	93.8%
Activity + assessments	0.55	90.8%	5875	80.5%
Activity + assessments	0.60	82.3%	5323	82.8%
Activity + assessments	0.65	73.3%	4741	85.6%
Activity + assessments	0.70	64.9%	4200	88.1%
Activity + assessments	0.80	46.9%	3036	92.9%
Activity + assessments	0.90	29.7%	1921	97.1%

Table 3 shows a consistent trade-off: stricter confidence thresholds retain fewer predictions but increase selective accuracy. For the week-8 activity-plus-assessment model, a threshold of 0.55 retained 90.8% of test cases with 80.5% accuracy. At a threshold

of 0.90, coverage fell to 29.7% (1,921 cases) while accuracy among retained cases rose to 97.1%. The activity-only model at the same threshold retained 18.8% of cases with 93.8% accuracy. These figures indicate that confidence filtering can identify a smaller subset of predictions with higher observed accuracy, but leave many students without a prediction; high confidence is not a guarantee of correctness.

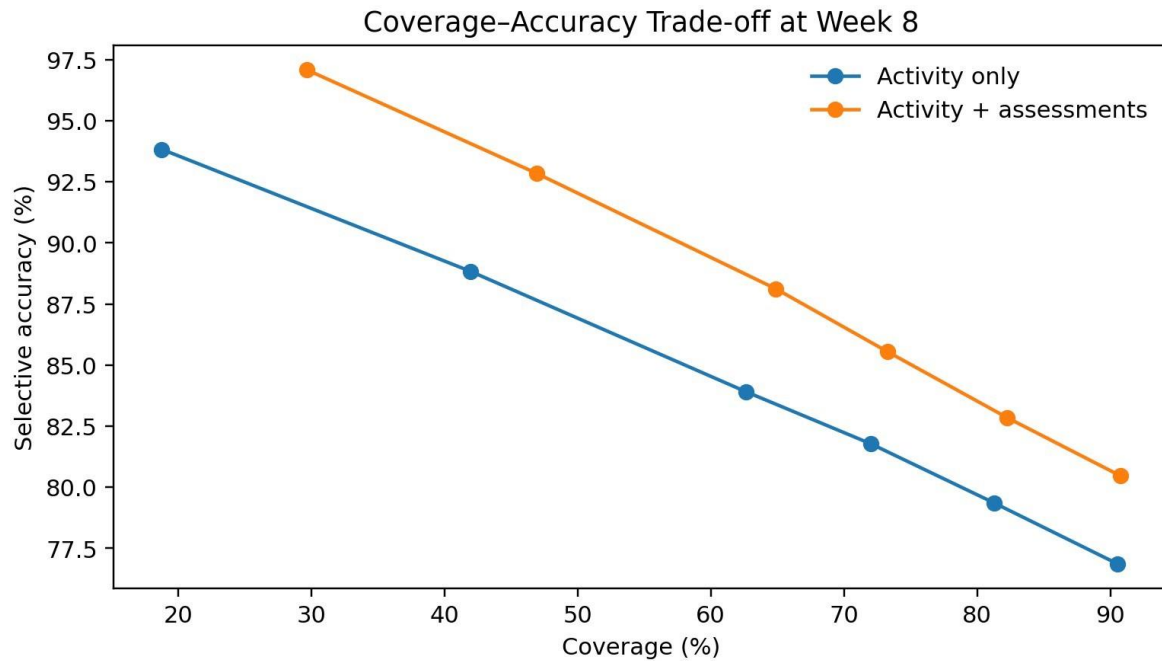


Figure -7: Coverage–accuracy trade-off at week 8. Coverage is the proportion of test cases retained at each confidence threshold

5.4 Robustness across random seeds

Table -4: Robustness across random seeds (mean ± SD)

Cutoff	Feature configuration	Accuracy mean ± SD	F1 mean ± SD	ROC-AUC mean ± SD
Week 2	Activity + assessments	0.7028 ± 0.0078	0.6809 ± 0.0074	0.7767 ± 0.0062
Week 2	Activity only	0.7028 ± 0.0062	0.6808 ± 0.0061	0.7749 ± 0.0060
Week 4	Activity + assessments	0.7426 ± 0.0059	0.7430 ± 0.0050	0.8260 ± 0.0044
Week 4	Activity only	0.7202 ± 0.0046	0.7033 ± 0.0031	0.7968 ± 0.0057
Week 6	Activity + assessments	0.7559 ± 0.0065	0.7578 ± 0.0059	0.8428 ± 0.0047
Week 6	Activity only	0.7336 ± 0.0037	0.7172 ± 0.0033	0.8121 ± 0.0038
Week 8	Activity + assessments	0.7752 ± 0.0032	0.7766 ± 0.0027	0.8629 ± 0.0029
Week 8	Activity only	0.7451 ± 0.0032	0.7302 ± 0.0021	0.8257 ± 0.0030

Table 4 summarizes performance across three random seeds. At week 8, activity-plus-assessment achieved mean ROC-AUC 0.8629 (SD 0.0029), compared with 0.8257 (SD 0.0030) for activity-only. The small standard deviations suggest that these results were relatively stable across the three tested splits. At week 2, the mean ROC-AUC difference was much smaller (0.7767 versus 0.7749). This supports the pattern that the assessment feature set was more beneficial at later cutoffs, subject to validation of feature timing.

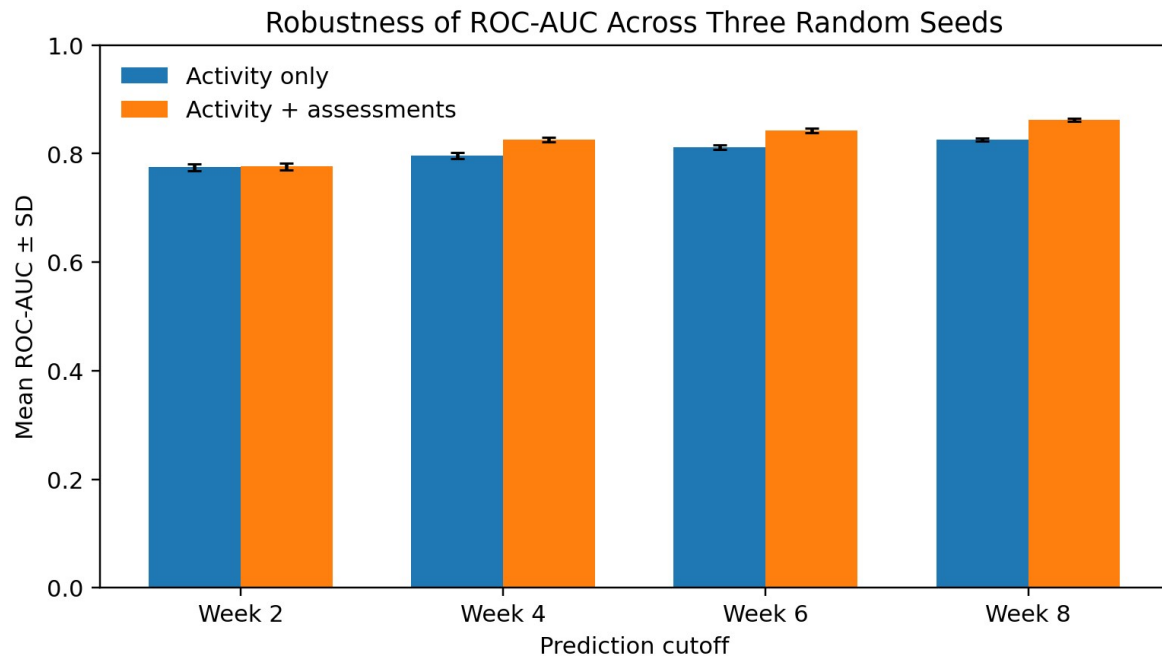


Figure -8: Mean ROC-AUC and standard deviation across random seeds 21, 42, and 84

5.5 Coefficient-based explainability

Table -5: Week-8 Logistic Regression coefficients

Feature	Week-8 coefficient
code_module_GGG	4.3248
code_module_DDD	-1.5191
code_module_FFF	-1.3144
code_module_CCC	-1.1289
missingindicator_week8_mean_score	-0.8107
week8_active_days	0.6854
week8_has_assessment	0.6691
week8_assessment_count	0.6247
week8_mean_score	0.5637
imd_band_?	0.4292
code_presentation_2013B	-0.4048
highest_education_No Formal quals	-0.3784

The largest absolute coefficient in the exported table was for code_module_GGG (+4.3248), followed by code_module_DDD (−1.5191), code_module_FFF (−1.3144), and code_module_CCC (−1.1289). Week-8 active days (+0.6854), presence of assessment information (+0.6691), assessment count (+0.6247), and mean assessment score (+0.5637) also had positive coefficients. The missing-score indicator had a negative coefficient (−0.8107). These values are conditional on the fitted model, preprocessing, target coding, and reference categories. In particular, the large module coefficient should not be interpreted as an inherent causal effect of a course; it may reflect differences in course structure, assessment, and student populations.

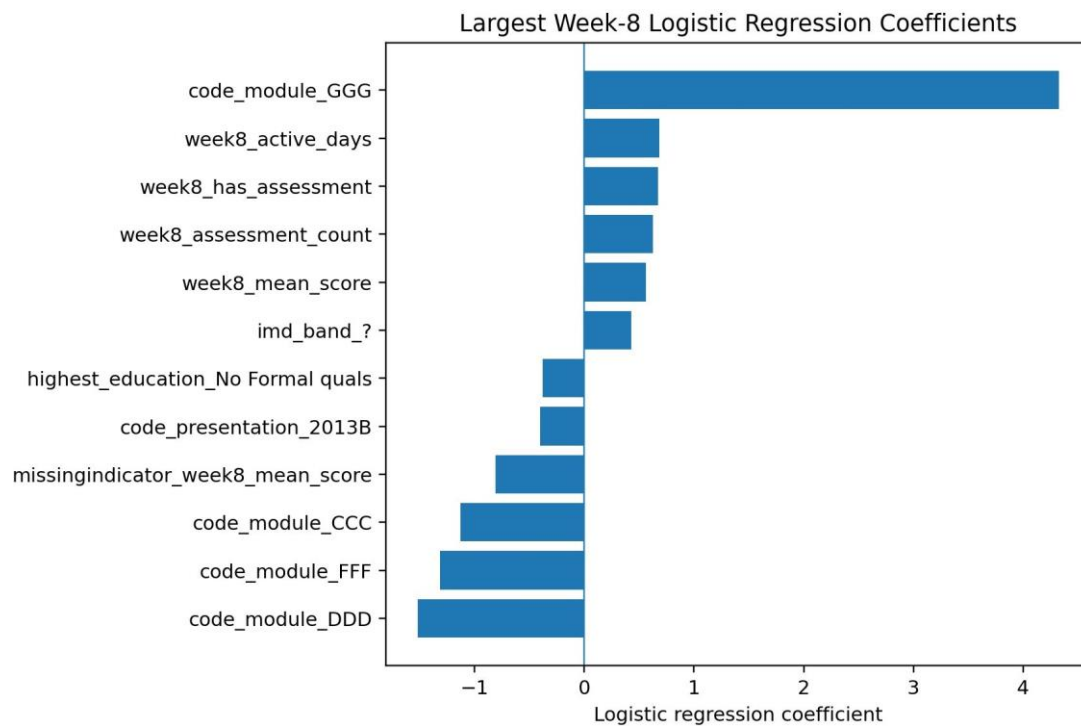


Figure -9: Twelve features with the largest absolute coefficients in the exported week-8 Logistic Regression coefficient table

6. DISCUSSION

The experiments indicate that predictive discrimination improved as the prediction cutoff moved from week 2 to week 8. This is expected in principle because later cutoffs provide more observed engagement and assessment information, although the precise pattern is empirical and dataset-specific. In the activity-only configuration, ROC-AUC rose from 0.7817 to 0.8292. The activity-plus-assessment configuration achieved higher scores at weeks 4, 6, and 8, with week-8 ROC-AUC of 0.8654 on the primary split.

The results also show that evaluating only classification metrics would be incomplete. At week 8, the activity-plus-assessment model had lower Brier score and log loss than the activity-only model, indicating better probability error on the test split. However, these metrics are not a complete proof of calibration across subgroups or future populations. The calibration curves and tabulated scores should be interpreted together.

Selective prediction provides an operational view of uncertainty. Increasing the confidence threshold increased accuracy among retained cases, but reduced coverage. For example, the week-8 assessment model reached 97.1% selective accuracy at a 0.90 threshold while covering only 29.7% of cases. In an educational support setting, this trade-off matters: withholding predictions for uncertain students may avoid overconfident labels, but the students left unclassified still require an appropriate support pathway.

The robustness results were consistent across three random seeds, with small standard deviations for the reported metrics. Nevertheless, this is limited split-based robustness, not independent replication. Finally, coefficient inspection offers a transparent summary of associations but cannot identify causes. The high magnitude of some course-module coefficients reinforces the need for contextual interpretation and fairness checks.

7. LIMITATIONS AND THREATS TO VALIDITY

- Temporal leakage risk: assessment dates and submission dates do not prove when grades became available. Assessment-score features must be checked against actual grade-release timing before the assessment-augmented results can support a strict early-prediction claim.
- Dataset generalizability: OULAD reflects a particular institution and learning context. Results may not transfer to other universities, courses, or student populations.
- Binary target: Pass and Distinction are combined, as are Fail and Withdrawn, which obscures differences among four original outcome categories.
- Model scope: Logistic Regression is interpretable but may not capture nonlinear relationships or interactions without explicit feature construction.
- Calibration and uncertainty: confidence-based filtering does not guarantee calibrated probabilities. Coverage can become low at high thresholds, limiting who receives a prediction.
- Robustness scope: three random seeds assess sensitivity to split choice but do not replace external validation, temporal validation, or subgroup testing.
- Fairness and ethical use: demographic and socioeconomic proxy variables may encode structural differences. Any intervention system should include fairness assessment, privacy protection, human review, and safeguards against stigmatizing students.

8. CONCLUSION

This study evaluated Logistic Regression for early prediction of binary academic outcomes using OULAD, comparing activity-only features with a configuration that also included assessment-related information. On the primary split, the highest reported classification performance was obtained at week 8 with activity-plus-assessment features (accuracy 0.7789, F1 0.7796, ROC-AUC 0.8654). Its Brier score was 0.1463 and log loss 0.4394. Selective prediction showed that higher confidence thresholds can improve accuracy among retained cases while substantially reducing coverage. Results across three random seeds were comparatively stable.

The main conclusion is conditional: the experiments suggest that later-semester information and assessment-related features can improve prediction, but the assessment-feature timing must be verified to rule out temporal leakage. The work should therefore be treated as a promising experimental workflow rather than a deployment-ready early-warning system. Future work should verify score-availability timestamps, validate on independent data, evaluate subgroup fairness, and assess whether predictions lead to beneficial support outcomes.

9. DATA AND REPRODUCIBILITY STATEMENT

The study uses the publicly available Open University Learning Analytics Dataset (OULAD). The project workflow used scikit-learn pipelines and exported model-comparison, calibration, uncertainty, robustness, and coefficient tables. The reported numerical values in this manuscript were taken from the uploaded research-results package. For full reproducibility, the final submission should archive the exact notebook, package versions, data-processing scripts, and final CSV outputs.

All authors have read and agreed to the published version of the manuscript.

10. REFERENCES

- [1] Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open University Learning Analytics dataset. Scientific Data, 4, 170171. <https://doi.org/10.1038/sdata.2017.171>
- [2] Open University Learning Analytics Dataset (OULAD). Dataset portal. https://analyse.kmi.open.ac.uk/open_dataset
- [3] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830. <https://jmlr.org/papers/v12/pedregosa11a.html>
- [4] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. Proceedings of the 22nd International Conference on Machine Learning. <https://doi.org/10.1145/1102351.1102430>

APPENDIX A. FULL WEEK-8 SELECTIVE-PREDICTION TABLE

Table -6: Full week-8 selective-prediction results

Feature configuration	Threshold	Coverage	Selected cases	Selective accuracy
Activity only	0.55	0.9056	5860	0.7686
Activity only	0.60	0.8125	5258	0.7936
Activity only	0.65	0.7201	4660	0.8178
Activity only	0.70	0.6263	4053	0.8391
Activity only	0.80	0.4196	2715	0.8884
Activity only	0.90	0.1879	1216	0.9383
Activity + assessments	0.55	0.9079	5875	0.8046
Activity + assessments	0.60	0.8226	5323	0.8285
Activity + assessments	0.65	0.7327	4741	0.8555
Activity + assessments	0.70	0.6490	4200	0.8812
Activity + assessments	0.80	0.4692	3036	0.9285
Activity + assessments	0.90	0.2969	1921	0.9708

APPENDIX B. INTERPRETATION NOTES

A positive Logistic Regression coefficient corresponds to increased log-odds of the encoded positive class, conditional on other variables and the model's coding scheme. It does not imply that changing the feature would cause an outcome change. The reference categories and full feature preprocessing should be preserved with the model if coefficients are interpreted in future work.

All reported classification and probability metrics are test-set results from the supplied experiment tables. Robustness statistics are means and standard deviations across the three specified random seeds. The paper deliberately preserves the temporal-leakage caveat associated with assessment-score features.