Predicting Student Dropout in University Courses Using Machine Learning Models

< Zhenqiang Tang >

Supervised by: < Felipe Campelo >

Abstract: Predicting student dropout at an early point is beneficial to institutions who want to take action to intervene. Student retention continues to be a hot topic in higher education; it impacts students' futures as well as institutions' success as a whole. This project centers on developing a predictive model for students who withdraw/drop out of university courses. This model will use historical data, provided by the Open University Learning Analytics dataset, and extract the variables that relate to students' characteristics, preferences, and their behavior on the virtual learning environment (VLE) such as frequency of logging in, submissions of assignments, and administrative interactions related to these variables. The dataset has many details on the interactions students made while studying course content, specifically their frequency of VLE access, their assessment performance, their demographic data which includes age, gender, and previous study achievement information. Aspects of the study want to determine which factors contribute most to students dropping out and build an accurate classifier that will determine if a student withdraws or drops out of a course. Additionally, we want to determine whether student dropout rates decreases by interventions initiated based on particular risk factors. We plan to use several machine learning algorithms including Logistic Regression, Random Forest, and XGBoost to predict the possibility of student dropout. The expectation is that the outcomes from the projects will build a better understanding of student retention for university settings, create some level of scalable intervention that targets student-specific needs. The model that we build can potentially be developed into a predictive system that can be employed in university contexts. This will support educators and administrators in identifying students at risk of withdrawal and dropping out before the students disengage. Ultimately, improving retention rates, supporting student experiences, and resource allocation in meaningful ways for academic institutions will be an important priority.

Ethics statement: This project:

- This project fits within the scope of ethics pre-approval process, as reviewed by my supervisor Felipe Campelo and approved by the faculty ethics committee as application 15208 < **Felipe Campelo**>.

I have completed the ethics test on Blackboard. My score is <12>/12.

Project plan:

In the past two decades, student retention has emerged as one of the biggest challenges faced by universities globally. With an increasing focus on improving educational outcomes and financial sustainability, understanding why students drop out and designing interventions to decrease dropout became of more importance. The ability to identify students at risk of dropping out at an early stage provides educational institutions with the timely ability to implement the appropriate interventions which can eventually improve student success and decrease attrition rates. This project focuses on predicting university course dropout using machine learning techniques and the Open University Learning Analytics dataset which provides rich detail about a students demographics, their academic journey, and the level of engagement a student has through the virtual learning environment (VLE).

The motivation for this project comes from the heightened need to improve student retention in higher education settings. Evidence suggests that many factors, such as academic performance, engagement with course materials, demographic variables and prior educational history, influence student dropout **(Del Bonifro et al., 2020)**. Knowing the many individual factors contributing to dropout enables universities to take proactive, early action to offer the correct academic support, resources, counselling and services to improve retention, and ultimately decrease dropout rates. Machine learning models for predicting student dropout rates, have increased in popularity, with positive results in many educational contexts **(Prenkaj et al., 2020)**. These studies acknowledge the value of predictive models to identify students who may be at risk of withdrawing from their course, before a decision to withdraw is made.

In fact I treat it as a binary classification problem to solve. So I decided to use classification method to encode, The proposed approach is to use sophisticated machine learning algorithms. this project will build a predictive model to identify which students are most likely to withdraw from their courses. This will be achieved using the Open University Learning Analytics dataset which contains a rich dataset of potential predictors; demographic details, assessment performance data and recorded behaviours via the VLE like, logging in, submitting assignments and engagement with course artefacts. The project will explore features and patterns that may foretell whether a student is going to dropout and classify which at-risk students are likely to withdraw.

The selected analytic method would include an explanation as to why those methods were chosen for the project. The benefits of machine learning methods will drive this aspect of the project. The applied algorithms will include logistic regression, random forests and XGBoost. Logistic regression, while very popular and useful as a rapid reference point given its simplicity, and capability of clear interpretation when predictive modelling in the past. Random Forests is an ensemble method and is useful as it has built-in protection against overfitting and is capable of identifying relationships between features, which may be very complex **(Doleck et al., 2020)**. XGBoost, the gradient boosting algorithm, is also gaining popularity as predicted high performance in machine learning competitions **(Rajni & Malaya, 2015)**, especially with structured/tabular data similar to the example used in this project.

Therefore, The planned method or methodology for the project will consist of a number of machine learning algorithms. In particular, a number of algorithms would be suitable to address the complex, manifold nature of the features in the dataset, including Logistic Regression, Random Forest and XGBoost. For the purposes of this analysis the three models will be compared to each other for performance. Logistic Regression is a widely accepted approach for binary classification tasks and as such will be our baseline model. Logistic Regression is a simple model, but easy to interpret and understand the inference of how predictor variables relate to the probability of dropout. Random Forest, an ensemble method deployed by captures complex structure in data and overfits less than other models.**(Kurni et al., 2023)** . XGBoost, an extension to boosting performance for continuous data, shown to be a powerful method for structured data (Models 1 and 2) and is able to deal with categorical and numerical features within the same dataset. The three algorithms were deployed because of the flexibility capacity, performance, and interpretability we had anticipated in the data.

An important consideration during these steps is also the data preprocessing and feature engineering to capture useful information from the dataset. Data cleaning procedures will address issues (i.e. missing values) in the training dataset, while normalizing numeric features, transforming categorical features in a manner that faitfully represents the categorical features for the use of machine learning algorithms (i.e. categorical features will utilize one-hot encoding features). Feature engineering will emphasize capturing the best predictors of dropout as raw features describing student behavior (i.e. demographic features - age, gender, prior academic achievement, engagement features - number of clicks from the VLE total, time doing course content, academic features - assignment score, number times attempts at a course). **(Márquez - Vera et al. 2016)** discussed student can be strongly classified early through engagement features (i.e. VLE) to suggest predicting students are at risk of dropout thus we saw importance to feature extraction in this project.

Once pre-processing is complete then the machine learning model will be trained using the training dataset, and evaluated on a separate validation dataset, ensuring evaluation is on data that has not been seen. Machine learning models commonly have multiple classification metrics to evaluate (e.g. accuracy, precision, recall, F1-score, AUC-ROC). The main classification metric for our project is recall, as fewest or minimizing false negatives of students classified as not dropping but do drop out later in the semester is most important in an early-system of students using early intervention systems. The feature importance analysis will provide a picture of likely factors that best related to dropout of students as described by **(Uskov et al., 2019)**, and we hope to created some insight of key risk factors that universities need to be aware and monitor.

In addition to building an accurate predictive model, a further objective of this project is to maintain model interpretability. This way, universities will have knowledge of the features that are most influential in the predictions of the model and have opportunities for sharing information with students and/or the allocation of resources/interventions with at-risk students in a more equitable way **(Bird et al. 2021)** ,indicate, the interpretability of higher education predictive models is a significant consideration in providing transparency to decision making and aids in ensuring the exploitability and fairness of the model's decisions. Therefore in this project we will strive to build not only a model that performs in the top percentiles, but also a functional artifact that real-life educators and administrators can use and benefit from in determining and recognizing at-risk students and supporting them in their academic lives.

We expect the findings of this study to provide meaningful insight into the dynamics of student retention. If we can predict which students are at risk of dropping out of their programs before they take action, it can provide universities an effective tool in more actively and efficiently engaging and supporting students, by appropriately targeting their interventions. Individually, what this research will also contribute to, is the greater body of work that is educational data mining and learning analytics, translating how predictive models can be deployed in order to improve education and reduce dropout rates, and how the introduction of machine learning (ML) techniques into educational institutions can improve student outcomes and increase inclusion and accessibility for higher education.

References:

[1]. Del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020). Student dropout prediction. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I 21* (pp. 129-140). Springer International Publishing.

[2]. Prenkaj, B., Velardi, P., Stilo, G., Distante, D., & Faralli, S. (2020). A survey of machine learning approaches for student dropout prediction in online courses. ACM Computing Surveys (CSUR), 53(3), 1-34.

[3]. Doleck, T., Lemay, D. J., Basnet, R. B., & Bazelais, P. (2020). Predictive analytics in education: a comparison of deep learning frameworks. *Education and Information Technologies*, *25*, 1951-1963.

[4]. Rajni, J., & Malaya, D. B. (2015). Predictive analytics in a higher education context. *IT Professional*, *17*(4), 24-33.

[5]. Kurni, M., Mohammed, M. S., & Srinivasa, K. G. (2023). Predictive analytics in education. In *A Beginner's Guide to Introduce Artificial Intelligence in Teaching and Learning* (pp. 55-81). Cham: Springer International Publishing.

[6]. Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, *33*(1), 107-124.

[7]. Uskov, V. L., Bakken, J. P., Byerly, A., & Shah, A. (2019, April). Machine learning-based predictive analytics of student academic performance in STEM education. In *2019 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1370-1376). IEEE.

[8]. Bird, K. A., Castleman, B. L., Mabel, Z., & Song, Y. (2021). Bringing transparency to predictive analytics: A systematic comparison of predictive modeling methods in higher education. *AERA Open*, *7*, 23328584211037630.

Appendix: Project Timeline

Week	**Tasks**
2-13/6	Familiarisation with project, definition of goals, requirements
16-27/6	Data loading and initial preprocessing (handling missing values, merging datasets) Feature engineering (creating student demographic, academic, and behavioral features)
30/6-11/7	Coding and evaluation: main goals， Model training (logistic regression, random forest, and XGBoost)

Week	Tasks
14-25/7	Coding and evaluation: secondary goals，Hyperparameter tuning and model evaluation，Model interpretation and explanation (SHAP, feature importance)
28/7-8/8	Final refinements and improvements
11/8-31/8	Final report writing and preparation for submission

Appendix: Risk Assessment

There are a number of risks associated with this project, all of which require consideration and mitigation plans. As was discussed previously, the major risk with this project is data quality issues, such as missing values, incorrect data formats, incomplete records, and so on. The fact that the project will rely on historical data from the Open University Learning Analytics dataset means that data quality issues will have significant effects on model accuracy. The likelihood of this risk occurring is moderate, because the dataset is large and covers a variety of factors. In order to mitigate the risk, data preprocessing steps will be conducted to clean and standardize the data set. Data preprocessing will include various steps such as imputation of missing values, removal of outliers, and reformatting data.

Another risk is that of model overfitting, whereby the machine learning models become overly complex and show very good performance metrics on the training set, yet fail to generalize and perform well on unseen data. The likelihood of this risk is medium likelihood due to the use of multiple models in this process, especially decision trees and ensemble methods. In order to mitigate this risk, we will use cross-validation, as well as hyper-parameter tuning, to allow for the models to be well-regularized and generalize well to new data.

A third risk is that of class imbalance, where the number of students who drop out will be significantly lower than the number of students who do not drop out, which will create bias in predicting this class. As this risk is likely high, hence lower dropout rates compared to total number of students. In order to mitigate this risk, we will try and apply techniques such as SMOTE (Synthetic Minority Over-sampling Technique), as well as change relevant class weights when training,

Finally there could be a technical risk arising from the potential complexity of integrating and deploying the model into a real-time system. The likelihood of this risk will be low likelihood as we can provide a limited deployment for the scope of this project. However, this may be an issue in the last stages of the project and could raise issues for subsequent previous project. In order to mitigate this risk we will do simple prototypes and local deployments first and not try to deploy before integrating the potential complexity.