MSc Project Proposal

Student Behavior Analysis and Prediction on Online Learning Platforms Based on Machine Learning

Geng Rui

Contents

Introduction

Online learning platforms have become central to modern education, offering flexible and scalable resources to diverse learners worldwide. According to OECD (2023), global enrollment in online and blended learning programs has grown by more than 15% in the past decade, highlighting the accelerating trend of education digitalization. These platforms generate vast amounts of interaction data, such as video viewing logs, quiz performance, discussion forum activity, and navigation history. When analyzed effectively, this data can provide deep insights into student engagement and learning outcomes.

However, most existing platforms lack predictive analytics tools to identify disengaged or at-risk students in a timely manner. This challenge is particularly urgent in vocational education, where learners often face unique circumstances such as balancing study with employment and needing workforce-oriented skills. Without data-driven intervention, these students are at a higher risk of dropping out or failing to achieve their learning goals.

This project aims to design and implement a machine learning–driven framework to predict student engagement using behavioral data. During the MSc study period, synthetic datasets will be used to build and validate a prototype, avoiding privacy concerns and data access restrictions. After returning to China, the framework will be extended using real-world data from the ICVE (Intelligent Cloud Vocational Education) platform, where the researcher has already developed multiple course repositories. This dual-phase approach ensures both short-term feasibility and long-term impact, contributing to the improvement of retention and performance in vocational education.

Problem Statement

Challenges Motivating This Project

Data Challenges

Online learning platforms generate complex, high-dimensional datasets containing thousands of behavioral features. However, each student only interacts with a small subset of these features, resulting in sparse and noisy data. Traditional statistical approaches often struggle to handle this complexity, leading to inaccurate predictions and unreliable insights. Moreover, missing or inconsistent data further complicates analysis, making it difficult for educators to act on the results with confidence.

Model Challenges

While advanced machine learning models, such as deep learning

architectures, can achieve high predictive accuracy, they often function as “black boxes.” This lack of interpretability poses a barrier to adoption in educational contexts. Teachers and administrators need transparent, understandable outputs to guide interventions. Without interpretability, even highly accurate models may be rejected by practitioners, preventing meaningful improvements in student support.

Contextual Challenges in Vocational Education

Most existing research in learning analytics focuses on research universities or MOOCs. Vocational education remains severely underrepresented, despite its vital role in preparing a skilled workforce. The behavior patterns of vocational learners differ significantly from those of university students, and models developed for MOOCs may not be transferable to vocational settings. This lack of tailored systems means teachers in vocational institutions cannot effectively monitor engagement or predict dropout risk, leading to inefficiencies and missed opportunities for early intervention.

Stakeholder Impact

The absence of predictive systems affects multiple stakeholders. Students are directly impacted when disengagement goes unnoticed, resulting in poor academic outcomes or dropout. Educators lack the tools to provide targeted, data-driven support, while institutions face reputational and financial risks due to low retention rates and inefficient resource allocation.

Importance of Solving the Problem

Addressing these challenges is critical for improving vocational education quality and equity. By developing a framework that balances accuracy and interpretability, this project will enable educators to act on reliable insights, improve student outcomes, and ultimately strengthen workforce preparation. In the long term, the system can serve as a scalable model for other educational contexts, contributing to broader digital transformation in education.

Aims and Objectives

Aim

The primary aim of this project is to design, implement, and evaluate a machine learning–driven framework for analyzing and predicting student engagement on online learning platforms.

The framework will first be developed and validated using synthetic datasets during the MSc study period, and later extended to real-world vocational education contexts through integration with authentic ICVE course data when

the researcher returns to the home institution.

Objectives

Prototype Development with Synthetic Data

Build a minimum viable prototype (MVP) consisting of three essential modules only:

1. Data pipelines,
2. Feature selection component,
3. A single core predictive model (e.g., Logistic Regression).

Use synthetic datasets to simulate typical online learning behaviors, including video interactions, quiz attempts, and forum participation.

Ensure the prototype demonstrates the full end-to-end workflow while maintaining a manageable scope during the MSc study period.

More advanced features, such as multiple model integration and complex dashboards, will be developed after returning to the home institution.

Feature Processing and Model Integration

Implement essential feature engineering and one core feature selection method to address the challenges of high-dimensional and sparse datasets common in online learning behavior data.

During the MSc phase, focus on a single core predictive model, such as Logistic Regression, to ensure transparency and feasibility.

Introduce an interpretability module using SHAP (SHapley Additive Explanations) to provide clear, visual explanations of how different behavioral features (e.g., quiz performance, video engagement, forum activity) influence predictions.

These insights will help educators understand and trust the system's outputs, bridging the gap between advanced machine learning techniques and practical teaching needs.

Define measurable performance targets for the MSc prototype to ensure technical robustness and well-defined objectives:

- Accuracy ≥ 80%
- F1-score ≥ 0.75
- AUC ≥ 0.80
- Dashboard load time < 3 seconds

Additional models with higher predictive performance but lower interpretability, such as Random Forest or Gradient Boosting, will be integrated after the MSc stage as part of future work.

The overall system will be built using modular architecture, ensuring that future models can be seamlessly added without compromising explainability or clarity.

Web-Based Dashboard and Visualization

Develop a simple and functional dashboard for visualizing model outputs and learning behavior summaries.

Provide at least two core visualizations:

1. Overall engagement trends,
2. Individual student prediction results.

Keep the interface lightweight during the MSc stage, while designing it for future expansion with more advanced visualization and interaction features.

Validation and Sustainability

Validation during the MSc study period

During the MSc study period, the system will be validated primarily using synthetic data.

Model performance will be assessed with standard machine learning metrics such as Accuracy, Precision, Recall, F1-score, and AUC.

Usability evaluation will be limited to simulated users, such as classmates or research colleagues, who will test the interface and basic workflows.

This approach ensures that the prototype can be evaluated without breaching privacy or requiring access to authentic vocational educators and real student data.

Explanation: Because the researcher is currently studying abroad and cannot access the ICVE platform's authentic data or real teacher users, usability testing will only focus on verifying basic interface logic and interaction during this stage.

Validation after returning to the home institution

Upon returning to Jiangxi Applied Engineering Vocational College, the researcher will conduct a full usability evaluation using authentic course data and real users.

Teachers will be invited to use the system to analyze actual learning behaviors from ICVE platform courses such as Mini-Program Application Development and

Java Programming Development.

Feedback will be collected through questionnaires and semi-structured interviews, focusing on the system's usefulness, ease of interpretation, and practical value in teaching decision-making.

A/B testing may also be conducted to compare teaching outcomes before and after using the system, ensuring robust validation of its impact.

Explanation: This two-phase validation plan balances current feasibility with long-term research goals and ensures that the system evolves from a synthetic-data prototype to a real-world, educator-supported application.

Timeline and Future Work

Overall Plan

MSc stage: Deliver an MVP prototype between September and November, focusing on core functionality only.

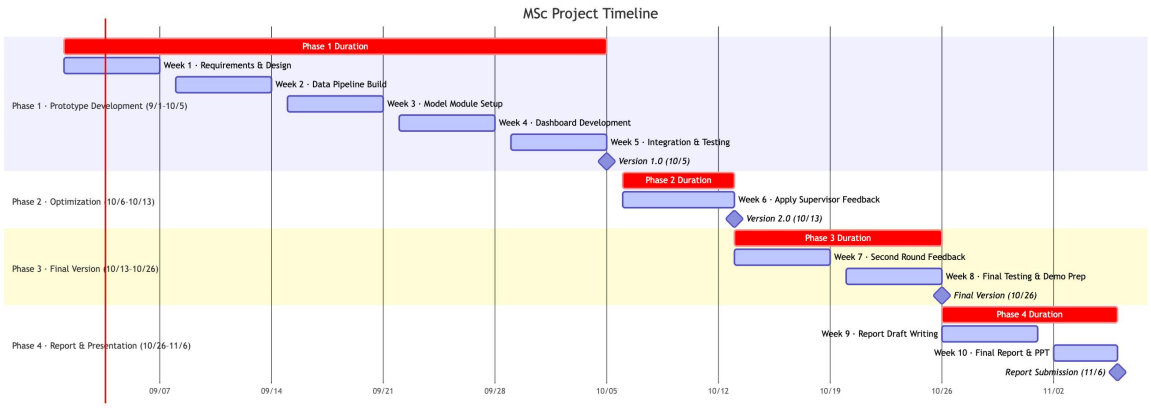
Post-MSc stage: Gradually add advanced models, extended visualizations, and large-scale usability testing.

This staged approach ensures short-term feasibility while laying the foundation for long-term scalability and real-world impact.

MSc Timeline

The MSc project will be completed in four phases between September 1st and November 6th, aligned with key milestones:

- Version 1.0: Initial prototype completed by October 5th
- Version 2.0: Updated version incorporating supervisor feedback by October 13th
- Final version: Fully tested and finalized by October 26th
- Report & Presentation: Final MSc project report and presentation materials completed by November 6th



Summary of Approach

By combining synthetic data prototyping with a structured plan for real-world data integration, this project achieves short-term deliverables while laying the foundation for long-term scalability.

The MSc phase will produce a functional prototype and evaluation report, while the post-MSc phase will focus on authentic data integration and deployment in vocational education settings.

Legal, Social, Ethical and Professional Considerations

Legal and Data Protection

During the MSc phase, only synthetic datasets will be used for system design and testing, ensuring that no identifiable student information is involved. This completely eliminates privacy concerns and reduces the risk of data breaches.

After returning to China, when real-world data from the ICVE (Intelligent Cloud Vocational Education) platform is integrated, strict compliance with relevant data protection laws will be followed, including:

- China's Personal Information Protection Law (PIPL) for handling and processing student-related data;
- EU's General Data Protection Regulation (GDPR) for ensuring privacy and security standards.

All collected data will be anonymized and encrypted before storage and analysis. Data access will be restricted to authorized researchers only, and a formal data usage agreement with ICVE will be signed to define data ownership, access permissions, and usage limitations. These measures ensure that the research process remains legally compliant and transparent.

Ethical Responsibilities

The predictive system developed in this project will act solely as a decision-support tool for educators rather than an automated decision-making system.

- Teachers will retain full authority over intervention strategies, ensuring that machine learning outputs do not directly influence student grades or progression.
- Ethical safeguards will be implemented to prevent algorithmic bias or discrimination, particularly toward vulnerable student groups.
- Prior to using real data, ethical approval will be sought from the relevant institutional review boards (IRB) to ensure full compliance with international

research ethics standards.

This approach builds trust among stakeholders and promotes responsible use of educational analytics.

Social and Educational Impact

The project has the potential to generate significant positive social and educational outcomes.

By identifying disengaged or at-risk students early, it supports targeted, data-driven interventions that improve student retention rates and learning outcomes.

It enhances educational equity, enabling teachers to provide better support for diverse learners, including those balancing study and employment in vocational education.

Robust privacy and security measures help prevent data misuse, reinforcing public trust in educational technologies.

The system design offers a scalable model for vocational institutions globally, contributing to the broader digital transformation of education.

Professional Integrity

The researcher will maintain the highest standards of professional integrity throughout the project:

- All data processing steps, model selections, and evaluation methods will be documented transparently to ensure replicability and accountability.
- Results will be reported honestly, without manipulation or selective disclosure.
- Collaboration with ICVE and other stakeholders will strictly follow professional codes of conduct, ensuring that all participants are fully informed about the project scope, potential risks, and expected benefits.

Additionally, to avoid any misunderstanding about the data included in this document:

- The screenshots and repository links presented within the main body of this report are included solely to demonstrate data accessibility and system design feasibility.
- These visuals do not contain any identifiable student information.
- During the MSc phase, only synthetic datasets will be used to develop and test the system. Integration of authentic ICVE data will occur after the researcher returns to China, and only after completing all required legal

agreements and ethical approval procedures.

Background

Introduction: Rise of Online Learning Platforms

Online learning platforms have rapidly transformed education by providing flexible, scalable, and accessible resources for learners worldwide.

According to OECD (2023), global enrollment in online and blended learning programs has increased by over 15% in the past decade, highlighting the growing importance of digital education infrastructures.

These platforms generate vast amounts of behavioral data, such as:

- Video viewing patterns,
- Quiz performance,
- Discussion forum participation, and
- Navigation histories.

This data has the potential to reveal critical insights into student engagement, motivation, and dropout risk. However, without effective analytics tools, raw data cannot be transformed into actionable strategies for teaching or timely interventions.

This challenge is especially urgent in vocational education, where learners often require personalized support to develop workforce-ready skills. Many vocational institutions lack predictive systems to identify at-risk students early, leading to poor retention rates and reduced learning outcomes (UNESCO, 2022).

This project addresses these challenges by developing a machine learning–driven framework to analyze and predict student engagement.

- During the MSc phase, synthetic datasets will be used to design and validate a prototype, avoiding privacy and data access issues.
- Upon returning to China, the researcher will extend the system using real data from the ICVE (Intelligent Cloud Vocational Education) platform, where they have already developed a comprehensive course repository.

This two-phase approach ensures short-term feasibility and long-term scalability.

State of the Art: Learning Analytics and Educational Data Mining

The fields of Learning Analytics (LA) and Educational Data Mining (EDM) form

the foundation for modern online learning research:

- Learning Analytics focuses on analyzing learner data to optimize both learning processes and teaching environments (Ferguson, 2012).
- Educational Data Mining develops algorithms to discover hidden patterns in educational datasets (Romero & Ventura, 2020).

Together, they enable predictive systems that can monitor engagement, predict outcomes, and support data-driven interventions.

Machine Learning Techniques for Student Prediction

Current models for predicting student performance and engagement fall into two main categories:

1. Traditional Models – Logistic Regression, Decision Trees

Strength: Highly interpretable and easy to implement.

Limitation: Struggle to capture complex, nonlinear relationships in high-dimensional datasets, leading to limited predictive accuracy.

2. Advanced Models – Deep Learning Architectures

Examples: Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks.

Strength: Capable of modeling sequential behavioral data, such as how past study sessions influence future actions (Tang et al., 2020).

Limitation: Require extensive computational resources and are often viewed as “black boxes” with limited transparency.

Persistent Challenges

Despite progress, two key challenges remain:

1. High-Dimensional and Sparse Data

Online platforms generate thousands of potential features, but each student interacts with only a fraction of them, creating sparse datasets with many zeros or missing values.

Zhao et al. (2021) emphasize the need for feature selection techniques to reduce noise and focus on the most informative predictors.

2. Accuracy VS Interpretability

Complex models achieve high accuracy but lack transparency.

Educators need interpretable systems they can trust and act upon (Knight et al., 2017).

Without interpretability, even accurate predictions may not be adopted in practice.

Implications for This Project

These challenges highlight the need for systems that:

1. Handle high-dimensional, sparse educational data
2. Provide accurate yet interpretable outputs

This project directly addresses both requirements by designing a machine learning framework tailored for vocational education platforms like ICVE.

Vocational Education Context and Research Gaps

While most LA and EDM studies focus on universities or MOOCs, vocational education remains underrepresented. Yet vocational institutions serve diverse learners who may balance study with work or lack strong academic foundations.

Predictive tools could help teachers identify at-risk students early and provide targeted interventions.

Global Perspective

Growth Trend: Vocational education enrollment worldwide has increased by over 15% in the past decade (OECD, 2023).

Digital Transformation: UNESCO (2022) highlights a shift toward online and hybrid learning platforms, driven by the need for flexible, workforce-oriented education.

Chinese Context

China has made major investments in vocational education modernization, including the ICVE platform, which hosts structured digital resources such as videos, quizzes, and interactive tools.

However, while ICVE generates rich behavioral data, there is little research on leveraging this data for predictive analytics and engagement monitoring.

Research Gaps

Existing predictive models are primarily designed for MOOC or university data, which differ significantly from vocational education patterns.

Transferability issues: These models may not generalize well to vocational contexts.

Teachers lack actionable insights to improve retention and performance.

Project Relevance

This project focuses on vocational education, a critically underserved area in current learning analytics research.

The researcher has played a central role in the development of multiple course repositories on the ICVE platform, serving as the lead instructor for Mini-Program Application Development and actively contributing to the design and resource construction for other courses such as Web Design and Production and Java Programming.

These repositories provide comprehensive digital learning resources, including videos, quizzes, and discussion forums, which will serve as the primary sources of authentic behavioral data for system validation after the MSc phase.

A link to the main course is provided here: [ICVE Course Repository – Mini-Program Application Development]

Note: This link can only be accessed within mainland China. For reviewers outside China, relevant screenshots are provided directly below as clear evidence of the course repository's structure and accessibility.

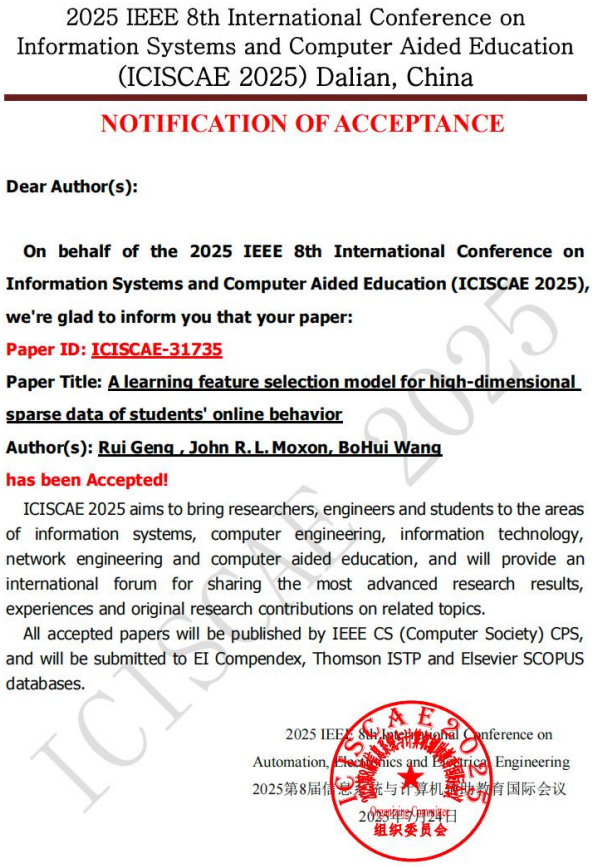
Project Innovation and Research Foundation

This project builds on prior research while introducing novel elements.

Research Foundation

The researcher's EI-indexed paper, A Learning Feature Selection Model for

High-Dimensional Sparse Data of Students' Online Behavior, accepted by ICISCAE 2025, provides the theoretical foundation for this project.



2025 IEEE 8th International Conference on
Information Systems and Computer Aided Education
(ICISCAE 2025) Dalian, China

NOTIFICATION OF ACCEPTANCE

Dear Author(s):

On behalf of the 2025 IEEE 8th International Conference on Information Systems and Computer Aided Education (ICISCAE 2025), we're glad to inform you that your paper:

Paper ID: ICISCAE-31735

Paper Title: A learning feature selection model for high-dimensional sparse data of students' online behavior

Author(s): Rui Geng , John R. L. Moxon, BoHui Wang

has been Accepted!

ICISCAE 2025 aims to bring researchers, engineers and students to the areas of information systems, computer engineering, information technology, network engineering and computer aided education, and will provide an international forum for sharing the most advanced research results, experiences and original research contributions on related topics.

All accepted papers will be published by IEEE CS (Computer Society) CPS, and will be submitted to EI Compendex, Thomson ISTP and Elsevier SCOPUS databases.

2025 IEEE 8th International Conference on
Automation, Electronics and Electrical Engineering
2025第8届信息系统与计算机教育国际会议
2025年7月24日
组织委员会

Engineering Village IEEE

This earlier work demonstrates the feasibility of handling complex educational datasets using specialized algorithms.

Innovation Highlights

Vocational Focus: Addresses the unique needs of vocational learners, filling a critical research gap.

Two-Phase Implementation:

- MSc phase: synthetic datasets for prototype development.
- Post-MSc phase: integration with real ICVE data for large-scale validation.

Accuracy + Interpretability: Incorporates feature attribution methods like SHAP or LIME to ensure outputs are both accurate and understandable.

Practical Integration: The system will be designed for seamless integration into ICVE, with dashboards for instructors to track engagement and risk levels.

Summary of Gaps and Contributions

This project operates at the intersection of learning analytics, machine learning, and vocational education.

It addresses the lack of predictive analytics tools for vocational contexts by providing a framework that balances technical complexity with practical usability.

The dual-phase design ensures immediate proof of concept and long-term scalability.

By leveraging ICVE's digital resources, the project has the potential to improve student retention, optimize resource allocation, and support evidence-based decision-making in vocational institutions.

References

Academic Studies

1. Geng R., Moxon J. R. L. , and Wang B. (2025). A Learning Feature Selection Model for High-Dimensional Sparse Data of Students' Online Behavior. Proceedings of the 2025 IEEE 8th International Conference on Information Systems and Computer Aided Education (ICISCAE 2025).
2. Ferguson, R. (2012). The state of learning analytics in 2012: A review and future challenges. Technical Report, Knowledge Media Institute, The Open University, UK.
3. Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1355.
4. Fredricks, J.A., Blumenfeld, P.C. and Paris, A.H. (2004). School engagement: Potential of the concept, state of the evidence. Review of Educational Research, 74(1), pp.59-109.
5. Baker, R.S., Lindrum, D., Lindrum, M.J. and Perkowski, D. (2021). Analyzing and predicting student engagement in online learning environments. Journal of Learning Analytics, 8(1), pp.34-52.
6. Zhao, J., Liu, X. and Zhang, H. (2021). Feature selection for high-dimensional student behavior data in online learning environments. Knowledge-Based Systems, 229, 107340.
7. Knight, S., Buckingham Shum, S. and Littleton, K. (2017). Epistemology, pedagogy, assessment and learning analytics. Learning, Media and Technology, 42(1), pp.7-25.

8. Tang, T., Song, L. and Chen, Y. (2020). Modeling student behavior sequences with recurrent neural networks for predicting learning outcomes. Computers & Education, 149, 103830.

Policy and Reports

9. OECD. (2023). Education at a Glance 2023: OECD Indicators. OECD Publishing.
10. UNESCO. (2022). Global Education Monitoring Report 2022: Building Skills for an Inclusive Workforce. UNESCO Publishing.

Student and First Supervisor Project Sign-Off			
	Name	**Signature**	**Date**
STUDENT: I agree to complete this project:			
SUPERVISOR: I approve this project proposal:			
Supervisor Comments/Feedback			