TITLE: **Predicting Student Dropout Using Machine Learning**

NAME: **Joe Achira**

DATE: **18/02/2025**

1. Introduction

1.1 Purpose of the Study

Study Group operates several International Study Centres across the UK and Dublin in partnership with universities, aiming to prepare a diverse pipeline of international students for degree study. These centres support students in adapting to the academic, cultural, and social aspects of studying abroad by offering personalised learning paths and flexible scheduling options to accommodate various learning styles and commitments.

1.2 Problem Statement

Study Group faces challenges with student dropout, as some students fail to complete their courses and progress to their chosen universities. This issue affects student retention, financial stability, revenue generation, institutional reputation, and overall student satisfaction.

The goal of this study is to employ and compare multiple predictive algorithms, such as XGBoost and a neural network-based model, to determine the most effective approach for predicting student dropout.

The data is divided into three stages:

- **Stage 1:** Applicant and course information (e.g. age, course details).
- **Stage 2:** Student engagement data (e.g. attendance records).
- **Stage 3:** Student academic performance (e.g. modules assessed, passed, and failed).

2. Methodology

2.1 Data Preprocessing and Exploratory Data Analysis (EDA)

Stage 1: Applicant and Course Information

Data preprocessing involved removing uninformative features, handling missing values, addressing high-cardinality variables, and applying appropriate encoding techniques.

EDA included analysing descriptive statistics and visualising histograms and distribution analysis.

The age variable required scaling as it was skewed and on a different scale than other features. Outliers were identified in the age distribution with a maximum age of 62.

The target variable (CompletedCourse) was imbalanced, with a majority of students completing their courses (85%) and a minority dropping out (15%). Class weights were applied to mitigate bias in model predictions.

3. Model Development and Performance Evaluation

3.1 Stage 1: Applicant and Course Information

Two models were trained:

1. XGBoost
2. Neural Network

Stage 1 Model Performance (Before Hyperparameter Tuning)

Model	Accuracy	AUC Score
XGBoost	69%	0.7151
Neural Network	68%	0.7642

Hyperparameter tuning was performed, yielding minor changes but not significantly altering performance.

Model	Accuracy	AUC Score
XGBoost (Tuned)	68%	0.7173
Neural Network (Tuned)	68%	0.7623

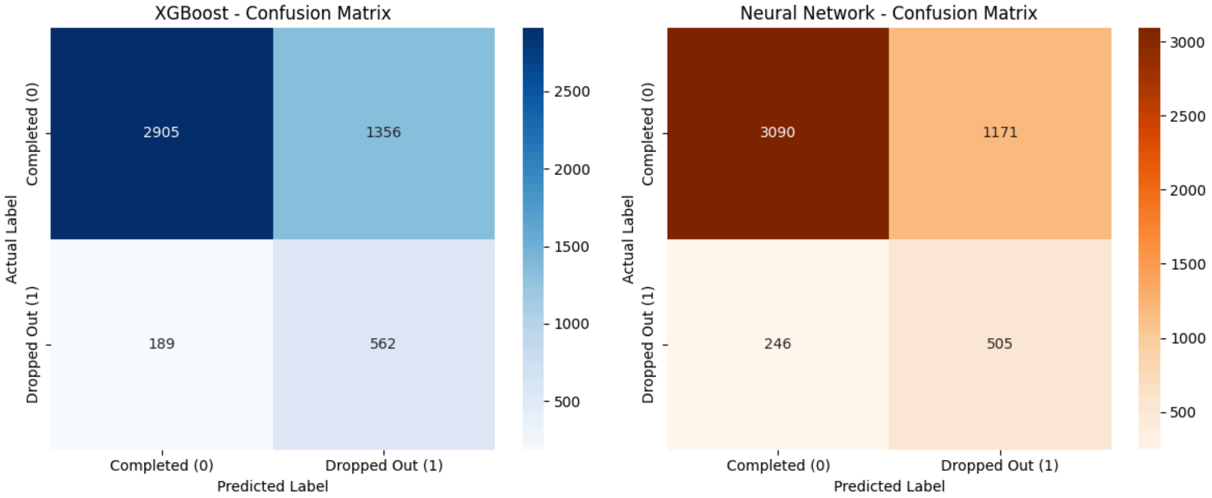
Key Observations

XGBoost and the neural network performed similarly.

Precision for the minority class (dropouts) was low (29%), highlighting difficulty in correctly identifying at-risk students.

The addition of class weights improved recall for dropouts (75%) but did not significantly enhance overall model performance. This was driven by a high number of false positives as seen on the heatmap below (Fig 1).

Fig 1 - Confusion Matrix HeatMap (XGBoost and Neural Network)



3.2 Stage 2: Student Engagement Data

New features included:

- AuthorisedAbsenceCount
- UnauthorisedAbsenceCount

These variables provided insights into student engagement and attendance patterns, which are strong indicators of commitment, discipline, and potential academic challenges.

Stage 2 Model Performance

Model	Accuracy	AUC Score
XGBoost	81%	0.7896
Neural Network	78%	0.856

Once again, hyperparameter tuning resulted in minor changes that did not significantly alter performance.

Model	Accuracy	AUC Score
XGBoost (Tuned)	80%	0.8045
Neural Network (Tuned)	79%	0.8565

Key Observations

Accuracy and AUC improved for both models, indicating that engagement data was useful in predicting dropout (Fig 2).

Precision for dropouts remained low but showed improvement.

Recall for dropouts remained high, suggesting the models were capturing at-risk students more effectively.

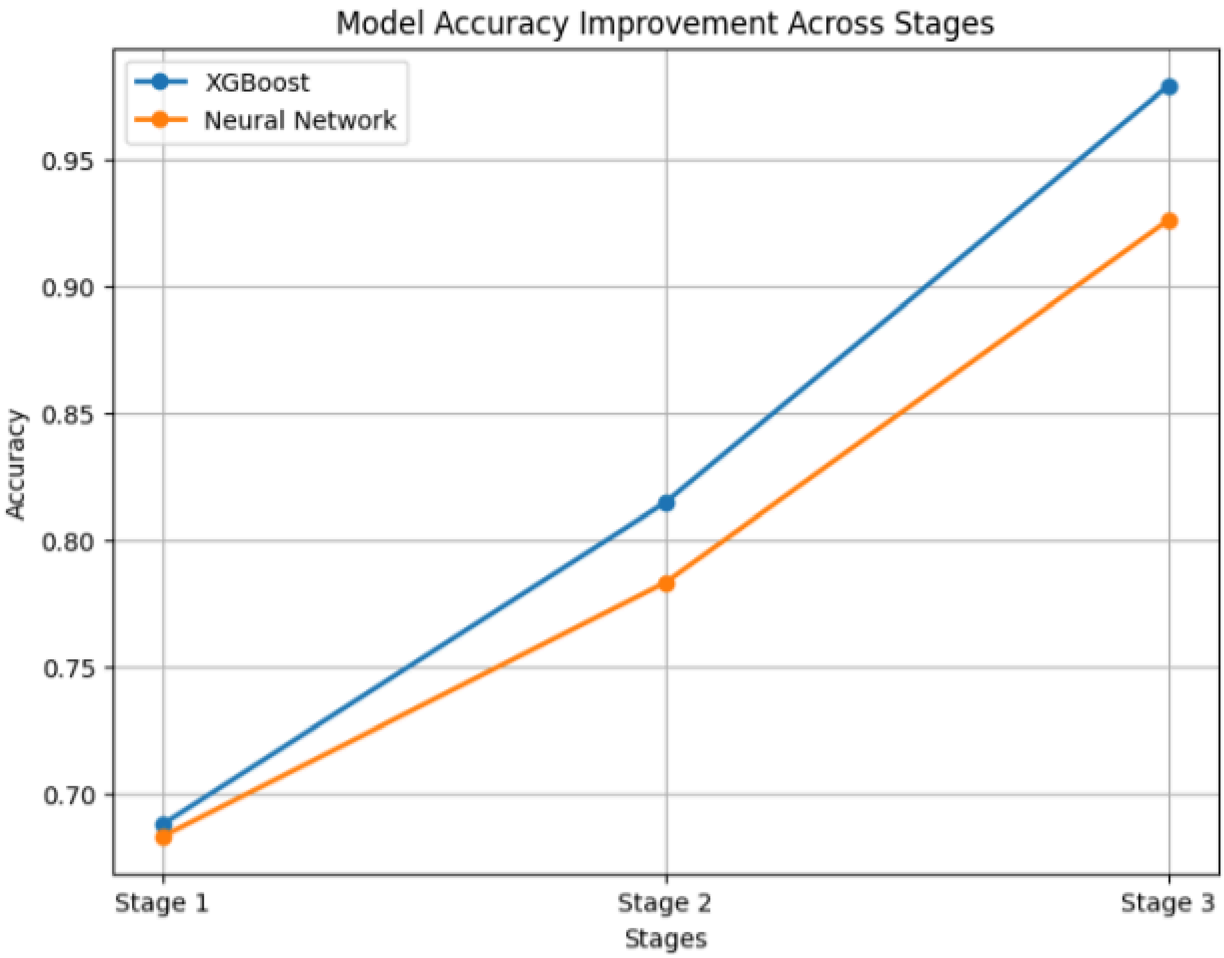


Fig 2 - Model accuracies improved across all 3 stages

3.3 Stage 3: Academic Performance Data

New features included:

- Number of modules assessed
- Modules passed
- Modules failed

These features were expected to provide a clearer picture of student progress and allow for better dropout prediction.

Stage 3 Model Performance

Model	Accuracy	AUC Score
XGBoost	0.9788	0.9629
Neural Network	0.9259	0.9926

Key Observations

Significant performance improvement in both models.

High recall for dropouts, meaning at-risk students were effectively identified.

Increase in area and the curve (AUC) scores has improved at each stage (Fig 3), signalling an improvement in the model's discriminative power between classes.

Precision for the dropout class remained a concern, as false positives increased.

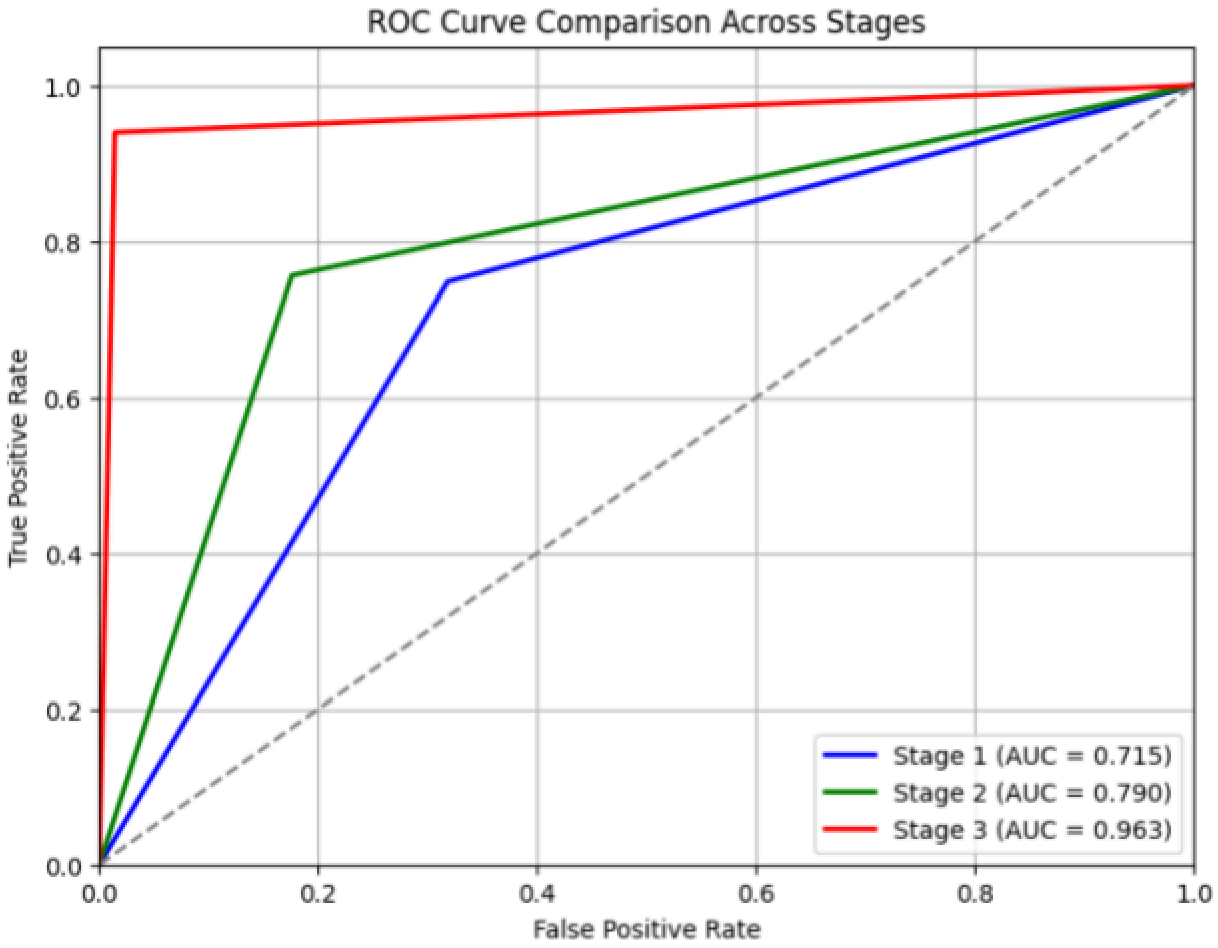


Fig 3 - Increasing AUC signals improvement in model's discriminative power.

Feature Importance Analysis (SHAP Interpretability)

Applying SHAP on the XGBoost model shows that the most important features impacting dropout are academic performance (late-stage), followed by engagement features (mid-course) and applicant information (early-stage).

Key Insights (Corresponds to Fig 4):

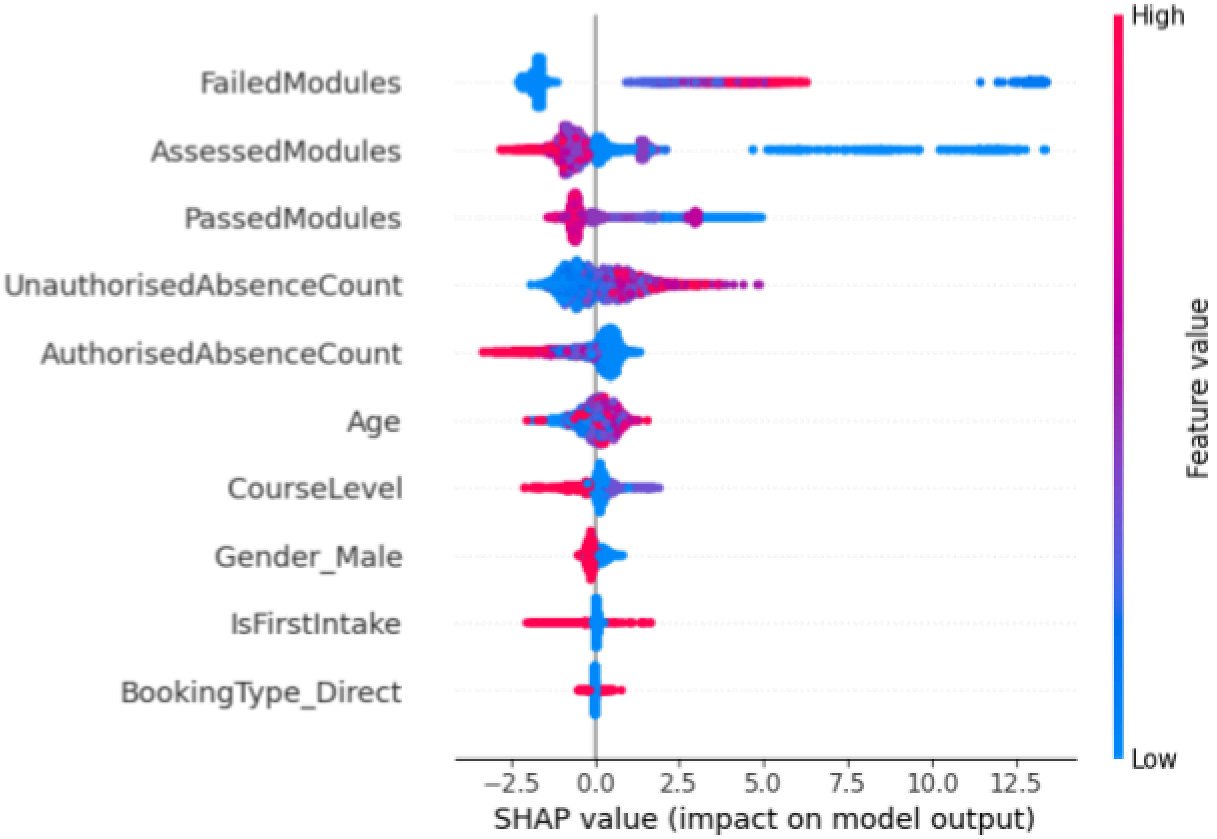
Failed modules and **assessed modules** significantly impact dropout risk:

- A **high number of failed modules** increases the risk of dropout.
- A **higher number of assessed modules** reduces dropout risk, indicating student commitment and ambition.

High numbers of unauthorised absences increase dropout risk.

Higher numbers of authorised absences reduce dropout risk, possibly indicating discipline and purpose-driven absences.

Fig 4 - SHAP Feature importance/value plot



These insights reinforce the importance of academic and engagement monitoring in predicting dropout and guiding intervention strategies.

4. Recommendations and Conclusion

4.1 Summary

Stage 1 models were limited in predictive power, as applicant information alone was insufficient to distinguish dropout risk effectively.

Stage 2 models showed improved performance, demonstrating that engagement features (e.g., attendance) play a crucial role in predicting dropout.

Stage 3 models showed **exceptional predictive power**, confirming that academic performance (failed/passed modules) is the strongest indicator of dropout risk.

4.2 Recommendations:

Early Intervention for At-Risk Students: Since academic performance is the most critical factor, early intervention should focus on students struggling with modules.

Attendance Monitoring: Unauthorised absences strongly predict dropout, so attendance tracking should be used to identify at-risk students.

Further Model Optimisation: Additional hyperparameter tuning, feature engineering, and alternative modelling approaches (e.g., ensemble methods) could further enhance predictive performance.

4.3 Conclusion

This study demonstrated the feasibility of using machine learning models to predict student dropout risk at different stages of the student journey. The inclusion of engagement and academic performance data significantly improved predictive power. Future work should focus on refining features, optimising model performance, and integrating these predictions into actionable strategies to support student retention and success.

By leveraging these models and insights, Study Group can better understand dropout risks and develop targeted interventions to improve student retention.