Applying Supervised Learning to Predict Student Dropout

Stakeholder Report

MAY 2025

PERFORMED BY

LAUREN BRIXEY

Table of Contents

1. Problem Statement & Objectives

Student dropout is a significant challenge for educational institutions, often resulting in lost revenue and reputational damage. It is critical to accurately predict which students are at risk for dropping out to allow for proactive intervention. The objective of this project is to develop a machine learning model that can predict student dropout based on early stage to late stage risks, with the primary focus on minimising false negatives, and maintaining balance of other metrics such as F1-score, precision and ROC-AUC. This will be achieved through the use of classification techniques (XGBoost and Neural Networks), class imbalance handling strategies and performance evaluation with recall as the primary metric.

2. Methodology

2.1. Data Overview

The prediction task is approached through three distinct datasets which represent the availability of student data throughout their academic journey and provides various risk identification points:

1. **Stage 1 (Early Stage):** Limited data available at the start of the academic journey, contains applicant and course information.
2. **Stage 2 (Mid Stage):** More information available midway through academic journey, contains student engagement data (‘AuthorisedAbsenceCount’, ‘UnauthorisedAbsenceCount’)
3. **Stage 3 (Late Stage):** Most informative dataset, contains academic performance data (‘AssessedModules’, ‘‘FailedModules’, ‘PassedModules’)

2.2. Data Preprocessing and Feature Engineering

To prepare the dataset for modelling, features with >50% missing values were dropped and median imputation was used to fill missing values of <20%. High cardinality categorical variables (>200 unique values) were dropped prior to one-hot encoding and label encoding the nominal and ordinal variables respectively. For input into the neural network models, numerical features were standardised using StandardScaler(). A new age feature was engineered using the ‘DateofBirth’ feature and the target ‘CourseCompleted’ feature was transformed to ‘DroppedOut’ and mapped appropriately to binary 1 (dropped out) or 0 (did not drop out) values to support intuitive interpretation of results.

2.3. Modelling Approach

To model student dropout, two machine learning models were explored: **XGboost** and a **neural network**. The XGBoost models were built using the XGBClassifier() instance from the XGBoost library and the baseline models were implemented with the default parameters (***Table 1***). To address the inherent class imbalance, the scale_pos_weight parameter was applied, calculated as the ratio of negative to positive cases.

The neural network models were constructed using the Sequential API from Keras. The baseline network consisted of an input layer tailored dynamically to the number of input features, two hidden layers containing 128 and 64 neurons respectively using the ReLU activation function, and an output layer containing a single neuron and a sigmoid activation function. The model was compiled using the binary cross-entropy loss function and the adam optimizer. To address the class imbalance, class weights were applied during training, along with early stopping and l2 regularisation to mitigate overfitting.

2.4. Hyperparameter Tuning

To optimise model performance and maximise recall, hyperparameter tuning was performed for both the XGBoost model and Neural Network trained on each stage's dataset. Hyperparameter tuning ranges for the XGBoost models were tested using the grid search function (GridSearchCV): learning_rate = [0.01, 0.05, 0.1], max_depth = [3, 5, 7] and n_estimators = [200, 500, 700] (**Table 1**).

Table 1: Hyperparameter values for **XGBoost** model for the baseline model (default) and optimised values applied in the tuned models.

Model	learning_rate	max_depth	n_estimators
Baseline	0.3	6	100
Tuned (Stage 1)	0.1	3	200
Tuned (Stage 2)	0.05	3	500
Tuned (Stage 3)	0.1	3	200

Hyperparameter ranges for the Neural Network models were tested using a nested for loop: optimizer= ['adam', 'rmsprop'], activation=['relu', 'tanh'], neuron_layer_1= [128, 64, 32], neuron_layer_2= [64, 32, 16] and batch_size= [32, 64, 128] (**Table 2**).

Table 2: Hyperparameter values for **Neural Network** model for the baseline model and optimised values applied in the tuned models.

Model	optimizer	activation	neuron_layer_1	neuron_layer_2	batch_size
Baseline	adam	relu	128	64	32
Tuned (Stage 1)	rmsprop	tanh	128	16	64
Tuned (Stage 2)	rmsprop	tanh	32	16	128
Tuned (Stage 3)	rmsprop	relu	64	64	64

3. Results & Interpretations

The model development process across Stages 1 to 3 shows cleared improvement in predictive performance for both XGBoost and Neural Network, driven primarily by the quality of the training features rather than hyperparameter tuning (**Figures 1-4**). In Stage 1, models relied on pre-enrollment data like demographics, resulting in modest recall (~0.77) and poor overall balance, highlighting the limits of early detection using indirect predictors. Stage 2 included student engagement features, which notably improved recall, though tuning only offered minor gains and often reduced precision. Stage 3 introduced academic performance data, the strongest predictor of student dropout, which led to the highest recall (up to 94.7% with XGBoost) and best overall performance. As in earlier stages, tuning slightly improved recall but lowered precision. Overall, while hyperparameter tuning can be useful, strong predictors are the most critical factor in building an effective and reliable dropout prediction model. While early stage indicators can provide a reasonable indication of student drop out, mid-late stage indicators provide the most accurate and actionable predictions of student dropout.

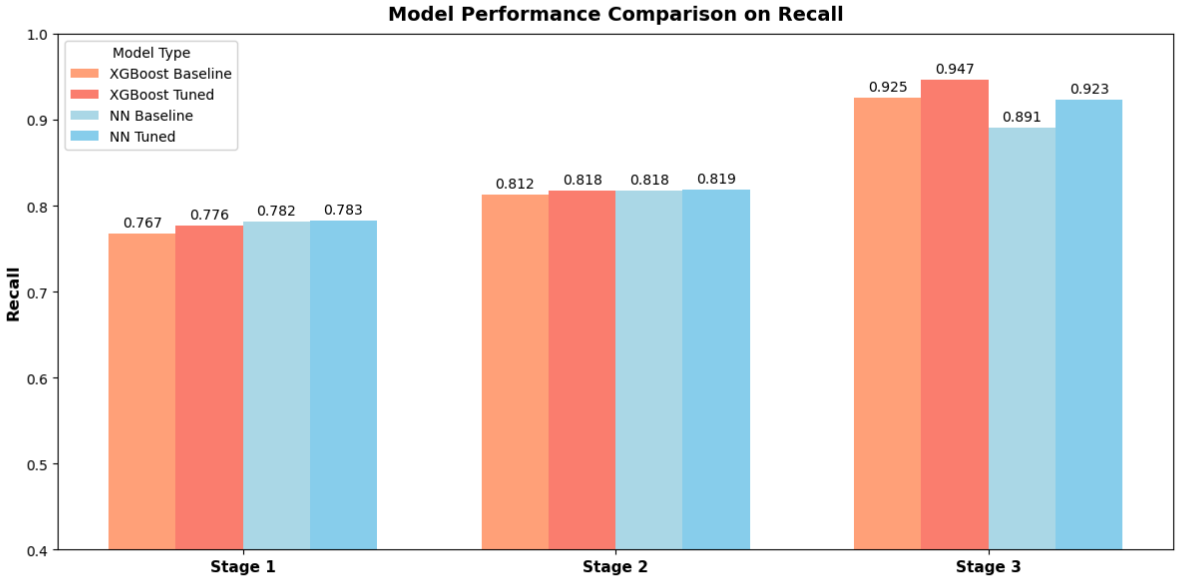

Figure 1: Comparison of model recall across Stages 1, 2, and 3 for baseline and tuned versions of XGBoost and Neural Network models.

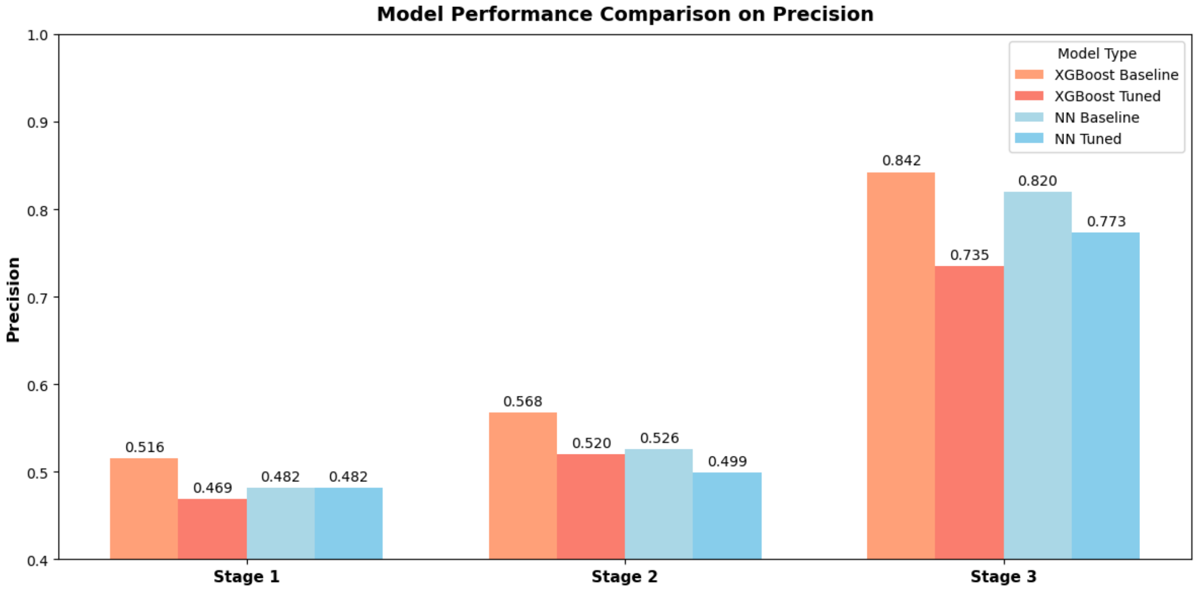

Figure 2: Comparison of model precision across Stages 1, 2, and 3 for baseline and tuned versions of XGBoost and Neural Network models.

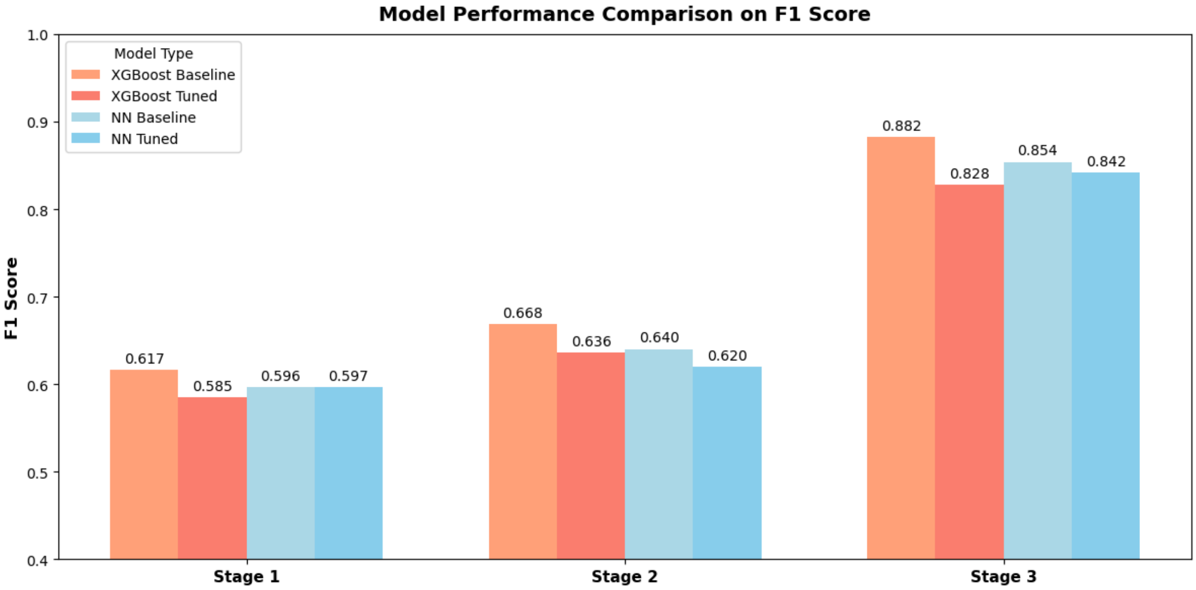

Figure 3: Comparison of model F1 Score across Stages 1, 2, and 3 for baseline and tuned versions of XGBoost and Neural Network models.

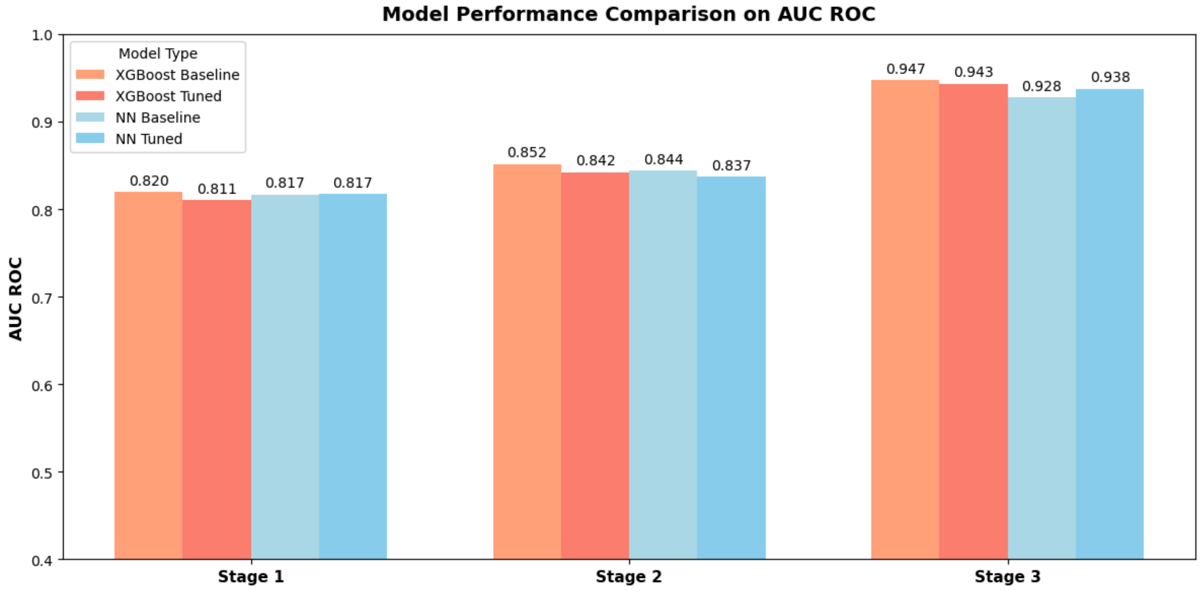

Figure 4: Comparison of model AUC ROC across Stages 1, 2, and 3 for baseline and tuned versions of XGBoost and Neural Network models.

XGBoost Tuned Model (trained on Stage 3 data) achieved the highest overall performance, with a recall of 0.925, precision of 0.842, and an F1 score of 0.882, demonstrating its effectiveness in accurately identifying at-risk students while maintaining balance and reliability. The confusion matrix (**Figure 5**) indicates it only missed 56 dropout cases and correctly identified 695.

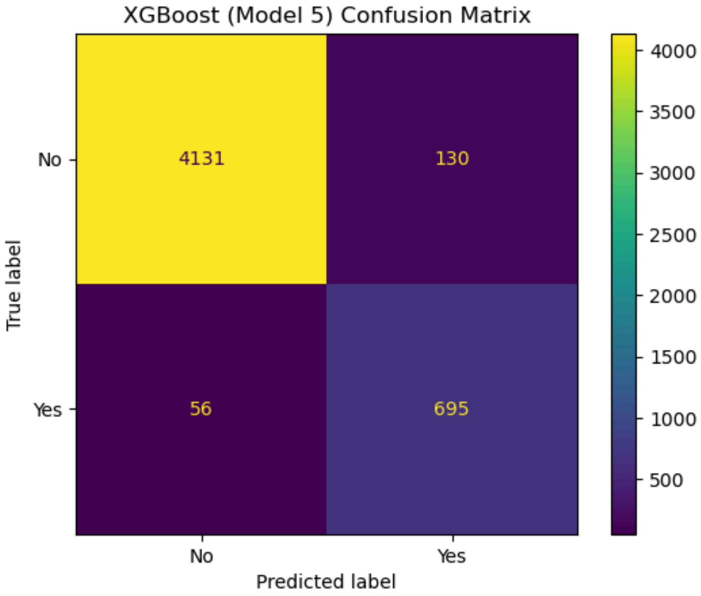

Figure 5: Confusion matrix for the best performing model (XGBoost Tuned Model, trained on Stage 3 data), illustrating the number of true positives, true negatives, false positives and false negatives in dropout prediction.

4. Conclusions & Recommendations

A variety of XGBoost and Neural Networks models were developed to predict student dropout risk using early, mid and late dropout predictors, with recall prioritised to minimise false negatives. Results indicate that model performance improved most significantly with stronger feature predictors (Stage 3 data), with the highest recall (and highest overall performance metrics) achieved. Overall, the best performing model was XGboost Model 5, which achieved a strong balance of recall (0.925) and precision (0.842). Its robust performance across all metrics make it a strong candidate for deployment in a production setting for supporting timely student interventions to dropout.