A DATA-ANALYTIC APPROACH TO EARLY DETECTION OF AT-RISK STUDENTS IN HIGHER EDUCATION

19 April 2025

CSTE Wuhan

Dr. Man Ho LING

Department of Mathematics and Information Technology

The Education University of Hong Kong

ACKNOWLEDGEMENT

- 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).

- This presentation is based on the work published in:

  Yip, J. C., Dong, C., Ling, A. M. H., Kwan, J. L. Y., Yu, P. L. H., Cheng, M. M. H., ... & Li, W. K. (2025, April). A Data-Analytic Approach to Early Detection of At-Risk Students in Higher Education. In *2025 7th International Conference on Computer Science and Technologies in Education (CSTE)* (pp. 228-232). IEEE. doi: 10.1109/CSTE64638.2025.11092031

OBJECTIVES

- **Identification:** To help leaders of academic programs identify at-risk students relatively early in those programs.

- **Enhancement:** To identify impactful courses within each program.

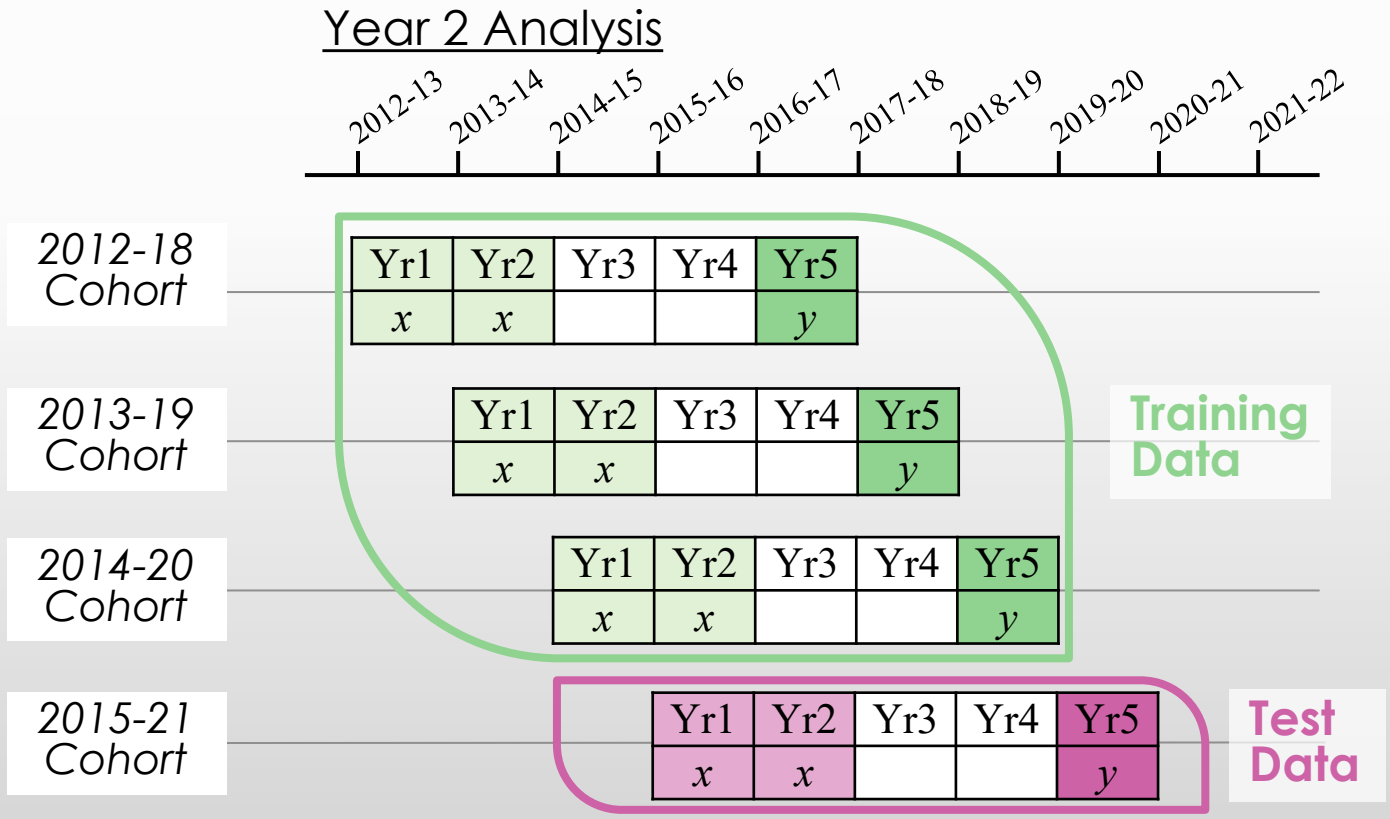
TYPES OF RECORDS

	Student Particulars	*Enrollment Info*	*Academic Records*	*Entrance Exams*	*Extracurricular Data*
Program-Level Data	Student ID (masked as Pseudo ID)	Program Code	Course Grades (all course taken)	Exam Type e.g. DSE, etc.	Hostel Status per Term
	Gender: F, M	Academic Term e.g. 201401	Program GPAs (end of each term)	Exam Subject e.g. CHN, EGB, MAB	On-Campus Job Hours per Year
	Residency: Local, Non-local	Program Year e.g. Year 2	Number of low grades	Exam Score 1, 4, 5*, etc.	Team/Org. Membership per Year
	Admission Route: JUPAS, Non-JUPAS, Mainland				Scholarships per Year

CROSS-VALIDATION USING PREVIOUS YEARS' DATA

Testing at Year 2 is early enough that program leaders can be notified to make necessary interventions for potentially at-risk students.

- LASSO models were trained on data from three consecutive cohorts and then tested on data from the next (fourth) cohort:



LASSO REGRESSION MODEL

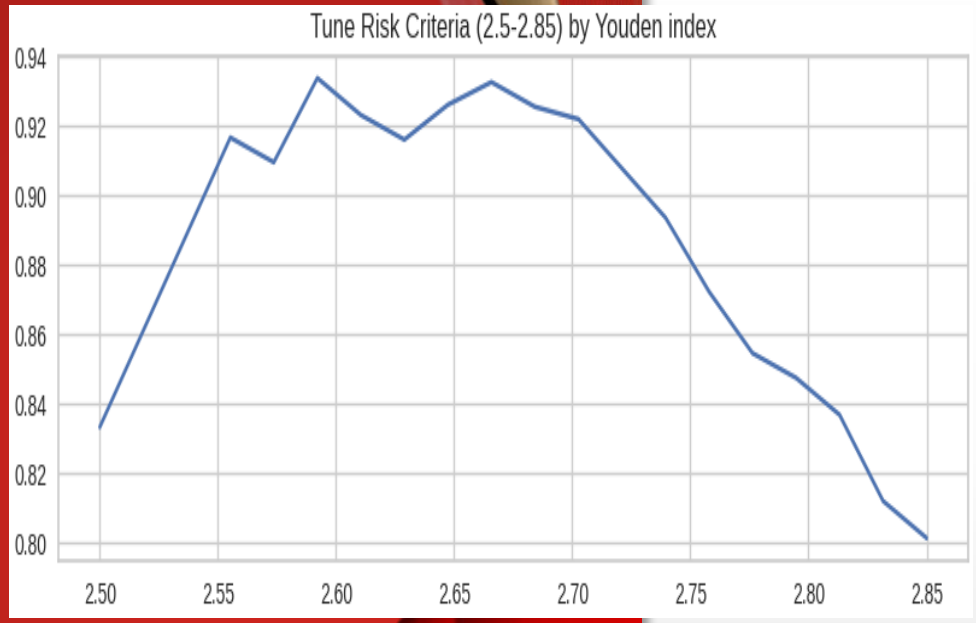
Training data were analysed using LASSO regression models, which provide our analysis with the following advantages

- **Efficiency:** LASSO models can be fit without stepwise fitting and testing
- **Parsimony and Interpretability:** LASSO model select the smallest set of predictors that explain the greatest amount of variation, and the relative strength of effects is explicitly indicated.
- **Validity:** LASSO modeling avoids overfitting of the data arising from a large number of predictors and multicollinearity issues due to predictors that are highly correlated.

DATA PRE-PROCESSING

- **Imputation and Standardisation:** Missing data values were filled in using *k*-NN imputation and then scaled using a standard scaling function.

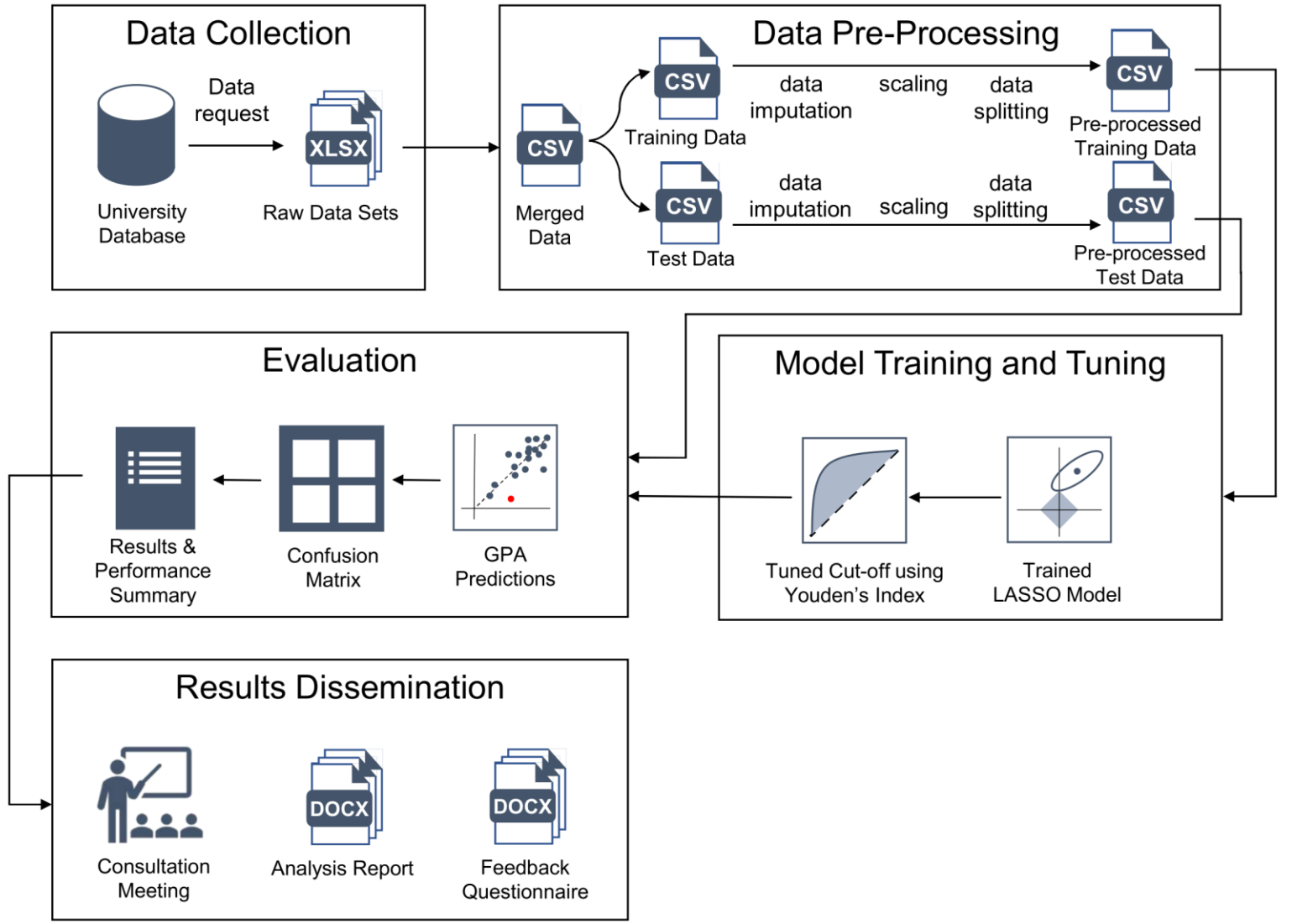
TUNING THE AT-RISK CUTOFF



At-risk classifications depended a cutoff that was tuned according to the accuracy of the model:

- The outcome variable was students' **cumulative GPA at graduation**, and the at-risk cutoff was based on a result of **2.50 or below** (Third-Class Honours).
- However, the actual cutoff point was adjusted by testing a range of possible cutoff points in the training data (predicted vs. actual outcome) and calculating their corresponding Youden's J scores.
- The **optimal cutoff** point was the outcome value with the highest Youden's index. This value was normally between 2.50 and 2.80.

OVERALL WORKFLOW

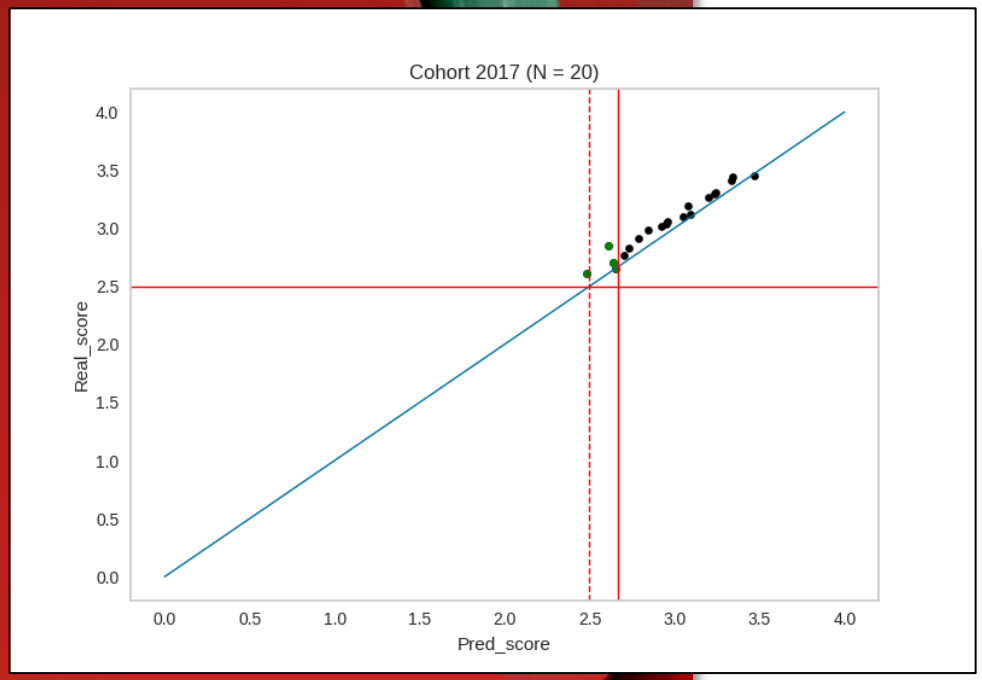


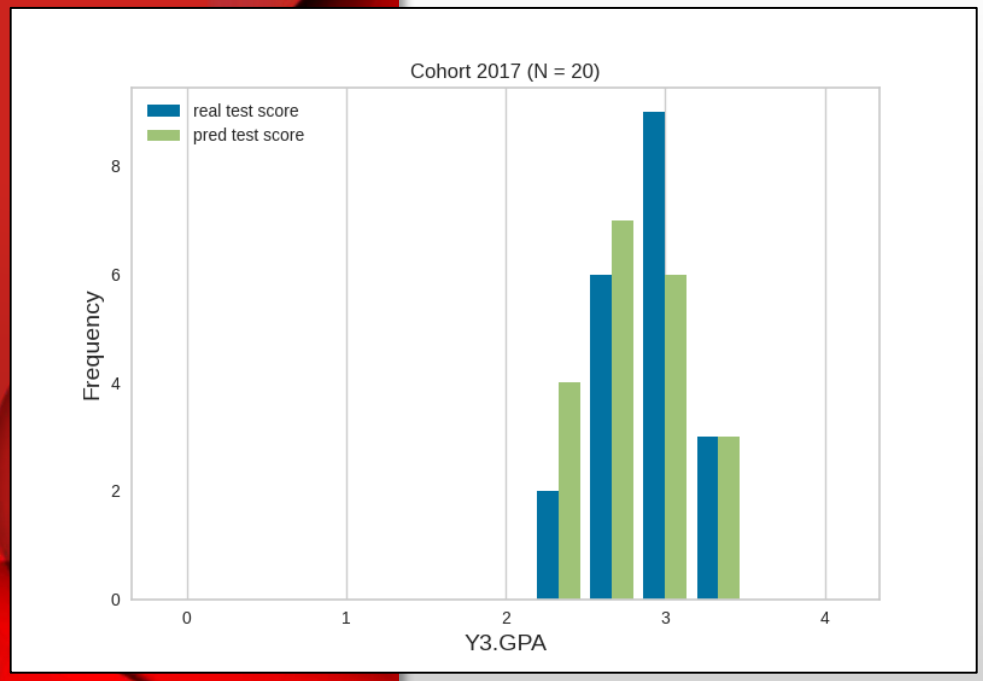
DEMONSTRATION OF METHODOLOGY WITH PROGRAM X

2017-22 COHORT OF PROGRAM X: [TRAINING & TESTING DATASETS]

Category	Variables	N	Mean	Std. Dev.	% Missing Data
Academics	Year 1 GPA (cumulative)	166	2.98	0.36	0.0%
	Year 2 GPA (cumulative)	166	2.99	0.37	0.0%
	Graduating GPA (cumulative)	166	3.05	0.3	0.0%
Entrance Exams	Sum of Best 5	113	20.21	1.06	31.9%
	Chinese Language subject	142	2.93	1.6	14.5%
	English Language subject	143	2.69	1.46	13.9%
	Liberal Studies subject	143	3.29	1.86	13.9%
	Mathematics subject	141	3.61	2.05	15.1%
	Mathematics optional module	63	3.21	1.36	62.0%
Scholarships	Scholarships (by Year 1)	166	0.12	0.39	0.0%
	Scholarships (by Year 2)	166	0.24	0.64	0.0%
Extracurriculars	Student Residence (in Year 1)	166	1.39	0.87	0.0%
	Student Residence (in Year 2)	166	1.28	0.93	0.0%
	On-Campus Job (hours in Year 1)	166	4.93	11.1	0.0%
	On-Campus Job (hours in Year 2)	166	43.36	99.67	0.0%
	Team/Organization in Year 1	166	0.17	0.38	0.0%
	Team/Organization in Year 2	166	0.35	0.58	0.0%

2017-22 COHORT OF PROGRAM X: PREDICTED VS. ACTUAL RESULTS



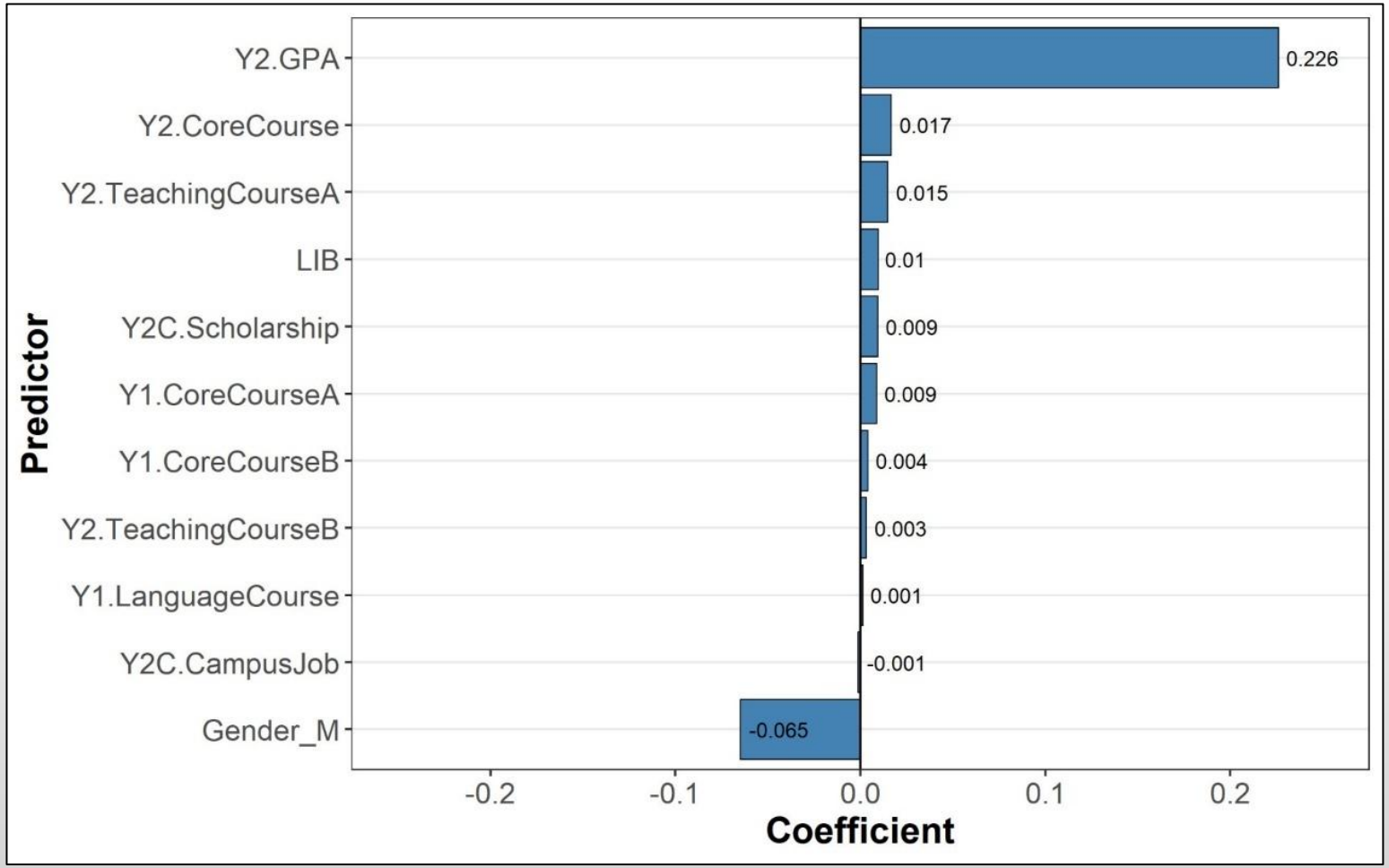


A comparison of predicted and actual graduating GPAs show a strong correlation between the two values.

- Correlation value = 0.9823
- r2 = 0.966
- MSE value = 0.00987

2017-22 COHORT OF PROGRAM X: SIGNIFICANT PREDICTORS

Feature importances for the 2017-22 model



RESULTS FROM YEAR 2 ANALYSIS OF PROGRAM X, 7 COHORTS

- The results over 7 years of testing indicates low rates of missed detections and slightly higher rates of false alarms.

Test Cohort	N	Correct Normal	Missed Detection	Correct At-Risk	False Alarm	Hit Rate	r^2-value
2015-20	56	48	**1**	6	1	0.964	0.942
2016-21	21	18	**1**	2	0	0.952	0.969
2017-22	20	16	**0**	0	4	0.800	0.966
2018-23	17	16	**0**	1	0	1.000	0.938
2019-24	77	75	**0**	2	0	1.000	0.921
2020-25	64	62	**1**	1	0	0.984	0.940
2021-26	71	66	**0**	3	2	0.972	0.950
Total	*326*	*301*	***3***	*12*	*7*	*n/a*	*n/a*
Mean	*46.6*	*43.0*	***0.4***	*1.7*	*1.0*	***0.953***	***0.947***

COMPARISON OF LASSO WITH OTHER MODELS OVER 3 COHORTS

The LASSO model outperforms other types of regression models over **3 cycles** of training and testing:

- Missed detections are minimized in the LASSO approach.
- Excess false alarms are tolerated to a reasonable degree.

Predictive Model	N	Correct Normal	Missed Detection	Correct At-Risk	False Alarm
Lasso regression	963	862	**11**	40	50
(variable selection)		89.5%	**1.1%**	4.2%	5.2%
Simple linear regression	963	872	14	37	**40**
(Year 2 GPA only)		90.6%	1.5%	3.8%	**4.2%**
Multiple linear regression	963	868	13	38	44
(all variables included)		90.1%	1.3%	3.9%	4.6%

REMARKS

- Given our approach, numbers of at-risk detections (true positives) are high, and numbers of missed detections (false negatives) are low.
- Year 2 analysis yields better results than Year 1 analysis.
- Impactful predictors (such as grades in required coursework) can be clearly identified from the LASSO model result.
- Overall, LASSO regression performs better than simple linear regression and multiple linear regression (no variable selection), with LASSO predictions being closer to actual outcome values than those of the other models.