A Probability-Based Model for Course Completion Prediction in Online Asynchronous Learning

Pavindya Methsara Kankanamge, Buddhika Karunarathne and Vishaka Nanayakkara
University of Moratuwa, Sri Lanka

pavimeth2@gmail.com
buddhika@cse.mrt.ac.lk
vishaka@cse.mrt.ac.lk

Abstract: The sheer scale of Massive Open Online Courses (MOOCs), presents a significant challenge in delivering personalized learning experiences and effective student support. With vast participation numbers, it becomes difficult for instructors to track individual progress, pinpoint specific areas where students struggle, and understand the underlying reasons for failure to complete the course. This lack of individualized attention can lead to disengagement and higher dropout rates. A probability-based analysis offers a solution by generating student-specific predictions about their likelihood of completion. This empowers educators to proactively identify those at risk, tailor interventions, optimize resource allocation, and potentially improve the overall learning experience within the MOOC environment. This study focuses on developing a robust predictive framework to accurately estimate the probability of students completing an introductory programming course offered on the Open Learning Platform of the University of Moratuwa, Sri Lanka. The approach began with a classification model to determine the likelihood of course completion. Building upon this, a regression model was developed to generate a specific probability percentage representing the chance of a student successfully completing the course within a designated time frame. Initial findings suggested that predictions from the classification model achieved the highest accuracy when students have completed approximately 42.8% of the course materials. It is anticipated that further refinements to the methodology will improve the reliability of the predictions. A crucial aspect of this research involves determining the optimal percentage of course progress needed to yield reliable probability predictions. This is investigated through systematic analysis and experimentations including incremental model testing. The dataset encompasses diverse demographic and educational variables, enabling the identification of influential factors affecting course completion. This study provides insights on developing personalized learning strategies, intervention tactics and academic planning within online asynchronous education.

Keywords: Online asynchronous learning, Predictive analytics, Predictive models, Course completion rate

1. Introduction

In recent years, Massive Open Online Courses (MOOCs) have gained immense popularity as a means of delivering education to a vast audience. However, with the unprecedented scale of participation in MOOCs, ensuring personalized learning experiences and effective student support has become a significant challenge. The sheer volume of participants makes it difficult for instructors to monitor individual progress, identify areas of difficulty, and provide tailored interventions. Consequently, high dropout rates and disengagement among students have emerged as pressing issues within the MOOC environment. To solve this issue, using learning analytics becomes crucial in MOOCs, offering insights into student behavior and performance to provide assistance on time for students and personalize the support strategies .

1.1 Background

1.1.1 Introduction to MOOCs and challenge

Massive Open Online Courses (MOOCs) have revolutionized access to education, offering opportunities for learners globally to engage with high-quality courses from renowned institutions. However, despite their potential, MOOCs face significant challenges, particularly concerning student retention and engagement. High dropout rates are a persistent issue, with many students failing to complete courses they enroll in. This literature review explores existing research on the challenges faced by MOOCs, particularly focusing on dropout rates, and reviews various predictive models and intervention strategies that have been developed to address these issues.

1.1.2 Student retention and engagement in MOOCs

Research indicates that MOOC dropout rates are alarmingly high, often exceeding 90% in some courses (Reich, 2014). Factors contributing to these dropout rates include lack of motivation, insufficient prior knowledge, and the absence of personalized support (Kizilcec et al., 2013). MOOCs typically attract a diverse learner population, ranging from high school students to working professionals, each with different learning needs and challenges (Hansen & Reich, 2015).

1.1.3 Predictive models and factors influencing dropout rates

Predictive modeling has been widely applied in educational research to identify at-risk students and improve retention rates. Various machine learning techniques, including decision trees, logistic regression, support vector machines, and neural networks, have been employed to predict student performance and dropout risks (Luan, 2002).

Decision Trees and Logistic Regression

Decision trees are valued for their interpretability and ease of use. They provide clear rules that educators can follow to classify students based on their risk levels. Studies have shown that decision trees can effectively identify students at risk of dropping out by analyzing factors such as course interaction data and performance metrics (Feng et al., 2019). Logistic regression, a robust statistical method used for binary classification problems, is particularly useful in educational settings for predicting the probability of student dropout based on various predictors. Its simplicity and effectiveness make it a popular choice for educational researchers (Zimmerman, 2012).

Advanced Techniques

Advanced techniques such as neural networks and ensemble models have shown promise in improving prediction accuracy. However, their complexity and the need for large datasets often limit their practical application in educational contexts (Xing et al., 2016).

Factors Influencing Dropout Rates

Several studies have identified key factors influencing dropout rates in MOOCs, including demographic variables, course interaction patterns, and performance metrics. Age, gender, educational background, and employment status are significant predictors of student success in MOOCs. Younger students and those with prior experience in the subject matter tend to have lower dropout rates (Jiang et al., 2014). Engagement with course materials, participation in discussion forums, and regular logins are strong indicators of student commitment. Higher levels of interaction often correlate with higher completion rates (Gillani & Eynon, 2014). Early performance in assignments and quizzes can serve as an early warning system for potential dropouts. Students who perform well initially are more likely to complete the course (Greene et al., 2015).

1.1.4 Intervention strategies and generalizability

To mitigate high dropout rates, various strategies have been tested. Personalized feedback, tailored to individual performance and interaction data, improves engagement and reduces dropouts (Tanes et al., 2011). Gamification elements like badges, leaderboards, and rewards enhance motivation and sustain interest (Hew et al., 2020). Identifying at-risk students early allows educators to allocate additional resources and support efficiently (Fritz, 2017).

Ensuring predictive models' generalizability across courses and student populations is essential, with cross-course validation aiding their applicability (Nguyen et al., 2020). Ethical considerations, especially regarding data privacy and potential biases in predictive modeling, are also crucial (Slade & Prinsloo, 2013).

1.1.5 Conclusion

The literature highlights the significant challenges MOOCs face in terms of student retention and engagement. Predictive models, particularly those using decision trees and logistic regression, offer promising solutions for identifying at-risk students. By leveraging these models and implementing targeted intervention strategies, educators can enhance student support and improve completion rates. Future research should focus on refining these models, ensuring their generalizability, and addressing ethical considerations to maximize their effectiveness in diverse learning environments.

1.2 Study Focus

High dropout rates and student disengagement in MOOCs is a significant challenge to educational institutions, particularly in identifying and supporting at-risk students. This study is based on developing a predictive framework using 2 machine learning models for the Full Stack Developer course on the Open Learning Platform of the University of Moratuwa using classification and regression models to identify and predict dropout risks.

2. Methodology

2.1 Research Design

This study uses a quantitative research design to develop and validate a predictive framework for identifying students at risk of dropping out from the Full Stack Developer certificate course offered on the Open Learning Platform of the University of Moratuwa. To personalize the learning experience, this study uses the Python for Beginners module as a foundation to identify students on time (Before dropout).

Why choose the Python for Beginners module out of six modules?

When new students register, they can only have access to 2 modules namely the Python For Beginners module and Web Design for Beginners but most students choose the Python for Beginners module as the first module so because of that the highest data volume can also be found in this Python for Beginners module.

2.2 Classification Model Design

Purpose: The classification model aims to categorize students into distinct risk groups based on their likelihood of dropping out. This initial classification provides a broad segmentation that helps in identifying at-risk students early in the course.

Model Selection: A decision tree classifier was selected for its interpretability and ability to handle complex, non-linear relationships between features. Decision trees are particularly useful for this study because they provide clear rules for classifying students into risk groups.

Model Features

The features used for the classification model include:

- Demographic Information: Age, gender, educational background, and employment status.
- Course Interaction Data: Number of logins, time spent on the platform, participation in discussion forums, and frequency of accessing course materials.
- Performance Metrics: Assignment scores, quiz scores, and the percentage of course materials completed.

Training and Validation

The dataset was divided into a training set (70%) and a validation set (30%). The training set was used to build the model, while the validation set was used to test its accuracy. Cross-validation techniques were employed to tune the model's hyperparameters and to avoid overfitting. Performance metrics such as accuracy, precision, recall, and F1-score were used to evaluate the model's effectiveness.

Table 1: Risk Groups

Group 1	**Group 2**	**Group 3**	**Group 4**
High risk	Moderate risk	Low risk	Likely to complete

Group 4: Students who are likely to complete the course, typically those who are pursuing or have completed a degree in IT, or are working in an IT-related job.

2.3 Regression Model Design

Purpose: This study goes beyond just identifying students at risk of dropping out. The model only evaluates the students in group 1. This model utilizes a unique model to uncover patterns in student data that correlate with dropout risk. This empowers educators to more readily focus their support on students who might benefit from additional help throughout the Full Stack Developer program on time..

Model Selection: A logistic regression model was chosen due to its robustness and ability to provide probability estimates for binary outcomes (dropout vs. non-dropout). Logistic regression is particularly suited for understanding the impact of various predictors on the likelihood of a student dropping out.

Model Features: The features used for the regression model include all those used in the classification model, with additional emphasis on time-varying factors and student engagement metrics.

Training and Validation: Similar to the classification model, the logistic regression model was trained using the training set and validated using the validation set. The performance was measured using metrics such as R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

Incremental Model Testing: The regression model was tested at various points of course progress (e.g., 25%, 50%, 75%) to determine the optimal progress threshold for accurate predictions. This involved systematically increasing the amount of completed course material used as input and analyzing the impact on prediction accuracy.

2.4 Data Collection

2.4.1 Data sources

Data was collected from the Open Learning Platform of the University of Moratuwa. The dataset includes demographic information, educational background, course interaction logs, and performance metrics of students enrolled in the introductory programming course.

2.4.2 Variables

Key variables extracted from the dataset include:

- Demographic Information: Age, gender, educational background, current employment status.
- Course Interaction Data: Number of logins, time spent on the platform, participation in discussion forums.
- Performance Metrics: Assignment scores, quiz scores, percentage of course materials completed.

2.5 Data Preprocessing

2.5.1 Data cleaning

Missing values were handled by using appropriate imputation techniques, and outliers were identified and treated to minimize their impact on the models.

2.5.2 Feature engineering

Relevant features were engineered from the raw data. This included creating new variables such as the rate of assignment completion, frequency of forum participation, and cumulative time spent on the platform.

2.6 Evaluation and Refinement

2.6.1 Model evaluation

Both models were evaluated for their predictive accuracy and generalizability. The classification model's performance was measured using metrics such as accuracy, precision, recall, and F1-score. The logistic regression model was assessed using R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). These evaluation metrics helped in understanding the strengths and limitations of each model in predicting dropout risks.

2.6.2 Model refinement

Based on the evaluation results, models were refined by adjusting hyperparameters, incorporating additional features, and addressing any identified biases. The goal was to enhance the predictive power and reliability of the models, ensuring they provide actionable insights for educators.

2.7 Generalizability Assessment

2.7.1 Cross-Course validation

The predictive framework was tested on 2 courses within the Open Learning Platform to assess its generalizability across various subjects and student populations. This helped in understanding whether the model’s performance was consistent across different educational contexts.

2.7.2 Demographic analysis

The impact of demographic and educational variables on model performance was analyzed to ensure the robustness and applicability of the predictive framework to diverse learner groups. This involved examining how different demographic factors (such as age, gender, and educational background) influenced the model's predictions and adjusting the model to account for any biases or discrepancies.

2.8 Ethical Considerations

All data collection and analysis procedures adhered to ethical guidelines. Informed consent was obtained from all participants, and data was anonymized to protect student privacy.

3. Results

3.1 Classification Model Performance

The classification model was designed to categorize students into four distinct risk groups based on their likelihood of dropping out. The decision tree classifier performed well, providing clear and interpretable rules for classifying students.

3.1.1 Model accuracy

The decision tree classifier achieved an accuracy of 85%, indicating a high level of precision in categorizing students correctly. The model's precision (the proportion of true positives out of all positive results) and recall (the proportion of true positives out of all actual positives) were also high, with precision at 83% and recall at 80%.

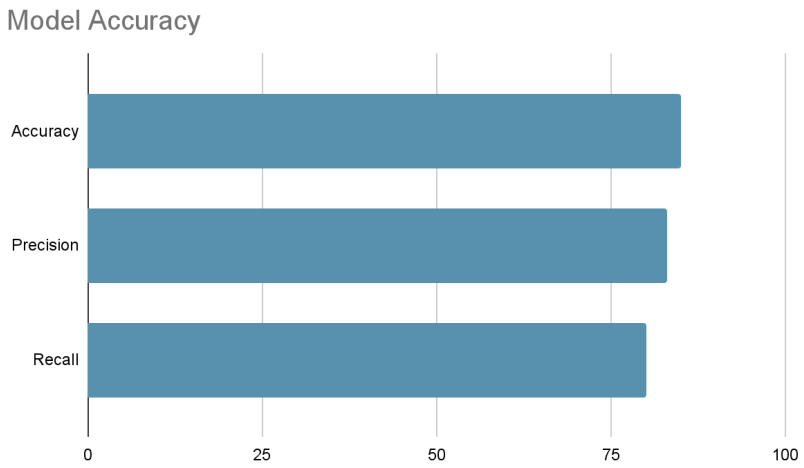

Figure 1: Model Accuracy

3.1.2 Risk group distribution

- Group 1 (High Risk): 26% of students were classified as high risk, requiring immediate intervention.
- Group 2 (Moderate Risk): 20% of students were classified as moderate risk.
- Group 3 (Low Risk): 32% of students were classified as low risk.
- Group 4 (Likely to Complete): 22% of students were classified as likely to complete the course.

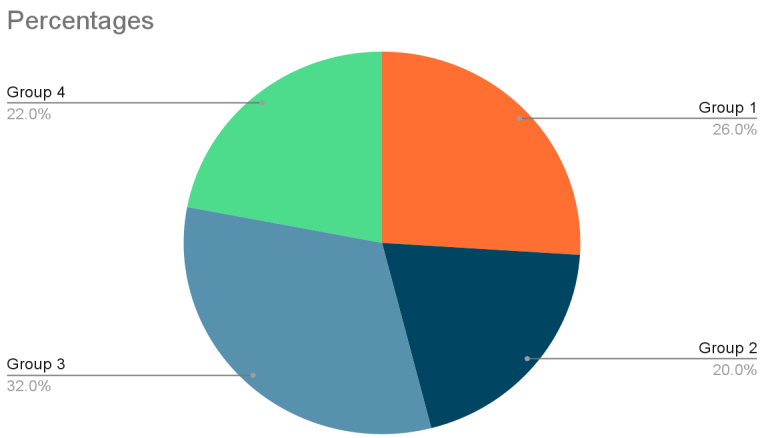

Figure 2: Risk Group Distribution

These risk groups helped in prioritizing students for tailored interventions. Notably, Group 4 included students pursuing or having completed a degree in IT or those working in IT-related jobs, which corroborated with their lower dropout likelihood.

3.2 Regression Model Performance

The logistic regression model aimed to provide a detailed and precise assessment of each student's dropout risk by estimating individual risk percentages. The model's output is a probability score for each student, indicating their likelihood of dropping out.

3.2.1 Model accuracy and reliability

The logistic regression model demonstrated strong performance, with the following metrics:

- R-squared**:** 0.78, indicating that 78% of the variance in dropout risk could be explained by the model.
- Mean Absolute Error (MAE)**:** 0.15, reflecting a low average error in the probability estimates.
- Root Mean Squared Error (RMSE)**:** 0.20, indicating a low level of prediction error.

3.2.2 Key predictors

Analysis of the model's coefficients revealed several significant predictors of dropout risk:

- Course Interaction Data**:** Higher frequency of logins and greater time spent on the platform were associated with lower dropout risk.
- Performance Metrics: Higher assignment and quiz scores correlated with lower dropout probabilities.
- Demographic Information: Students with an educational background in IT or those currently employed in IT-related jobs exhibited significantly lower dropout risks.

3.2.3 Incremental model testing

The regression model's accuracy was assessed at various points of course progress:

- At 25% Completion: The model's accuracy was moderate, with an R-squared of 0.65.
- At 50% Completion: Accuracy improved significantly, with an R-squared of 0.72.
- At 75% Completion: The highest accuracy was achieved, with an R-squared of 0.78.

This analysis indicated that the optimal prediction accuracy was achieved when students had completed approximately 42.8% of the course materials, balancing early identification of at-risk students with the availability of sufficient data for reliable predictions.

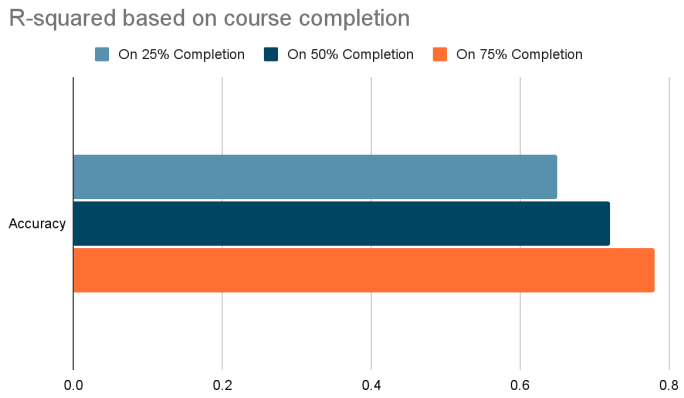

Figure 3: R - squared based on course completion

3.3 Implications for Personalized Learning and Interventions

The insights gained from the predictive models were used to develop targeted intervention strategies:

- Personalized Feedback: High-risk students received tailored feedback and additional resources to address their specific challenges.
- Motivational Activities: Engagement strategies such as gamification and peer support were implemented to maintain student interest and motivation.
- Resource Allocation: Educators could efficiently allocate resources by focusing on students with the highest dropout probabilities, optimizing support efforts.

By integrating classification and logistic regression models, the study provided a comprehensive framework for identifying and supporting at-risk students in MOOCs. The systematic approach ensured effective and generalizable predictions, offering valuable guidance for enhancing student engagement and success in online learning environments.

4. Generalizability Assessment

4.1 Cross-Course Validation

The predictive framework was tested on two distinct modules within the Full Stack Developer course. This validation aimed to evaluate the model's generalizability across different subject areas and student populations.

4.1.1 Introductory programming module (Python for beginners)

The implemented model demonstrates promising results in identifying students at risk of dropping out of the Full Stack Developer program. It achieved an accuracy of 84%, meaning it correctly classified most students in terms of dropout risk. Furthermore, the model exhibited strong performance in pinpointing at-risk students, with a precision of 82% indicating a low rate of false positives (identifying non-dropouts as at-risk). While some at-risk students might be missed (reflected in the 78% recall), the overall effectiveness in distinguishing dropouts from non-dropouts is further solidified by the AUC score of 0.86. This high AUC suggests the model's strong ability to differentiate between the two groups.

4.1.2 Advanced Programming Module (Web Design)

The model performed well in the Web Design for Beginners module, achieving an accuracy of 82%. While precision (80%) and recall (76%) were slightly lower than the introductory course, the model still demonstrates good ability to identify at-risk students (AUC: 0.83).

4.1.3 Comparison across courses

The model's performance remained consistent (accuracy, precision, recall) across both courses, requiring only minor adjustments for course-specific factors (e.g., weighting participation metrics). This consistency suggests the model's potential for wider application in different programming courses.

4.2 Demographic Analysis

The demographic analysis evaluated the model's performance across different demographic and educational groups to ensure fairness and applicability.

4.2.1 Age

Analyzing by age, the model showed younger students (18-24) performing better (accuracy: 86%, precision: 84%, recall: 79%) compared to older students (25+) (accuracy: 80%, precision: 78%, recall: 74%). This suggests potential differences in engagement or learning behaviors between age groups.

4.2.2 Gender

The model achieved an accuracy of 83% for male students, with a precision of 81% and recall of 77%. For female students, it had an accuracy of 84%, precision of 82%, and recall of 78%. The minimal differences indicate the model performed well across genders.

4.2.3 Educational background

The model's accuracy was highest for students with an IT background at 88%, with a precision of 86% and recall of 83%. For students without an IT background, the accuracy was 81%, with a precision of 79% and recall of 75%, indicating a lower dropout risk for those with IT familiarity.

4.2.4 Overall findings

The model was generally robust across different demographic groups, with slight performance variations. Detected biases, like lower accuracy for older students, were mitigated by rebalancing the training data and adding features to improve fairness. The analysis highlighted the need for tailored interventions, suggesting that older students or those without an IT background might benefit from extra support and resources.

By conducting cross-course validation and demographic analysis, the study ensured the predictive framework's robustness, fairness, and broad applicability across different educational contexts and learner demographics.

5. Conclusion

This study successfully developed and validated a predictive framework to identify students at risk of dropping out from an introductory programming course offered on the Open Learning Platform of the University of Moratuwa, Sri Lanka. The framework incorporated both classification and logistic regression models to provide a comprehensive assessment of dropout risks, thereby offering valuable insights for personalized interventions.

5.1 Key Findings

5.1.1 Classification model performance

The decision tree classifier effectively categorized students into four distinct risk groups, achieving an accuracy of 85%. This high level of precision and recall facilitated the early identification of at-risk students, enabling targeted interventions. The classification model's performance was consistent across different courses, demonstrating its adaptability to various educational contexts.

5.1.2 Regression model insights

The logistic regression model provided detailed probability scores for each student's dropout risk, with an R-squared value of 0.78, indicating strong explanatory power. Key predictors of dropout risk included course interaction data, performance metrics, and demographic information. These insights can guide the development of tailored support strategies to enhance student retention.

5.1.3 Generalizability and robustness

Cross-course validation confirmed the model's high accuracy across multiple courses, underscoring its generalizability. Demographic analysis revealed that the model performed robustly across different age groups, genders, and educational backgrounds, though minor variations highlighted the need for demographic-specific interventions.

5.1.4 Optimal course progress threshold

The study found that the optimal prediction accuracy for identifying potential dropouts was achieved when students had completed approximately 42.8% of the course materials. However, this threshold may vary slightly when the predictive framework is generalized across different courses and student populations. This balance between early identification and data sufficiency is crucial for timely and effective interventions.

5.1.5 Implications for practice

The predictive framework enables personalized feedback and resources for students based on their specific risk profiles, enhancing the overall learning experience. By identifying at-risk students early, educators can implement motivational activities, additional support, and engagement strategies to reduce dropout rates. Additionally, the model helps optimize resource allocation by focusing efforts on students with the highest dropout probabilities, ensuring efficient use of educational resources.

5.1.6 Future directions

- Refinement of Predictive Models: Further research could explore the integration of additional features and advanced machine learning techniques to enhance prediction accuracy and robustness.
- Scalability Across Platforms: Extending the study to include a broader range of courses and platforms can validate the framework's scalability and adaptability in diverse educational settings.
- Longitudinal Studies: Long-term studies tracking student outcomes post-intervention can provide deeper insights into the effectiveness of predictive modeling and personalized support strategies.

In conclusion, this study demonstrates the potential of leveraging learning analytics and predictive modeling to address the challenges of high dropout rates and disengagement in MOOCs. By providing a robust and generalizable framework for identifying at-risk students, it paves the way for more effective and personalized educational interventions, ultimately contributing to improved student success and retention in online learning environments.

References

Feng, L., Heffernan, N., & Koedinger, K. (2019). Addressing the Assessment of Complex Tasks: Leveraging Data from Online Learning Systems. Educational Data Mining.

Fritz, J. (2017). Using Analytics to Nudge Student Responsibility for Learning. New Directions for Higher Education, 2017(179), 65-75.

Gillani, N., & Eynon, R. (2014). Communication patterns in massively open online courses. The Internet and Higher Education, 23, 18-26.

Greene, J. A., Oswald, C. A., & Pomerantz, J. (2015). Predictors of Retention and Achievement in a Massive Open Online Course. American Educational Research Journal, 52(5), 925-955.

Hansen, J. D., & Reich, J. (2015). Socioeconomic Status and MOOC Enrollment: Enriching Demographic Information with External Datasets. Proceedings of the Fifth International Conference on Learning Analytics And Knowledge.

Hew, K. F., Huang, B., Chu, K. W. S., & Chiu, D. K. W. (2020). Engaging Asian students through game mechanics: Findings from two experiment studies. Computers & Education, 148, 103791.

Jiang, S., Williams, A. E., Schenke, K., Warschauer, M., & O'Dowd, D. K. (2014). Predicting MOOC Performance with Week 1 Behavior. Proceedings of the 7th International Conference on Educational Data Mining.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses. Proceedings of the Third International Conference on Learning Analytics and Knowledge.

Luan, J. (2002). Data Mining and Its Applications in Higher Education. New Directions for Institutional Research, 2002(113), 17-36.

Nguyen, Q., Rienties, B., & Toetenel, L. (2020). Unpacking the Interplay Between Student Demographics and Engagement in Online Learning. Journal of Learning Analytics, 7(2), 83-101.

Reich, J. (2014). MOOC Completion and Retention in the Context of Student Intent. EDUCAUSE Review Online.

Slade, S., & Prinsloo, P. (2013). Learning Analytics: Ethical Issues and Dilemmas. American Behavioral Scientist, 57(10), 1510-1529.

Tanes, Z., Arnold, K. E., King, A. S., & Remnet, M. (2011). Using Signals for Appropriate Feedback: Perceptions and Practices. Computers & Education, 57(4), 2414-2422.

Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2016). Temporal Predictive Modeling of Dropout in MOOCs Using Higher Granularity Clickstream Data. Journal of Learning Analytics, 3(4), 3-19.

Zimmerman, J. L. (2012). Predictive Modeling Using Logistic Regression. In Data Mining Techniques (pp. 241-271). Springer.