Spring 7-10-2024

Driving Student Success through a Data-Driven Approach in Higher Education

Yun Chen
University of Salford, Y.Chen@salford.ac.uk

Kate Han
University of Salford, k.han3@salford.ac.uk

Driving Student Success through a Data-Driven Approach in Higher Education

Dr Yun Chen and Dr Kate Han

Salford Business School, The University of Salford, UK

Abstract

The project aims to explore a data-driven approach to enhance student engagement and achievement in Higher Education (HE),with the ultimate goal of promoting student success. The project utilised a case study approach at the University of Salford, employing data analysis and Machine Learning (ML) techniques to understand corelation between students' engagement and academic performance to support strategy of students support in their learning process. Being a project in progress, this paper delves into the initial phase of our research findings. This phase focuses on the data collected from a pilot module, specifically pertaining to student engagement and progression data. Additionally, the paper presents a prototype of a ML algorithm that aims to facilitate decision-making in the realm of student support. Moving forward, the next stage of the project aims to automate the entire process, spanning from data analysis to student intervention. It aims to use this automation to drive student success throughout their HE journey.

Keywords: Higher Education (HE), student engagement, data-driven approach, student support strategy, data analysis, Machine Learning (ML) algorithm

1.0 Introduction

The concept of "Student Engagement" encompasses the degree of enthusiasm, attentiveness, and dedication that students invest in their educational experiences (Kahn, 2014). Strong student engagement within higher education (HE) can yield numerous positive outcomes, benefiting not only individual students but also the wider academic community (Bryson, 2014). These advantages transcend the confines of the classroom and contribute to personal, academic, and institutional achievements.

It's essential to recognise that student engagement is a shared responsibility, not solely borne by students themselves. Educational institutions also play a pivotal role in fostering and sustaining an environment conducive to student engagement (Bond & Bedenlier, 2019). This involves employing effective teaching techniques, implementing mentoring programs, establishing monitoring mechanisms, and providing ongoing support. When institutions prioritise and invest in strategies to enhance student engagement, the rewards can be substantial for all stakeholders within the realm of HE.

It is believed that there is generally a positive correlation between student engagement and academic performance, but the strength and nature of this correlation can be based on several factors, such as interventions by academic tutors (Fredricks et al., 2019). UK universities are increasingly acknowledging the pivotal role of personal tutoring in bolstering student engagement within the HE milieu and promoting future success. It is widely acknowledged that personal tutors play a crucial role in cultivating a sense of belonging, which is fundamental to students' development of a learner identity and sustained engagement (Lochtie & Walker, 2022; Ross et al., 2014)

At the University of Salford, a progression framework has been established since 2016, identifying three key facilitators vital for student engagement: fostering a sense of belonging, offering support, advice, and guidance, and building academic confidence. A dedicated group of academic staff members, known as Academic Progress Tutors (APTs), has been appointed to champion these initiatives. APTs have access to university systems that provide data for a comprehensive understanding of student engagement and achievement. They use this information to organise APT meetings with relevant students, offering academic references throughout their program and after graduation. These meetings are designed to facilitate robust personal academic development and professional growth through active engagement. They place a significant emphasis on increasing learner autonomy and challenging students to expand beyond their current capabilities.

This developmental paper introduces a continuing research project conducted at Salford Business School. The primary objective of this research is to tackle the obstacles associated with implementing a data-driven approach to enhance student engagement. In this paper, we will begin by examining the challenges we have encountered, then proceed to explore the potential solutions we have developed, and finally outline the next steps for this project.

2.0 Students Support Challenges in HE

2.1 Data Integration

The pandemic has expedited the adoption of a variety of information systems and learning platforms within HE institutions. These platforms provide invaluable data insights by leveraging existing data resources. However, a significant challenge arises due to the disparate storage of data across various information systems within universities (e.g., QlikView, Jigsaw, Blackboard). Institutions encounter a significant challenge when attempting to integrate various systems for monitoring student engagement, including absence data, academic alerts related to tutorials, and case management. This fragmented data landscape leads to issues of data integration, resulting in unclean, irrelevant, and redundant data. Furthermore, when these systems fail to communicate seamlessly, gaining a comprehensive overview of students at risk becomes a formidable task. Consequently, this poses a substantial challenge for academic tutors who rely on this data to identify low-engaged students and implement interventions to enhance their performance (Aldowah, H., Al-Samarraie, H. and Fauzy, W.M., 2019).

A prevalent research focus aims to tackle this challenge through initiatives like the development of information systems for monitoring student engagement (JISC 2022; StDREAM, 2023). Another approach involves proposing models that aid decision-makers in identifying crucial factors contributing to elevated graduation rates, as discussed by Addison and Williams in 2023.

2.2 Complexity in Student Support

Although information systems and machine learning have enhanced prediction based on data analytics to enable early intervention in instances of academic risks, the most current models often focus solely on subjective student factors without examine the external environmental and objective elements. (Qin, et al., 2023). The COVID-19 global pandemic has introduced substantial disruptions to HEIs teaching and learning, posing challenges that existing models struggle to address. It enhances the importance of the role and operational procedures of personal tutors within universities. In current HE, monitoring student engagement across multiple systems has the potential to impose an increased workload and add barriers for the APT team. Considering the substantial number of students in Higher Education, APT team members are now tasked with overseeing student behaviour and activities across a multitude of systems and platforms, with consideration on external factors who might be the reason to impact students' performance. (Gajewski, E. M., 2023) These systems and platforms generate distinct sets of data, complicating the work of the APT team as they must develop proficiency in comprehending and interpreting these diverse data sources and insights. Furthermore, they are required to amalgamate data from these platforms before making informed decisions, rather than relying on a single system. These additional demands place barriers in the path of the APT team, requiring not only data analysis skills but also strong communication skills for effective student support.

In addition to the two aforementioned challenges, diverse HEIs face unique situations, adding further complexity. This variability encompasses aspects such as dataset capture, adopted information systems, and factors influencing student performance. Consequently, a pilot case study was conducted specifically for Salford Business School, which will be discussed in the following session.

3.0 Pilot Case Study

To address the aforementioned challenges, an internally funded project initiated by the University of Salford has been underway since February 2023. Its primary aim is to explore the integration of data from diverse learning platforms and the meticulous cleansing of this data to eliminate inaccuracies. This process is intended to provide a clearer understanding of the correlation between student engagement and their academic performance. Once the data pipeline is established, it opens the door for the adoption of ML algorithms, ultimately influencing decision-making and strengthening responsive student support.

The project started with interviews involving the APT team at Salford Business School in May 2023. Two interviews took place—one with the Head of the APT team and another with the Academic Student Success Lead. These interviews served to validate the data integration and process complexity issues discussed in the prior session. Furthermore, both interviewees expressed concerns about the time-consuming nature of consistently assessing and monitoring the advancement of disengaged students. They also highlighted the challenge of determining the optimal timing for intervention.

During the interviews, another focal point was to pinpoint the essential attributes that the project should utilise for predicting students' performance, particularly for Salford Business School. This will be elaborated upon in the later part of this session.

After conducting interviews, we selected a pilot module, i,e, Level 7 in Information Systems and Digital Transformation, to develop our machine learning algorithm for data analysis and predicting students' performance. This work will establish the groundwork for upcoming phases.

3.1 Establishing the Data Pipeline

The establishment of this data pipeline forms the bedrock of our research and sets a noteworthy precedent for the consistent management of extensive datasets in HE. Operating as a pilot dataset, we gathered data from both QlikView and VLE (i.e., Blackboard). The objective was to integrate data from these two systems while filtering out extraneous information from the raw data. This effort was aimed at uncovering correlations between student engagement and academic performance data with consideration of their personal backgrounds. The selection of relevant features was guided by input from the APT team during interviews, taking into account aspects such as gap awards and Equality Diversity and Inclusion (EDI) considerations.

As indicated in Table 1 below, the integrated dataset for the pilot module consists of selected features. The Blackboard dataset encompasses data on student weekly active time and assignment performance scores, whereas the QlikView dataset contains student personal information and registration data. Both of these initial raw data have undergone a data cleaning process, which involved feature selection and integration. This process was carried out to create a dataset that is well-suited for ML classification purposes which is explained in section 3.2.

Data features	**Data source**
Assignment Score	Blackboard Data
Last Access of Blackboard	
Activities during w/c 08/11/2021	
Activities during w/c15/11/2021	
Activities during w/c 22/11/2021	
Activities during w/c 29/11/2021	
Activities during w/c 06/12/2021	
Activities during w/c 13/12/2021	
Degree	QlikView data
Directorate	
Disability Description	
Domicile	
Ethnicity	
Gender	
Greater Manchester Indicator	
HQE (previous degree)	
Nationality	
Postcode on Entry	
Previous Institution Name	
Program Title	
Region	
Residency Summary	
UK Region	
Year of Program	

Table 1. Pilot Dataset Features

3.2 Developing Machine Learning Algorithms

Following the establishment of the data pipeline, we progressed to the subsequent stage of creating training, testing, and validation datasets, with the objective of applying ML algorithms for classification. The target is the final assessment score, with features encompassing student personal information and Blackboard weekly active time.

Once the dataset was prepared, we applied various classic state-of-the-art machine learning classification algorithms, including Random Forest (Rigatti, S.J., 2017), AdaBoost (Schapire, R.E., 2013), Gradient Boosting (Natekin, A. and Knoll, A., 2013), and the Voting Classifier (Ruta, D. and Gabrys, B., 2005). For the selected module with a small dataset, we achieved promising results (see Figure 1).

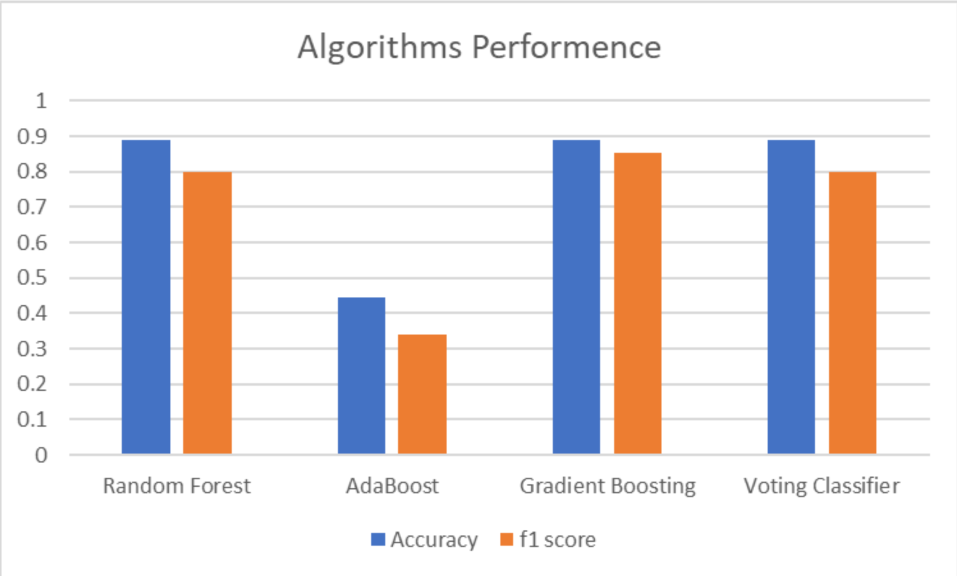

Figure 1. 2Algorithms Performance Comparison

Figure 1 shows the performance of the four ML algorithms that we employed on the pilot dataset. Among these algorithms, Random Forest, Gradient Boosting, and Voting Classifier all achieved an Accuracy score of 0.8888 within the range of 0 to 1. However, AdaBoost fall behind with an Accuracy score of 0.4444 within the range of 0 to 1. In terms of the F1 score within the range of 0 to 1, Gradient Boosting outperformed the others, attaining the highest value of 0.8518. This showcases the promising potential of the developed ML algorithm, which has room for improvement through further training with a diverse range of datasets in future research.

4.0 Conclusion and Future Work

In this research, we successfully accomplished the integration of raw data from QlikView and VLE (i.e. Blackboard) used at the University of Salford. Additionally, we extract information APT interview results to facilitate feature selection for data pipeline and achieved promising outcomes through the application of cutting-edge ML algorithms. The next step of this project entails the acquisition of a more comprehensive dataset to enhance the training of the ML algorithm. This extended dataset will encompass modules ranging from Level 4 to Level 7 to refine the accuracy of the developed model. Building upon a more precise ML algorithm, the next project phase also seeks to automate various functions within the APT team's workflow. This includes automating interventions based on data analysis, progress monitoring, and generating feedback reports.

The future work will be implemented in the following three phases:

1. **Establishment of Dataset Benchmark**

Considering the results obtained in this project, it is clear that the expansion of the data pipeline to establish a benchmark dataset is not only imperative but also holds significant potential. This benchmark dataset will include modules of different sizes and levels of complexity. In addition to the current module-specific active time tracking, we will also gather general student engagement data, including library learning time and email activity for benchmark dataset set up in Academic Year 2023-24. This diversification is vital as it guarantees that the research findings derived from the benchmark will maintain their robustness and adaptability in the face of ever-changing real-world scenarios. Additionally, it is planned to make this dataset available to the wider community and potentially organise a competition to stimulate research interest within the field of ML.

2. **Development of Ensemble Learning Algorithms**

As a natural progression, following the establishment of the benchmark dataset, the following phase involves delving into the application of ensemble learning algorithms. These algorithms will be specifically geared towards generating classification and prediction outcomes in scenarios involving incomplete data, a frequent occurrence in real-world situations. It is expected these algorithms can facilitate swift decision-making by providing timely prediction results.

3. **Creation of APT Process Automation**

After engaging in preliminary conversations with the APT team, we have acquired valuable insights into their operational workflow and the difficulties they encounter. As the last stage of this project, we aim to provide assistance to by initially replicating their processes and subsequently enhancing them through the application of the enhanced ML algorithms. Additionally, we intend to work in close partnership with the APT team, employing an agile approach to continually assess and propose process automation process.

References:

Addison, L. and Williams, D. (2023). Predicting student retention in higher education institutions (HEIs), in *Higher Education, Skills and Work-Based learning*, Vol. 13 No. 5, pp. 865-885. Available at: https://doi.org/10.1108/HESWBL-12-2022-0257

Aldowah, H., Al-Samarraie, H. and Fauzy, W.M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. Telematics and Informatics, 37, pp.13-49.

Bond, M., & Bedenlier, S. (2019). Facilitating Student Engagement Through Educational Technology: Towards a Conceptual Framework. *Journal of Interactive Media in Education : JiME*, *2019*(1). Available at: https://doi.org/10.5334/jime.528

Bryson, C. (2014). *Understanding and developing student engagement*. Routledge. Available at: https://doi.org/10.4324/9781315813691

Fredricks, J. A., Reschly, A. L., & Christenson, S. (Eds.). (2019). *Handbook of student engagement interventions : working with disengaged students*. Elsevier.

Gajewski, E. M. (2023). The Effects of the Covid-19 Pandemic on Students in the School Setting: The Adaptations Needed for Tiered Supports. PhD Thesis, National-Louis University.

JISC (2022). *Enhancing student engagement using technological solutions*. Available at: https://www.jisc.ac.uk/reports/enhancing-student-engagement-using-technological-solutions

Kahn. (2014). Theorising student engagement in higher education. British Educational Research Journal, 40(6), 1005–1018. Available at: https://doi.org/10.1002/berj.3121

Lochtie, Stork, A., & Walker, B. W. (2022). *The Higher Education Personal Tutor's and Advisor's Companion: Translating Theory into Practice to Improve Student Success (1st ed.)*. Critical Publishing.

Natekin, A. and Knoll, A., (2013). G*radient boosting machines, a tutorial. Frontiers in neurorobotics*, 7, p.21.

Qin, K., Xie, X., He, Q. and Deng, G. (2023). Early Warning of Student Performance with Integration of Subjective and Objective Elements, in *IEEE Access*, vol. 11, pp. 72601-72617, 2023.

Rigatti, S.J. (2017). *Random forest. Journal of Insurance Medicine*, 47(1), pp.31-39.

Ross, Head, K., King, L., Perry, P. M., & Smith, S. (2014). The personal development tutor role: An exploration of student and lecturer experiences and perceptions of that relationship. *Nurse Education Today*, 34(9), 1207–1213. Available at: https://doi.org/10.1016/j.nedt.2014.01.001

Ruta, D. and Gabrys, B., (2005). Classifier selection for majority voting. *Information fusion*, 6(1), pp.63-81.

Schapire, R.E., (2013). Explaining adaboost. *In Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik* (pp. 37-52). Berlin, Heidelberg: Springer Berlin Heidelberg.

StREAM (2023). Website: https://www.solutionpath.co.uk