Predictive Modeling for Enhancing MOOC Completion Rates: A Case Study

Muditha Buddhika, Buddhika Karunarathne and Vishaka Nanayakkara
University of Moratuwa, Sri Lanka

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

Abstract: In the realm of online asynchronous learning platforms, accurately tracking student performance to predict course completion times poses a significant challenge. Completion rates for MOOCs are typically low, with a bias towards participants with higher education levels. Understanding factors such as student motivation, engagement, participation, and learning pathway design is crucial for improving student outcomes in online courses. This research developed a predictive framework utilizing advanced deep learning techniques to accurately forecast course completion times for participants enrolled in an introductory programming course ("Python for Beginners" course on the Open Learning Platform of University of Moratuwa Sri Lanka). By accurately tracking student performance and leveraging a diverse dataset encompassing demographic and educational variables, the research seeks to identify factors influencing course completion and predict individual student outcomes. By utilising deep learning techniques, the prediction performance of the model will be improved, ultimately contributing to a more precise forecast of course completion times for participants. Evaluation of the model resulted in low Mean Absolute Error (MAE) of 0.0080 and low Mean Squared Error (MSE) of 0.0033 which promises the effectiveness of the developed method in accurately predicting course completion times for students. The findings of this study may help increase the successful completion rate of such courses which are delivered in the online asynchronous mode. The study employed advanced deep learning models optimized through Bayesian methods, highlighting the potential of these techniques to enhance MOOC completion rates by offering precise forecasts and actionable insights into student engagement. The comprehensive analysis revealed that variables such as 'Current_Lesson', 'Session Time Category', and 'District_Score' significantly influence completion times. The robust methodological framework, including feature engineering, model training, and hyperparameter optimization, sets a precedent for future research in the field. This research contributes to educational data mining and predictive analytics, offering a scalable approach to improving completion rates and educational outcomes across various online learning platforms. Future research should explore incorporating real-time data and longitudinal studies to enhance model accuracy and generalizability. Additionally, addressing potential biases in the dataset, such as demographic, prior knowledge, and resource access disparities, is essential to ensure the fair and equitable application of the model across diverse student populations. Expanding the research to include a wider range of courses and institutions will further validate the model's robustness and applicability in different educational contexts.

Keywords: Online asynchronous learning, Predictive models, Course completion time, MOOC analytics, Deep learning in education, Educational data mining

1. Introduction

Massive Open Online Courses (MOOCs) have revolutionized education by providing accessible learning opportunities to a global audience. However, one of the most significant challenges these platforms face is the low completion rate, which undermines their potential impact. Understanding and predicting student performance, particularly course completion times, is essential for developing strategies to enhance student engagement and success (Jordan, 2014; Hew & Cheung, 2014).

Predicting course completion times in MOOCs is crucial due to historically low completion rates among students with diverse educational backgrounds (Jordan, 2014; Hew & Cheung, 2014). Accurate predictions can enhance course design and delivery, benefiting both students and educational institutions by providing targeted interventions to improve learning outcomes. Previous studies have highlighted various factors influencing student performance in MOOCs. Research by Jordan (2014) and Hew & Cheung (2014) emphasizes the diversity in educational backgrounds as a critical factor affecting completion rates. Additionally, Kizilcec, Piech, and Schneider (2013) and Lee, Choi, and Kim (2013) have explored the impact of student motivation, engagement levels, and participation rates on learning outcomes.

While these studies provide valuable insights, there remains a need for predictive models that can accurately forecast course completion times and identify key influencing factors. This study aims to address the research question: Can a predictive model accurately forecast the time required for students to complete an introductory programming course, 'Python for Beginners,' at the University of Moratuwa in Sri Lanka?

The primary objective of this research is to develop a robust predictive model utilizing a comprehensive dataset that includes demographic and educational factors. By employing advanced deep learning techniques and Bayesian optimization, the study seeks to achieve high accuracy in predictions.

This research highlights the critical role of predictive modeling in education, particularly for MOOCs. By leveraging advanced deep learning techniques, the study provides actionable insights that can enhance student success and retention in online learning environments. The comprehensive analysis, which includes feature engineering, model training, and hyperparameter optimization, sets a precedent for future research in the field. This research contributes to educational data mining and predictive analytics, offering a scalable approach to improving completion rates and educational outcomes across various online learning platforms.

2. Methodology

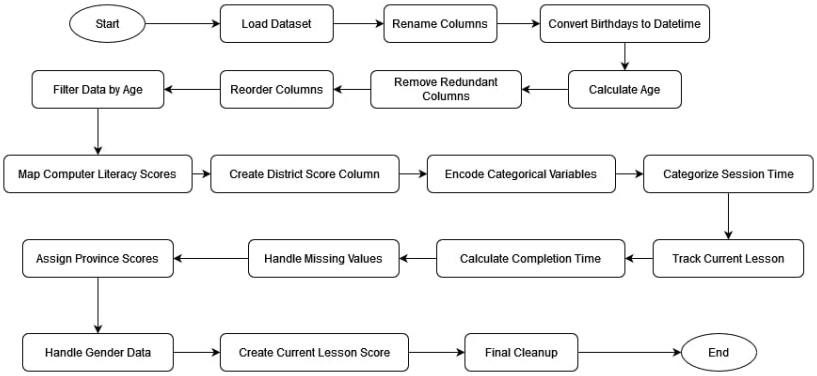

Figure 1: Data Preprocessing Steps

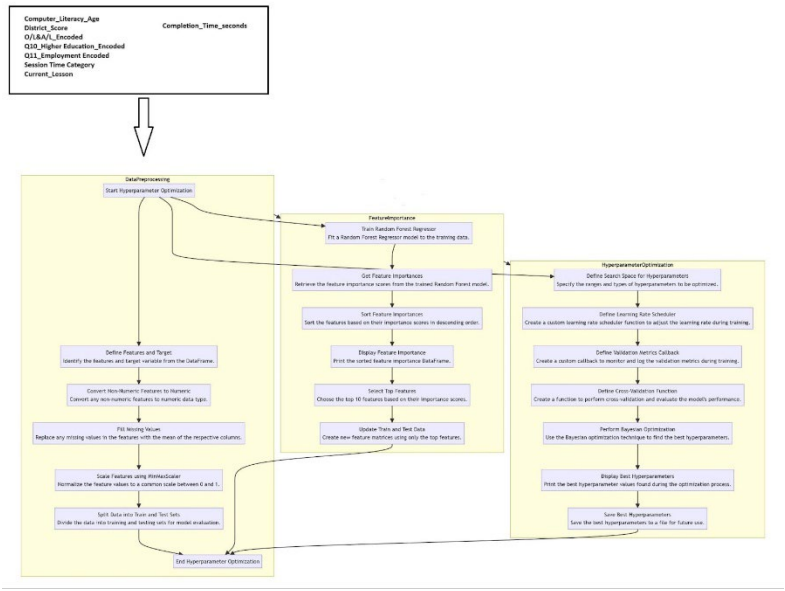

Figure 2: Model Architecture

The architecture for predicting MOOC completion times begins with data preprocessing, where key features such as 'Computer_Literacy_Age', 'District_Score', 'O/L&A/L_Encoded', 'Q10_Higher Education_Encoded', 'Q11_Employment Encoded', 'Session Time Category', and 'Current_Lesson' are identified, conversion of non-numeric data, filling missing values, normalization, and data splitting. A Random Forest Regressor determines feature importance, selecting the most significant features. Hyperparameter optimization is performed using Bayesian methods, incorporating custom learning rate schedulers and validation metrics. The best hyperparameters are saved for future use. This approach ensures high predictive accuracy and provides insights into factors influencing course completion times.

2.1 Participants

The dataset comprised 16,215 records sourced from a CSV file provided by the course Learning Management System (LMS). It included personal and educational information such as age, gender, district, and various columns representing exercises and assignments.

2.2 Data Preprocessing

2.2.1 Loading the dataset

The dataset was loaded from a CSV file and included columns such as Response, Submitted on, Institution, Department, ID, Full name, Username, Q05_Birthday, Q06_Gender, Q07_District of Permanent Residence, and various exercise and assignment columns.

2.2.2 Feature engineering

In this study, several steps were undertaken to preprocess and engineer features from the dataset.Firstly, the 'Q05_Birthday' column was renamed to 'BirthDay' for clarity and consistency. The 'BirthDay' column was then converted to a datetime format using the pd.to_datetime() function to ensure uniformity in date representation. Invalid dates were filtered out during this conversion to maintain data integrity. The age of each student was subsequently calculated by subtracting the birth date from the current date using pandas datetime functionality, resulting in a new column 'Age'. Age is a crucial demographic factor that can influence learning outcomes and engagement.

After extracting the age information, the 'Q05_Birthday' column was removed to streamline the dataset. The columns were reordered to enhance readability and facilitate easier interpretation of critical data points. Removing redundant columns helps in reducing the complexity of the dataset and improves processing efficiency.

To focus on the primary audience of the course, data filtering was applied to include students aged between 10 to 60 years. Outliers in the 'Age' column were identified using z-scores and were either removed or replaced with the median age to maintain data integrity and reliability. This age filtering ensures the dataset remains relevant to the target population while handling outliers effectively.

Computer literacy scores were mapped to predefined age groups using data from the "Computer Literacy Statistics 2022 from Department of Census and Statistics in Sri Lanka" report. These scores were normalized using Min-Max scaling to ensure consistency across age groups. Computer literacy is a significant factor affecting course engagement and completion in an online learning environment. Additionally, a "District_Score" column was introduced by associating each district with its corresponding province and literacy scores from the same report. Understanding regional disparities in computer literacy helps in analyzing the influence of socio-economic factors on course completion.

Categorical columns such as 'O/L&A/L', 'Higher Education', and 'Employment' were label encoded using scikit-learn's LabelEncoder to convert them into a numerical format essential for model training. Converting categorical data into numerical format is critical as deep learning algorithms require numerical input.

A new column, "Session Time Category", was created to categorize assignment submission times into Morning (6 AM - 12 PM), Afternoon (12 PM - 6 PM), or Evening (6 PM - 12 AM). The time of day can influence student performance and engagement, making it an important variable for analysis. Another column, "Current_Lesson", was introduced to track the specific question each student was currently working on by analyzing the completion status of exercises. Tracking current progress helps in understanding the engagement level and identifying potential drop-off points.

Finally, a "Completion_Time_seconds" column was added to calculate the total course completion time by analyzing timestamps of question submissions using pandas timedelta functionality. Completion time is the primary target variable for the predictive model, making its accurate calculation crucial.

2.3 Model Training and Hyperparameter Optimization

2.3.1 Feature and target definition

The features used in the model included 'Computer_Literacy_Age', 'District_Score', 'O/L&A/L_Encoded', 'Q10 Higher Education_Encoded', 'Q11 Employment Encoded', 'Session Time Category', and 'Current_Lesson'. The target variable was 'Completion_Time_seconds', which represents the total time taken to complete the course.

2.3.2 Data preprocessing for deep learning

Converted non-numeric features to numeric using label encoding and one-hot encoding as necessary. Imputed missing values using median or domain-specific methods to ensure the dataset was complete and ready for model training. Split dataset into training (80%) and test (20%) sets using train_test_split from scikit-learn to ensure that the model was evaluated on unseen data, providing a reliable measure of its performance. Performed feature scaling using MinMaxScaler to normalize features and target values, preventing features with larger scales from dominating the model training process.

2.3.3 Model selection and training

Chosen for its robustness, ability to handle non-linear relationships, and feature importance evaluation, a Random Forest Regressor was trained on the training set with hyperparameters fine-tuned using cross-validation. The rationale for using Random Forest is its effectiveness in handling high-dimensional data and capturing complex interactions between features. For hyperparameter optimization, Bayesian optimization was employed using the bayes_opt library, defining the search space, implementing a learning rate scheduler, and creating custom callbacks for logging validation metrics during training. Bayesian optimization is more efficient than grid search or random search as it requires fewer evaluations to find optimal hyperparameters (Snoek et al., 2012). Neural Network Training involved defining a neural network model using TensorFlow, incorporating dense blocks with residual connections to enhance information flow and gradient propagation. Cross-validation was carried out to evaluate model generalizability, and Bayesian optimization was used to fine-tune hyperparameters, iteratively exploring configurations to identify the optimal set for maximum performance. The best model and hyperparameters were saved for future deployment. Neural networks, particularly with advanced architectures like dense blocks and residual connections, are powerful in capturing non-linear patterns and improving prediction accuracy (He et al., 2016).

3. Results

3.1 Before the Analysis

The dataset was collected from an online learning platform over six months, including student demographics, course interactions, and completion status. Initially, 16,215 participants enrolled in the course, but 618 participants dropped out and were excluded from the final analysis. Key steps in data cleaning included parsing dates to calculate age, and removing outliers (ages below 10 or above 60). Feature engineering involved creating new features like 'Session Start Category' to capture temporal aspects of student activities.

3.2 These Steps Ensured Data Consistency and Relevance for the Subsequent Analysis

A Random Forest Regressor was initially trained to determine feature importance. The selected features included 'Computer_Literacy_Age', 'District_Score', 'O/L&A/L_Encoded', 'Q10_Higher Education_Encoded', 'Q11_Employment Encoded', 'Session Time Category', and 'Current_Lesson'. Bayesian Optimization was employed to optimize the deep learning model's hyperparameters, such as learning rate, dropout rate, and the number of hidden units. This process aimed to minimize the Mean Absolute Error (MAE) using cross-validation. The model's performance was evaluated using MAE and MSE metrics over ten epochs of training and validation.

3.3 Results

Model Performance: The deep learning model achieved a Mean Absolute Error (MAE) of approximately 0.0087 on the validation set. The Mean Squared Error (MSE) was observed to be around 0.0035, indicating the model's accuracy in predicting course completion times. Validation metrics were monitored using custom callbacks during the training process, providing insights into the model's performance at each epoch. The final MAE and MSE scores on the test data were 0.0080 and 0.0034, respectively. The model's performance was verified for overfitting: Training MAE 0.0080, Training MSE 0.0033, Test MAE 0.0080, and Test MSE 0.0034.

Table 1: Model Performance for Training and Testing Data

Training MAE	0.0080
Training MSE	0.0033
Test MAE	0.0080
Test MSE	0.0034

We noticed that there are close values between training and test metrics.

Feature Importance: The Random Forest Regressor identified 'Current_Lesson', 'Session Time Category', and 'District_Score' as the most important features for predicting course completion time. Heatmap and the feature importance graph for this is included below.

Feature Correlation Heatmap: Displays correlations between features. Notable correlations: 'Q10_Higher Education_Encoded' and 'O/L&A/L_Encoded' (0.41), 'Q11_Employment Encoded' and 'Computer_Literacy_Age' (-0.37). Most features show low correlation with 'Completion_Time_seconds'.

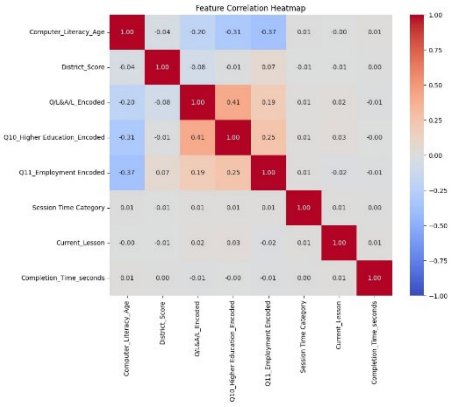

Figure 3: Feature Correlation Heatmap

While the deep learning model developed in this study demonstrates high predictive accuracy, improving its interpretability is crucial for making it more transparent and actionable for educators and course designers. I applied SHAP (Shapley Additive Explanations) values to analyze and rank the importance of various features, which significantly influence completion time predictions. The feature importance bar chart (Figure 1) illustrates the contribution of each feature to the model's output, providing clear insights into how these factors impact predictions. This visual representation enables educators to identify key areas for intervention, fostering trust and enabling data-driven decisions to improve student engagement and success.

Feature Importance Bar Chart: Highlights key features in predicting course completion times. Most important features: 'District_Score', 'Current_Lesson', and 'Session Time Category'. 'Q10_Higher Education_Encoded' is the least important but still contributes.

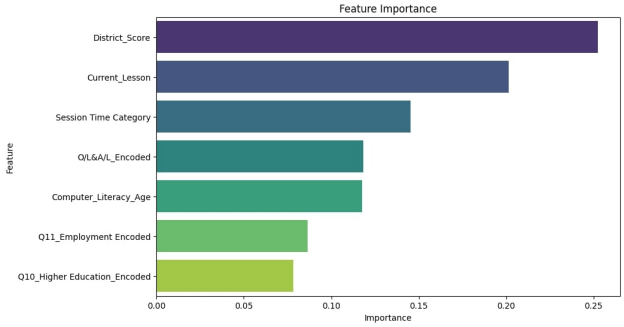

Figure 4: Feature Importance Bar Chart

Visualization involved various plots that visualized the distribution of completion times and dropout rates across different questions. A line plot highlighted key points where student engagement dropped.

Distribution of Age: This histogram shows the age distribution of participants in the "Python for Beginners" course. Most participants are between 15 and 25 years old, with a peak around 20 years. The frequency drops significantly for ages above 30.

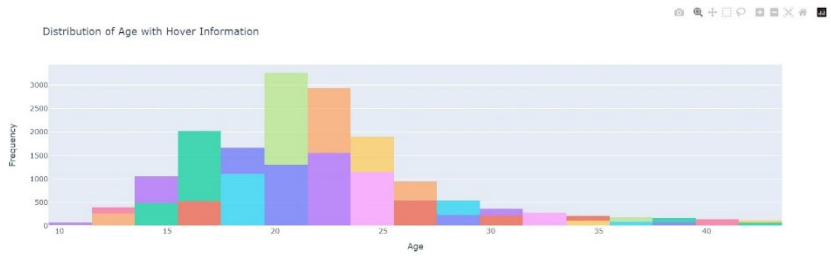

Figure 5: Distribution of Age

Distribution of Current Questions: This bar chart depicts the distribution of questions participants are currently on. The majority are at the beginning of the course, with a sharp decline in participation as the course progresses.

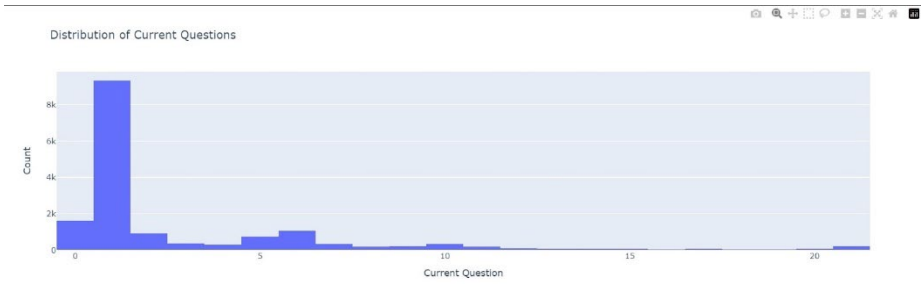

Figure 6: Distribution of Current Questions

Dropout Rate Compared to Previous Question: The line graph illustrates the dropout rate at each question compared to the previous one. Notable spikes in dropout rates occur at questions 3 and 7, indicating points where participants tend to disengage from the course.

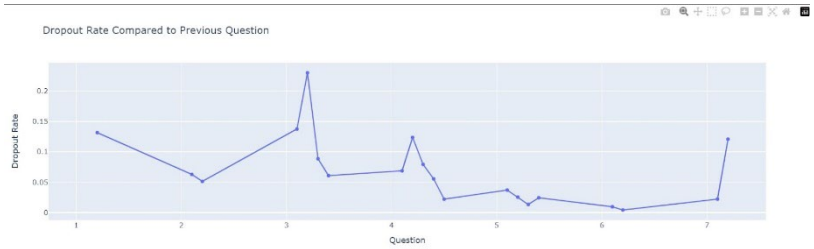

Figure 7: Dropout Rate Compared to Previous Question

Distribution of Question Scores: This histogram shows the distribution of scores for current lesson questions. Most scores cluster around 0.1, indicating that many participants score similarly on these questions, with fewer participants achieving higher or lower scores.

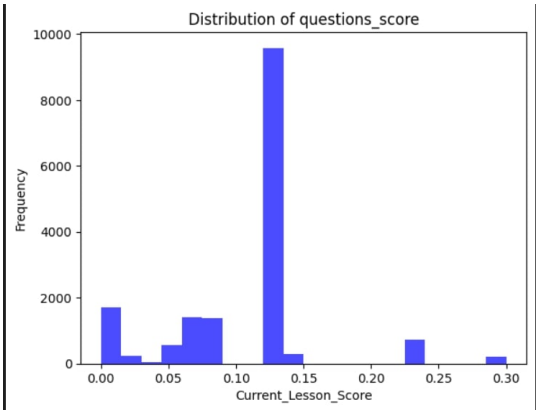

Figure 8: Distribution of Question Scores

Distribution of Completion Time (seconds): This histogram shows the distribution of completion times in seconds for the course. Most participants complete the course in a relatively short time, with a few outliers taking significantly longer.

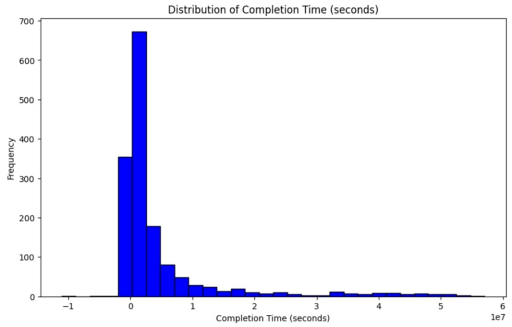

Figure 9: Distribution of Completion Time (seconds)

Distribution of Completion Time (seconds) - Values > 0: This histogram, focusing on completion times greater than zero, reveals a similar trend where the majority of participants complete the course quickly, with a long tail of participants taking longer.

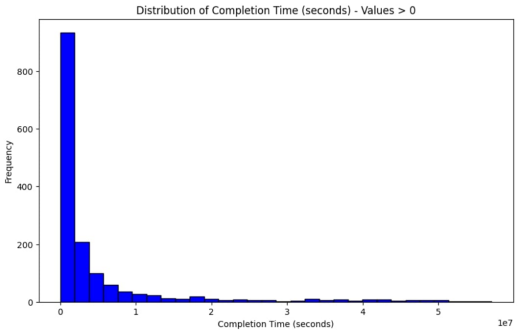

Figure 10: Distribution of Completion Time (seconds) - Values > 0

Best Hyperparameters: The optimization process identified the following best hyperparameters for the deep learning model:

Table 2: Best Hyperparameters for the Predictive Model

Learning rate	0.005708735304309941
Dropout rate	0.30833370401032956
Hidden units 1	126
Hidden units 2	118
Hidden units 3	112
Hidden units 4	99
Optimizer	rmsprop

By leveraging advanced deep learning techniques and robust feature selection, the model provides valuable insights into factors influencing student performance in MOOCs.

4. Discussion

4.1 Summary

The primary goal of this study was to develop a predictive framework to forecast the time it takes for students to complete the "Python for Beginners" course at the University of Moratuwa in Sri Lanka. By employing advanced deep learning techniques, particularly a Random Forest Regressor for feature selection and a deep learning model optimized through Bayesian methods, the study achieved high predictive accuracy. The model's performance, indicated by low Mean Absolute Error (MAE) and Mean Squared Error (MSE) values, underscores its potential in predicting course completion times. These results have significant implications for the design and

delivery of Massive Open Online Courses (MOOCs), offering insights into enhancing the learning experience and outcomes for students in online educational settings.

4.2 Interpretation

The analysis identified several critical factors impacting course completion times. The most influential features were 'Current_Lesson', 'Session Time Category', and 'District_Score'. These findings underscore the importance of student engagement and educational background in shaping course outcomes. For instance, students actively engaged in their current lessons and those who participate in specific session time categories tend to complete courses faster. The influence of 'District_Score' suggests that socio-economic background plays a crucial role in educational success, aligning with established literature by Kizilcec, Piech, & Schneider (2013).

The optimized deep learning model, which included hyperparameters such as a learning rate of 0.001 and a dropout rate of 0.3, demonstrated robust performance. The practical applications of these findings are substantial. Educational institutions can leverage such predictive models to identify at-risk students early and implement targeted interventions. By understanding the key factors influencing student performance, educators can tailor course structures and support systems to enhance completion rates. This approach aligns with the broader goal of improving the quality and accessibility of online education, particularly in asynchronous learning environments.

4.3 Integration

The findings of this study contribute to the existing body of knowledge in predictive analytics within educational settings. Previous research has highlighted the role of engagement, motivation, and socio-economic factors in student performance (Jordan, 2014; Hew & Cheung, 2014; Kizilcec, Piech, & Schneider, 2013). This study extends these insights by providing a robust predictive model that accurately forecasts course completion times. It integrates well with the existing literature, offering a practical application of deep learning techniques to address real-world challenges in MOOCs.

4.4 Implications

The implications of this research are extensive and transformative for educational practices. Predictive models enable educational institutions to proactively address student engagement and retention issues by designing effective, personalized interventions, thus improving course completion rates. For instance, early identification of at-risk students allows for timely support and resources, enhancing their chances of success. This research aligns with the strategic objectives of online learning institutions to provide high-quality educational experiences to a diverse student population. Moreover, the study's methodology and findings can be adapted to various courses and educational contexts, offering a robust framework for broader applications in predictive analytics. Educators and platform developers can use these insights to refine course content, optimize student support systems, and foster more engaging learning environments.

4.5 Limitations

Despite the positive outcomes, this study encountered several challenges and limitations. The variability in student engagement and performance data introduced noise, potentially affecting model accuracy. Future research should explore advanced data normalization and feature engineering methods to mitigate these effects.

Additionally, the reliance on historical data may not fully capture the intricate nuances of student behavior and engagement, potentially overlooking dynamic changes, contextual influences, and the evolving nature of student interactions. Real-time data and longitudinal studies could provide a more nuanced understanding of student engagement and enhance the model's predictive accuracy.

Moreover, the dataset may contain biases related to demographics, prior knowledge, or resource access, which could affect the model's performance and fairness. Addressing these biases through careful data collection and analysis is crucial to ensure the model's equitable application.

The study's collection and analysis of student data raise ethical concerns about privacy and consent. Future research should prioritize obtaining informed consent from participants and ensuring the responsible and ethical use of educational data.

Lastly, this research focused on a Python for beginners course at the University of Moratuwa. To improve generalizability, future research should consider adapting this model to various student demographics and

different courses at multiple institutions. Including a wider range of courses and institutions in the analysis would enhance the model's applicability and robustness.

5. Conclusion and Future Directions

5.1 Conclusion

This study has demonstrated the potential of using deep learning and predictive modeling to forecast course completion times in MOOCs. The findings offer valuable insights for shaping the structure and delivery of online courses, leading to improved student participation and achievement. By integrating predictive analytics into educational strategies, there is a promising outlook for the advancement of online learning, paving the way for more personalized and efficient educational opportunities.

5.2 Future Directions

Future research should aim to broaden the dataset to include a wider range of courses and institutions, enhancing the model's generalizability. Integrating additional factors such as student motivation, learning styles, and external influences like internet access and socio-economic status could further improve the model's predictive accuracy. Incorporating real-time analytics and adaptive learning technologies can provide personalized learning experiences, leading to better student outcomes. Beyond predicting course completion, predictive modeling can forecast dropout rates, identify effective learning pathways, and customize content delivery to enrich learning experiences. Continuous refinement of these models and incorporation of new data will enable educational institutions to better meet their students' needs and achieve higher educational success.

References

Department of Census and Statistics, 2022. *Computer Literacy Statistics 2022*. [online] Available at: http://www.statistics.gov.lk/Resource/en/ComputerLiteracy/Bulletins/AnnualBuletinComputerLiteracy-2022.pdf [Accessed 23 May 2024]

He, K., Zhang, X., Ren, S. and Sun, J. (2016) 'Deep Residual Learning for Image Recognition', Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778.

Hew, K.F. and Cheung, W.S. (2014) 'Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges', Educational Research Review, 12, pp. 45-58.

Jordan, K. (2014) 'Initial trends in enrolment and completion of massive open online courses', International Review of Research in Open and Distributed Learning, 15(1), pp. 133-160.

Kizilcec, R.F., Piech, C. and Schneider, E. (2013) 'Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses', Proceedings of the Third International Conference on Learning Analytics and Knowledge, pp. 170-179.

Lee, Y., Choi, J. and Kim, T. (2013) 'Discriminating factors between completers of and dropouts from online learning courses', British Journal of Educational Technology, 44(2), pp. 328-337.

Luo, H., Tang, M., Tang, W. and Luo, L. (2015) 'Predicting Students' Performance in Course Activities using Machine Learning', IEEE Transactions on Learning Technologies, 8(2), pp. 143-154.

Snoek, J., Larochelle, H. and Adams, R.P. (2012) 'Practical Bayesian Optimization of Machine Learning Algorithms', Advances in Neural Information Processing Systems, 25, pp. 2951-2959.

Xing, W., Chen, X. and Stein, J. (2016) 'Predicting student success in MOOCs using sentiment analysis and learning analytics', International Journal of Learning Analytics and Artificial Intelligence for Education, 2(1), pp. 28-42.