IIAI Open Conference Publication Series
IIAI Letters on Institutional Research
Vol. 003, LIR089
DOI: https://doi.org/10.52731/lir.v003.089

Using Azure Machine Learning to Detection of At-Risk Students

Dun-Cheng Chang [*], Shinyi Lin [†]

Abstract

An important issue in higher education is the systematic monitoring of student achievement stability in programs; an Institutional Research (IR) department found that students were falling behind in grades, so IR would provide an academic counseling mechanism. This study uses Azure Machine Learning (AML) cloud to establish an early alert system that enables teachers to provide tutorial support to students struggling with their coursework. The required subject and elective subject determine the student's total semester grade. Students often raise their overall grades by giving loose grades to elective subjects. This study will eliminate the disruptions caused by lenient electives and provide timely support for students who need to be alerted to required subjects.

Keywords: Institutional Research, Azure Machine Learning, early alert system, required subject

1 Introduction

Institutional Research (IR) focuses on University administration and data-driven decision-making (DDDM) as the primary focus. The IR demand for data analysis technology and the progress of data quality are constantly evolving. From academic analytic (AA), which supports campus decisions, to learning analytic (LA), which integrates student learning experiences to improve the quality of teaching and learning, all rely on evidence-based decision-making. Incorporating the concept of education data mining (EDM), which collects many students' learning experiences into IR. [1][2][3] With the growing trend of Artificial Intelligence (AI) and the popularity of Machine Learning (ML), more and more new methods are being brought into the analysis of student learning outcomes, improving the value of school research applications and decision-making capabilities and enhancing the efficiency of student counseling. Due to the advances in machine learning and AI in the last decade, Universities around the world have developed students early alert systems (EAS) by other names such as "early warning systems," The advancement and extensive application of information technology has also driven the trend of using learning analytics to develop early warning systems in higher education overseas, such as in North America and Australia [4], with breakthroughs in both practice and scale, and more research on the overall operational mechanism [6][7]. To improve the quality of learning through the school's support system (counseling center or teaching center) to avoid failing or the student dropping out

[*] Executive Master of Business Administration, National Taichung University of Education, Taichung, Taiwan
[†] Master Program of Business Administration, Department of Creative Design and Management, National Taichung University of Education, Taichung, Taiwan

[8]. Consider those learning outcomes are highly correlated with student learning counseling issues such as student pass rates, learning experiences, retention rates, and course participation. [4][9][10][11][12][13][14][15].

2 Problems

The junior (third year) of higher education in Taiwan is the most stressful stage for students in university courses, completing half of the university education but also facing the critical point of future education or employment.

2.1 Research Questions

Therefore, we built the model for juniors, and if the prediction was good. We gradually made the EAS for the required subjects in first- and second-year students (see Figure 1) and did it in advance for the compulsory courses of junior students early warning. This EAS is built on the AML cloud platform to solve the above problems, using its powerful and highly flexible AutoML cloud computing capabilities. We believe in promoting the most efficient student learning counseling mechanism by combining mandatory academic performance with proper algorithms in AML with empirical data.

2.2 At-risk Students Detection in AML

To precisely identify at-risk students' required subjects for EAS, this study was conducted by junior students enrolled in the early years. A list of at-risk students who are in the bottom 25% of the class in their junior year of study is produced for each department in Figure 1. First, 80% of the training data were sent to the AML platform to build the EAS used in this study, and then the remaining 20% of the data were used as the test set to validate the model. The test set in this study is statistically representative of the population due to randomization. Once the EAS for data training was established, the subjects of this study: junior college students whose historical data in first- and second-year required subjects, were entered into the EAS.

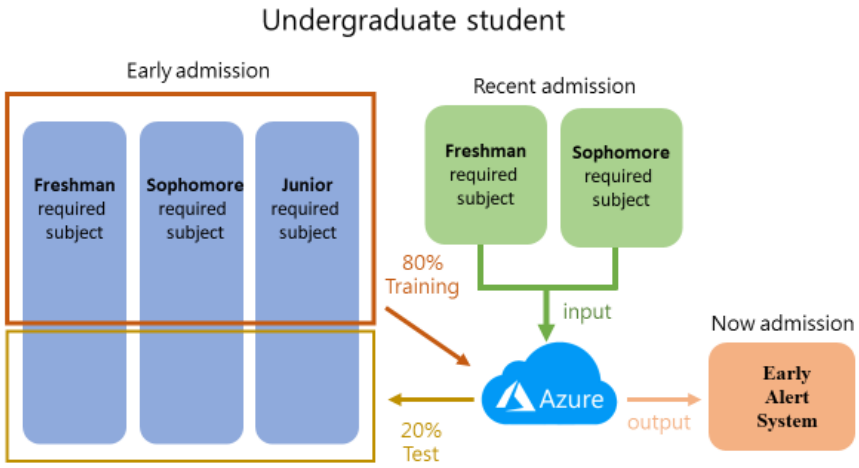

Figure 1: Proposed Early Alert System structure in Azure Machine Learning platform

3 Related works

In recent years, the COVID-19 pandemic has promoted the transformation of enterprise cloud business and industrial digitalization, thereby promoting the growth of the overall market. Various international corporations have announced cloud computing platforms; the famous trio of Amazon AWS, Google Cloud Platform (GCP), and Microsoft Azure dominate the market [14][16][17][18]. Every University's IR also developed a cloud-based EAS. Ravikumar et al. used students' mid-semester and end-of-semester grades, the SGPAs (semester GPA), and the CGPA (cumulative GPA) in the development of an Early Warning System (EWS) [19], added peer-tutoring help to reducing attrition of "at-risk" students at the earliest point in time for detecting curated data from six semesters Building an AI system through intelligent tutoring systems (ITS) development of an EWS, from demographic data (e.g., gender, semester, student id, age, admission channel, etc.), self-reported questionnaires, continuous assessment results, set a threshold for evaluating the quality of the predictive model that can detect at-risk students[18]. Also, to set up mathematical modeling in evidence-based analytics to explain students' learning performance and the relationship in big data [16].

4 Azure Machine Learning

Azure Machine Learning is a cloud service that helps data scientists and developers build, deploy, and manage high-quality models speedy and more confidently. Two environments, Designer and AutoML, were used. For illustration, the Department of Early Childhood education (AEC) data set will be used as an example with the analysis process and final results of the operation ML.

4.1 Designer Environment

The AML designer is a drag-and-drop interface used to train and arrange models. Figure 2 illustrates the tasks performed by the designer for this study. Because of the normal vs. at-risk students, a classical binary classification, we use built-in three algorithms in the AML designer workspace, the two-class boosted decision tree, the two-class averaged perceptron, and the two-class logistic regression.

4.2 AutoML Environment

Azure Automated Machine Learning (AutoML) is a process that automates machine learning model development's time-consuming and iterative tasks. AutoML has several advantages that allow data scientists, analysts, and developers to build ML models with high scalability, efficiency, and productivity while maintaining model quality.

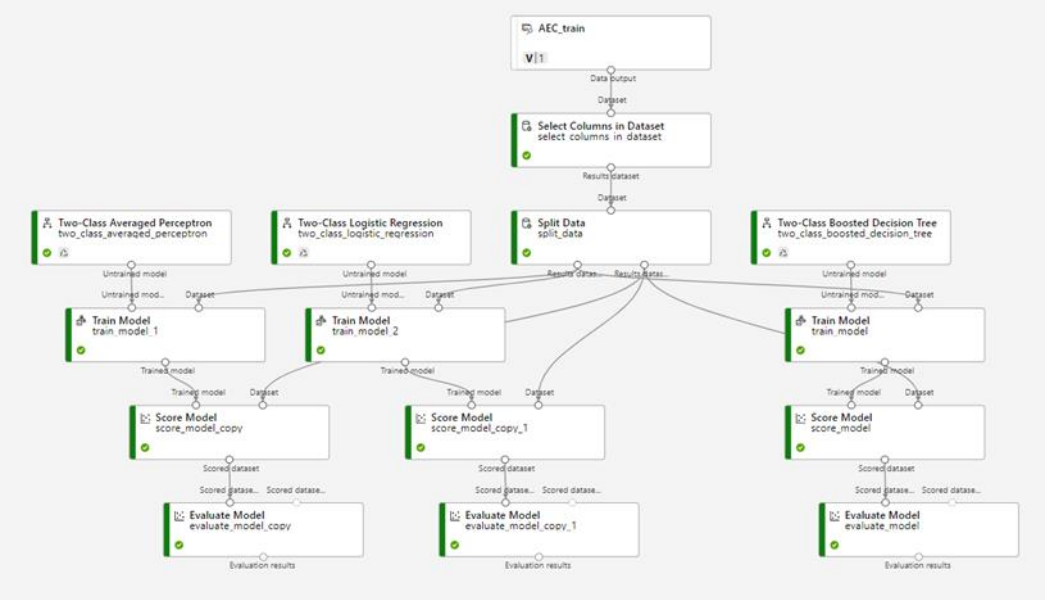

Figure 2: The EAS for the dataset AEC (Department of Early Childhood Education) in an AML designer environment.

4.5 Data Resource

The NTCU provides its faculties and researchers with a database to encourage literacy of evidence-based decision-making: the NTCU data mart. The student background database includes: gender, nation, aboriginal, and household registration address belonging to school_background.csv, collected by the IR dataset. Table 1. is the Department of Early Childhood Education variables. Respectively department, class, semester, student id, the department required subjects (ex: AEC25080), college required subjects (ex: ZCE00020), and admission channel belong to score of students.csv.

Table 1: school_background.csv and score of students.csv of AEC

score of students.csv		
Variables	**Data-type**	**Description**
AEC	String	Department of Early Childhood Education
std_id	String	student identification number
AEC25080	Double	Health and Safety of Young Children
AEC21100	Double	Literature for Young Children
AEC25150	Double	Professional Ethics of Early Childhood Practitioners
AEC25070	Double	Observation Methods of Young Children's Behavior
AEC25060	Double	Development and Care of Young Children
AEC20030	Double	Education for Young Children with Special Needs
ZCE00020	Double	Enhancement and Practice of Mind and Body Potential
ZCE00030	Double	Application of Technology in Education

school_background.csv		
gender	String	gender of students
nation	String	nations of students
aboriginal	Integer	1: yes, 0: no
race	String	race of aborigines
entr_prog	String	entrance program of students
in_year	Double	entrance year of students

5 Exploratory Analysis and Results

To improve the accuracy, we designed each department separately with its own EAS model, and its performance is organized in Table 2.

5.1 The AUC and ROC

The AML evaluates the model using the AUC (area under the ROC curve) and ROC (receiver operating characteristics). Figure 3 shows that the area under the ROC curve is 0.923 (also see in Table 2). Different color lines represent the ROC curve of the AEC dataset with varying calculations of weighting.

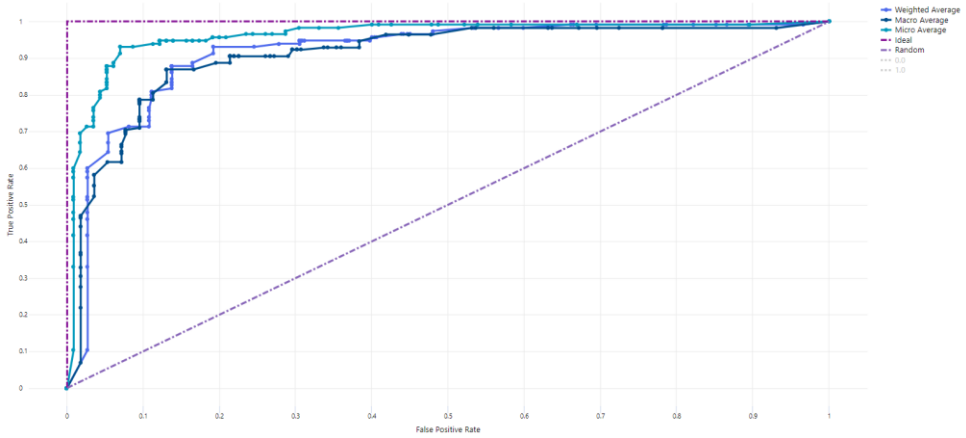

Figure 3. The ROC cure of AEC in different weighted average methods, by various different colors.

5.2Experiments Results

In this study, the critical indicators of the EAS model performance were organized in Table 2. The AUCs of AEL, APE, ASO, and ATA all exceeded 95%. The ACC, AEL, APE, ASO, and ATA departments exceeded 95% regarding accuracy. For example, ACC has an AUC of 94.7% and an accuracy rate of 95.5%.

Table 2: Performance evaluation of each department's EAS in AML

Department	Accuracy	Precision	AUC	Algorithm
ACA	0.852	0.857	0.759	ExtremeRandomTrees
ACC	0.955	0.750	0.947	VotingEnsemble
ACS	0.857	0.667	0.904	LightGBM
ADT	0.846	0.667	0.724	XGBoostClassifier
AEC	0.931	0.600	0.923	LightGBM
AEL	0.957	1.000	0.978	VotingEnsemble
AEN	0.857	1.000	0.674	VotingEnsemble
AIB	0.784	0.700	0.803	VotingEnsemble
ALA	0.929	1.000	0.720	LightGBM
AMU	0.909	0.667	0.746	XGBoostClassifier
APE	0.963	1.000	0.978	VotingEnsemble
ASO	1.000	1.000	1.000	ExtremeRandomTrees
ASP	0.828	1.000	0.864	XGBoostClassifier
ATA	0.955	1.000	0.950	VotingEnsemble

5.3 Feature Importance

AutoML provides a visualization of feature selection as in Figure 4. The course Health and Safety of Young Children (AEC25080) had the highest effect of 0.75, and the second highest effect was on Observation Methods of Young Children's Behavior (AEC25070) at 0.49. The third highest was Professional Ethics of Early Childhood Practitioners (AEC25150) at 0.44. High feature importance values indicate the course AEC25080 (Health and Safety of Young Children) has a strong impact on the Department of Early Childhood education model's prediction.

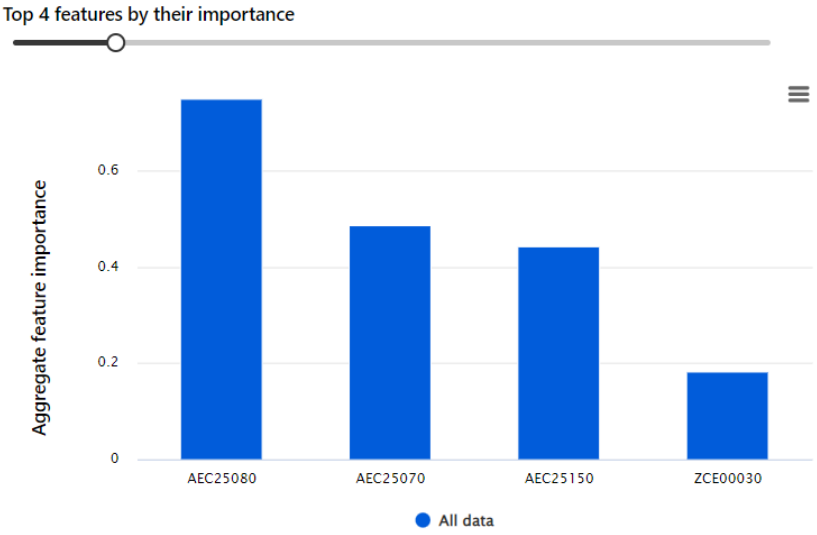

Figure 4: The confusion matrix evaluating the performance EAS of AEC.

6 Conclusion

This paper reports on the research findings of the EAS, which we believe contributes to our understanding of the issues related to the required subjects of higher education. Specifically, our research shows:

a) EAS built with required subjects reduces the need for students to raise their semester totals by lenient electives, interfering with the proper identification of at-risk students.

b) The feasibility of an early-alert prototype for higher education, detailing the design logic and process of implementing this system.

Departments can reorganize the core curriculum structure according to each department's feature importance and the required courses with high weighting and provide feedback to departments for review and adjustment of future curriculum teaching priorities.

Acknowledgement

I sincerely thank Mr. Zhang He-Jun, who provided invaluable assistance in the data organization and AutoML manipulation.

References

[1] R. Ferguson, "Learning analytics: Drivers, developments and challenges," International Journal of Technology Enhanced Learning, vol. 4, pp. 304-317, 01/01 2012, doi: 10.1504/IJTEL.2012.051816.

[2] K. Verbert, N. Manouselis, H. Drachsler, and e. duval, "Dataset-driven Research to Support Learning and Knowledge Analytics," Educational Technology & Society, vol. 15, 01/01 2012.

[3] E. B. Mandinach and E. S. Gummer, "A Systemic View of Implementing Data Literacy in Educator Preparation," vol. 42, no. 1, pp. 30-37, 2013, doi: 10.3102/0013189x12459803.

[4] Y.-H. Hu and H.-Y. Yu, "Improving Retention Rate Through Educational Data Mining: The Design of Placement Program for Newly Enrolled Students," (in Traditional Chinese), Journal of Research in Education Sciences, vol. 65, no. 4, pp. 31-63, 2020, doi: 10.6209/jories.202012_65(4).0002.

[6] J. L. Hung, M. C. Wang, S. Wang, M. Abdelrasoul, Y. Li, and W. He, "Identifying At-Risk Students for Early Interventions—A Time-Series Clustering Approach," IEEE Transactions on Emerging Topics in Computing, vol. 5, no. 1, pp. 45-55, 2017, doi: 10.1109/TETC.2015.2504239.

[7] S. M. Jayaprakash, E. W. Moody, E. J. M. Lauría, J. R. Regan, and J. D. Baron, "Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative," Journal of Learning Analytics, vol. 1, no. 1, pp. 6-47, 05/01 2014, doi: 10.18608/jla.2014.11.3.

[8] R. Villano, S. Harrison, G. Lynch, and G. Chen, "Linking early alert systems and student retention: a survival analysis approach," Higher education, vol. 76, 11/01 2018, doi: 10.1007/s10734-018-0249-y.

[9] K. Arnold, "Signals: Applying Academic Analytics," EDUCAUSE Quarterly, vol. 33, 01/01 2010.

[10] M. h. Abdous, W. He, and C.-J. Yen, "Using Data Mining for Predicting Relationships between Online Question Theme and Final Grade," Educational Technology & Society, vol. 15, 01/01 2012.

[11] M. Saqr, U. Fors, and M. Tedre, "How learning analytics can early predict under-achieving students in a blended medical education course," Medical Teacher, vol. 39, no. 7, pp. 757-767, 04/19 2017, doi: 10.1080/0142159X.2017.1309376.

[12] C. Lacave, A. Molina Díaz, and J. Cruz-Lemus, "Learning Analytics to identify dropout factors of Computer Science studies through Bayesian networks," Behaviour & Information Technology, vol. 37, pp. 1-15, 06/11 2018, doi: 10.1080/0144929X.2018.1485053.

[13] E. Er, E. Gómez-Sánchez, Y. Dimitriadis, M. Bote-Lorenzo, J. Asensio-Pérez, and S. Alvarez Alvarez, "Aligning learning design and learning analytics through instructor involvement: a MOOC case study," Interactive Learning Environments, 05/01 2019, doi: 10.1080/10494820.2019.1610455.

[14] A. Huang, O. Lu, J. Huang, C. Yin, and S. Yang, "Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs," Interactive Learning Environments, vol. 28, pp. 1-25, 07/01 2019, doi: 10.1080/10494820.2019.1636086.

[15] Q. Nguyen, B. Rienties, and J. T. E. Richardson, "Learning analytics to uncover inequality in behavioural engagement and academic attainment in a distance learning setting," Assessment & Evaluation in Higher Education, vol. 45, no. 4, pp. 594-606, 2020/05/18 2020, doi: 10.1080/02602938.2019.1679088.

[16] M. S. A. Razak, S. Abdul-Rahman, and Y. Mahmud, "Mathematics Performance Monitoring System Using Data Analytics," in 2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS), 8-9 Sept. 2021 2021, pp. 1-6, doi: 10.1109/AiDAS53897.2021.9574210.

[17] I. Khan, A. R. Ahmad, N. Jabeur, and M. N. Mahdi, "An artificial intelligence approach to monitor student performance and devise preventive measures," Smart Learning Environments, vol. 8, no. 1, p. 17, 2021/09/08 2021, doi: 10.1186/s40561-021-00161-y.

[18] D. Baneres, M. Rodríguez, A.-E. Guerrero, and A. Karadeniz, "An Early Warning System to Detect At-Risk Students in Online Higher Education," Applied Sciences, vol. 10, p. 4427, 06/27 2020, doi: 10.3390/app10134427.

[19] R. Ravikumar, F. Aljanahi, A. Rajan, and V. Akre, Early Alert System for Detection of At-Risk Students. 2018, pp. 138-142.