A Data-Analytical Framework for the Early Detection of At-Risk Students in Higher Education

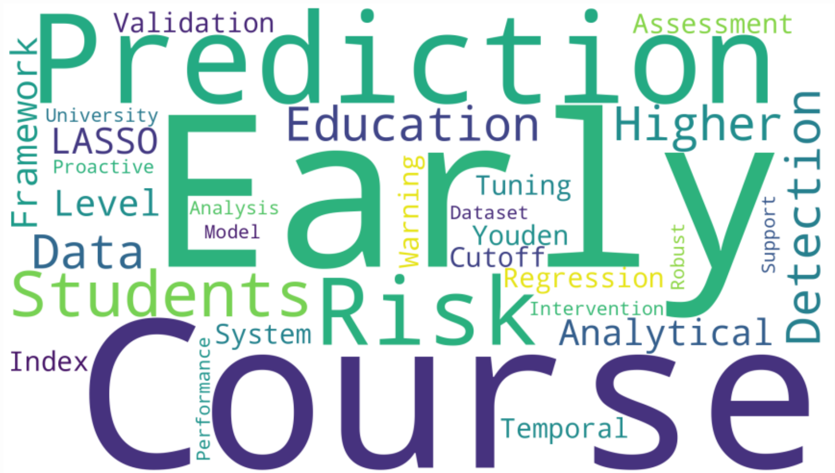
Course-Level Model

2025 – March – 15

DONG Chenxi
Department of Mathematics and Information Technology
The Education University of Hong Kong

Content

- Introduction
  - Motivation
  - Limitations of Existing Research
- Objectives
- Methodology
  - Overview
  - Dataset
  - Model Mechanism
- Model Performance
- Discussions
- Conclusion & Future Directions



Introduction

1. Motivation
2. Limitations of Existing Research
3. Our Study
4. Impact

Introduction

- **Motivation:** Enable early detection for at-risk students in university courses.
- **Limitations of Existing Research:**
  - **Limited Practical Application:** Non-temporal validation approaches (train-test in the same cohort); models not validated on future data.
  - **Model Overfitting Risk:** High due to non-temporal validation; unrealistic performance.
  - **Interpretability Gap:** Difficulty identifying key factors in high-dimensional data.
  - **Fragmented Insights:** Separate grade prediction (Regression) and at-risk detection (Classification); lacks integrated actionable insight.
- **Our Study:** A temporally validated predictive framework that delivers accurate course grades by mid-semester
- **Impact :** Early At-Risk Student Detection -> Intervention on At-Risk Students -> Enhanced Student Outcomes

Objectives

1. Develop a Mid-Semester Early Warning System for the Course
2. Create a Dual-Output Prediction Framework
3. Validate through Temporal Cross-Cohort Testing
4. Interventions to Support At-Risk Students

Objectives

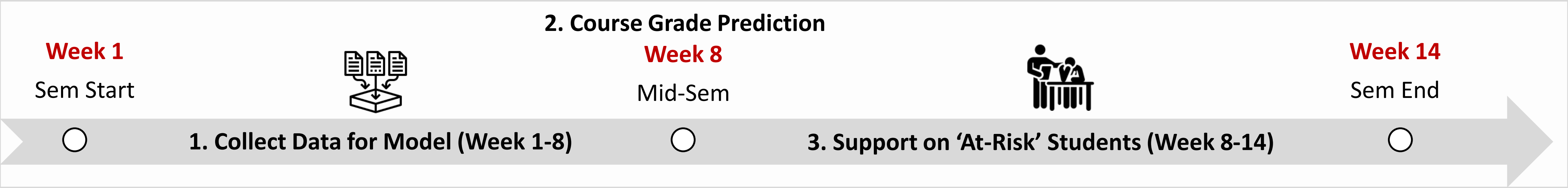


Fig 1. Early (Week 8) At-Risk Students Detection in a University Course

1. **Develop a Mid-Semester Early Warning System for Course**
   - Timing: Week 8 of the 14-week semester (when intervention is still effective)
   - Definition: At-risk = students projected to receive grades below 2.33/4.33
2. **Create a Dual-Output Prediction Framework**
   - Continuous variable prediction: Course grade point estimates (e.g., 2.65/4.33)
   - Binary classification: At-risk status (0 or 1) with optimized at-risk detection cutoff
3. **Validate through Temporal Cross-Cohort Testing**
   - Train on the previous cohort data (historical data, i.e, 2021)
   - Test on the current cohort data (current data, i.e., first 8 weeks in 2022)
4. **Interventions to Support At-Risk Students**

Methodology

1. Overview
2. Dataset
3. Model Mechanism

Methodology: Overview

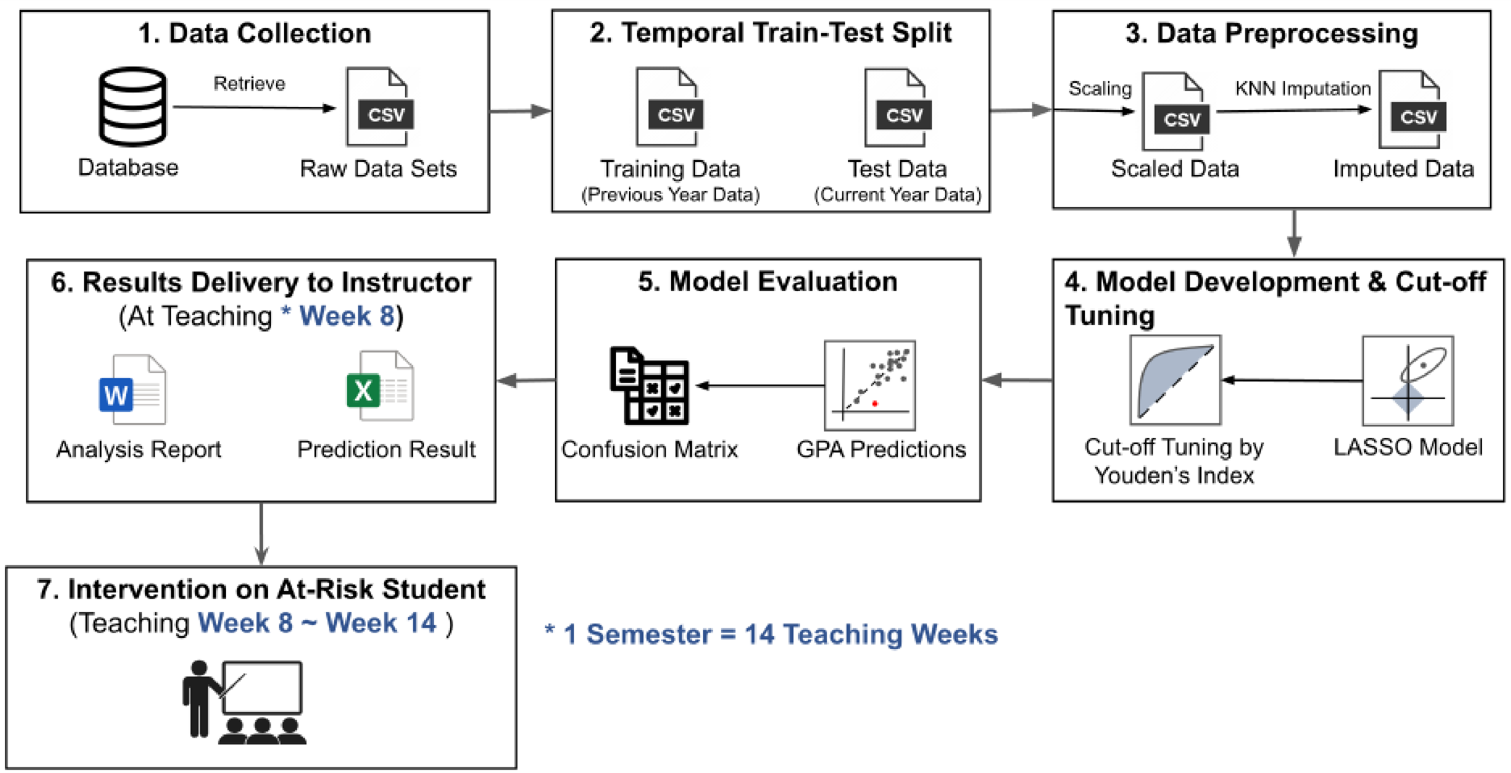


Fig 2. Methodology Overview for Early At-Risk Students Detection

1. Data Collection
   - Retrieve from university database: Academic, demographic, & engagement data.
   - Gather in-course assessment data from the course instructor.
2. Temporal Train–Test Split
   - Historical data (e.g., 2021) for training; Current data (e.g., 2022) for testing.
   - Simulates real-world early prediction for future cohorts.
3. Data Processing
   - Apply scaling and KNN imputation to handle missing values.
4. Model Development & Cut-off Tuning
   - Train LASSO regression for GPA prediction.
   - Tune optimal at-risk GPA cutoff using the Youden Index.
5. Model Evaluation
   - Assess course grade point prediction (R^2, MSE) & classification (confusion matrix).
6. Results Delivery to Instructor
   - Provide analysis reports and at-risk predictions by Week 8.
   - Enables proactive support.
7. Intervention on At-Risk Students
   - Offer assistance (extra tutorials, consultations) before the course ends (Week 14).

Methodology: Dataset

Case Study Context

- **Course:** An Undergraduate Course
- **Prediction target:** Final course grade point (0-4.33 scale; At-risk cutoff: <2.33)

Dataset Feature Categories

- **Prior Academic Performance:** Term GPA, Cumulative GPA from university records
- **Demographic Factors**: Gender, residency status, entry path,…
- **Engagement Metrics:** LMS activity patterns, scholarships
- **In-Course Assessment:** Mid-term quiz scores (Week 7) from course instructor

Temporal Train-Test Design

- Training: 2021 cohort (N=60) → Testing: 2022 cohort (N=30)
- Simulates real-world implementation conditions

Table 1. List of features in course-level dataset (training/testing)

Feature Name	Description	Data Type
SID	We masked student IDs to protect students' privacy.	Categorical
Term	The course term code representing the semester is formatted as YYYYMM	Categorical
TGPA.Prev	Student's previous term GPA ranges from 0 to 4.33, indicating past performance.	Numerical: Continuous
CGPA.Prev	Student's cumulative GPA before the current term ranges from 0 to 4.33, measuring overall achievement.	Numerical: Continuous
Gender	Categorizes students as Male or Female.	Categorical
Local	1 = Local students; 0 = Non-local students	Categorical
SenYrEntr	Senior year entry students' status (1 if student joined in 3rd year after diploma, 0 otherwise).	Categorical
Hostel	1 for on-campus, 0 for off-campus residence.	Categorical
Scholarship	The total number of scholarships received before this course, indicating past recognition.	Numerical: Discrete
Mdl_ts	Normalized cumulative time on Moodle for this course in 10 quantiles ranges from 1 to 10, measuring engagement.	Numerical: Discrete
Midterm Quiz	A midterm quiz for the course ranges from 0 to 100	Numerical: Continuous
Course grade point	*The Target variable for prediction ranges from 0 to 4.33*	*Numerical: Continuous*

Methodology: Model Mechanism

- **LASSO Regression** (Tibshirani, 1996) for Course Grade Point Prediction:
  - $y_{pred} = \widehat{w}X$, where $\widehat{w} = Argmin_w\ (RSS(w) + \lambda|w|)$
  - Strength: Automatic predictor selection, minimizes overfitting, interpretable.
- **Youden Index (J)** (Youden, 1950) for Cutoff Tuning:
  - Challenge: Default cutoff (c) misses borderline at-risk students
  - Solution: Data-driven cutoff optimization by Youden Index
  - $\mathrm{J} = Sensitivity\ (c) + Specificity\ (c) - 1$, Larger the better.
  - Finding the Optimal Cutoff: $c^* = Argmax_c\ J(c)$ (See Fig 3)
- **At-Risk Detection** - Binary Classification based on Tuned Cutoff c^*
  - $y_{at-risk} = 1\ if\ y_{pred} < c^* otherwise\ 0$ (See Fig 4)
- **Evaluation Metrics:**
  - Regression: R², Mean Squared Error (MSE)
  - At-Risk Classification - Confusion matrix
    - Missed detections (false negatives) - most critical error type
    - False alarms (false positives) - acceptable when minimizing missed cases

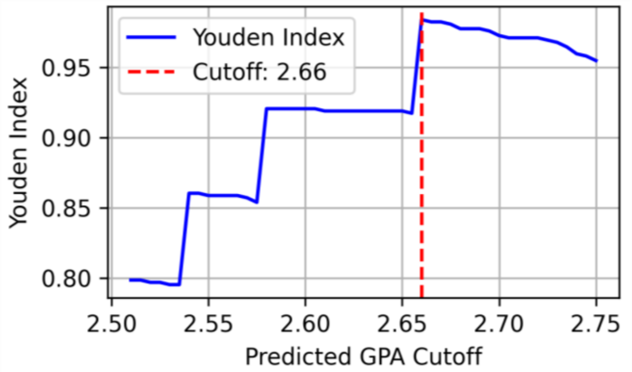


Fig 3. Using the Youden Index to Tune the Optimal Cut-Off

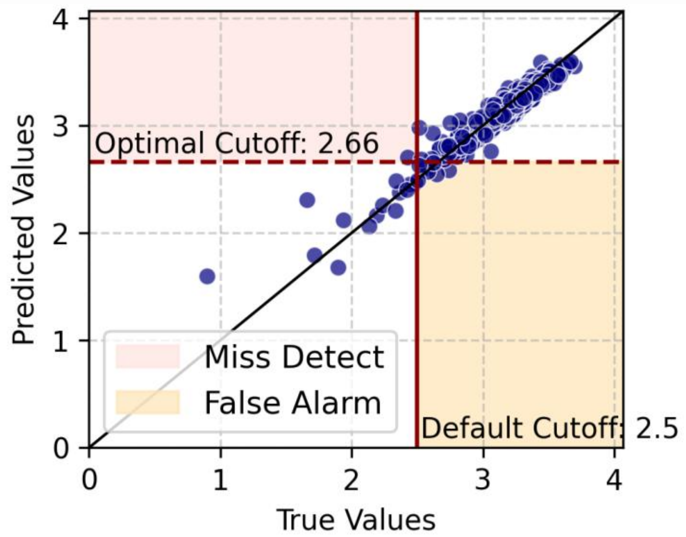


Fig 4. Model Performance using Tunned Cutoff – Actual vs. Predicted Score

Model Performance

1. Training and Testing Set Performance
2. Discussion

Model Performance

Table 2. Model Performance for a UG Course

Dataset	N	R^2	MSE	Accuracy	No. At-Risks	Miss Detects	False Alarms
Training Set (Year 2021)	60	0.72	0.19	83.3%	3	1	9
Testing Set (Year 2022)	30	0.63	0.21	80.0%	2	0	6

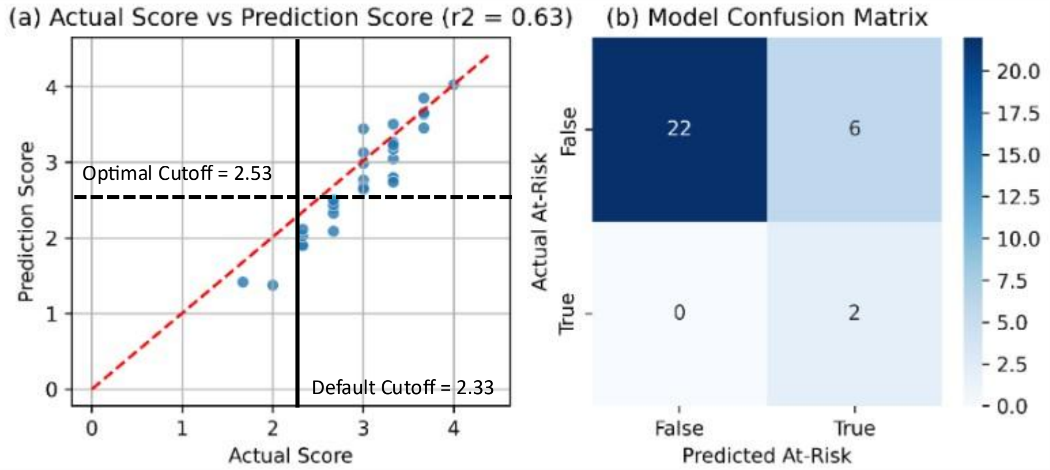


Fig 5. Model performance on the test set (a): actual-predicted score scatter plot; (b): confusion matrix for at-risk students' classification - tunned cutoff 2.53

Key Observations:

- Maintained performance despite small sample size (N=30)
- Strong out-of-sample prediction on future, unseen cohort
  - 0 missed detections in the test cohort
  - Acceptable false positive rate (6 students)
- Temporal validation confirmed real-world applicability
- Minimal performance degradation between training and testing (Δ MSE = 0.02)

Discussion: Importance of Course Assessment in Course Grade Point Prediction

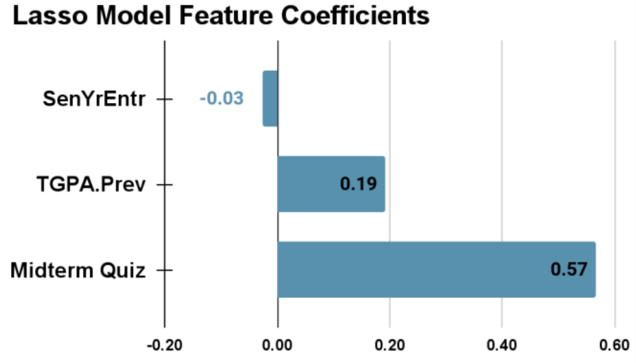


Fig 6. Lasso model coefficients - Predictor Significance

Table 3. Test set model performance (With/Without Course Assessment)

Dataset Include Course Assessment	N	R^2	Miss Detects (Total 2 At-Risks)
Yes	30	0.63	0
No	30	0.03	2

Key Observations:

1. **Course assessments are essential for accurate grade point prediction:**
   - Most impactful predictor (Fig 6).
   - In-course assessments before week 8 significantly enhance prediction quality (See Table 3)
2. **High-quality assessments are crucial:**
   - Discrimination power needed to differentiate students.
   - Varied difficulty levels (Easy, Medium, Difficult) ideal.
3. **Practical recommendations for course instructors**
   - Implement substantial assessments before Mid-Sem
   - Ensure rapid grading to enable Mid-Sem prediction

Discussion: Intervention Strategies & Instructor Feedback

- **Instructor-Led Interventions**
  - Course instructors decided and implemented interventions (See Table 4) to support at-risk students.
- **Positive Instructor Feedback**
  - Positive Impact: The instructor reported, *"To a large extent, the intervention can help the at-risk students.“*
  - Student Engagement: The instructor reported, *"Some at-risk students asked questions about course exercises by email and attended a supplementary class on course projects.“*

Table 4. Interventions Implemented by Course Instructors

Intervention Strategy	Course 1	Course 2	Course 3	Course 4
1. Individual academic advising		Yes		
2. Provide extra learning materials			Yes	
3. Arrange peer tutors/TA for consultations	Yes	Yes	Yes	
4. Organize supplementary classes				Yes

- **Quantitative Evidence**: Significant Grade Lift after Intervention
  - Average 0.72 Point Lift: Statistical analysis shows actual course grades were significantly higher than predicted grades ($p < 0.05$), with an average grade improvement of 0.72 points across all students.
  - This indicates a positive impact on overall student performance after intervention.

Conclusion

1. Key Contributions
2. Limitations & Future Directions

Conclusion & Future Directions

Key Contributions:

- **Practical Framework:** We developed a LASSO & Youden Index model for early at-risk student detection.
- **Dual Prediction Output:** The model provides both Course Grade (Regression) & At-Risk Label (Classification).
- **Robust Performance:** Demonstrated strong out-of-sample prediction accuracy on cross-cohorts student data, showcasing real-world applicability.
- **Positive Impact for Intervention**
- **Model Explanation Provided:** Lasso model feature coefficients

Limitations & Future Directions:

- **Small data set**
  - Future Direction: merge dataset, course-agnostic model
  - Possible solution: synthetic data augmentation
- **Excessive manual work required in the data pipeline**
  - Manual collection of assessment score data from course instructors
  - Changes in course structure and instructors over time necessitate manual data cleaning and processing
  - Data pipeline automation opportunities
- **Systematic evaluation of intervention effectiveness:** Which intervention strategy is the most effective?

Thank You

This work was supported by the Central Reserve Allocation Committee and the University-level Teaching Development Grant, The Education University of Hong Kong (Project No. 03ABL and T0282).

香港教育大學
The Education University of Hong Kong

This presentation is based on the work published in:

Dong, C., Yip, J. C., Ling, A. M. H., Kwan, J. L. Y., Yu, P. L. H., Cheng, M. H. M., Lee, J. C., & Li, W. K. (2025). ***A Data-Analytical Framework for the Early Detection of At-Risk Students in Higher Education***. In M. Della Ventura & Z. Zhan (Eds.), *Exploring Innovations in Educational Technology: The ICEIT'25 Collection* (Chapters 1–2). Springer Nature. https://link.springer.com/book/9789819508716