Applying supervised learning to predict student dropout

Contents

Problem Statement

Student dropout is a major challenge for higher education, affecting revenue, reputation, and outcomes. Study Group aims to identify at-risk students early to support timely intervention.

This project uses supervised learning to predict dropout using real student data, simulating the staged availability of features: applicant details, engagement metrics, and academic performance.

Summary of Approach

The modelling followed a three-stage framework, reflecting the staged availability of student data. Each stage introduced new features to assess how additional information affected predictive performance.

Stage 1 used applicant and enrolment data (e.g. nationality, start date, age). After preprocessing, both XGBoost and a neural network were trained and evaluated.

Stage 2 added engagement data such as attendance and contact hours, using the same modelling process.

Stage 3 introduced academic results and average grades. XGBoost was tuned using grid search; the neural network kept a fixed architecture.

Models were evaluated using accuracy, precision, recall, AUC, and confusion matrices. Training behaviour was assessed using loss, accuracy, and AUC curves.

Stage 3 Model Evaluation

Stage 3 includes the richest dataset, offering the best chance for dropout prediction. This section evaluates the tuned XGBoost and Neural Network models trained on it.

XGBoost Performance

Tuning was performed using grid search over learning_rate, max_depth, and n_estimators. The best configuration, selected based on cross-validation accuracy (0.9750), is shown in Table 1.

Parameter	Value
learning_rate	0.3
max_depth	8
n_estimators	100

Table 1: Optimal hyperparameter values for XGBoost after grid search tuning

Tuning delivered modest gains, particularly in precision and false negatives. Both models performed strongly, as shown in Table 2 and Figure 1.

Metric / Component	Untuned	Tuned
Accuracy	0.9739	0.976
Precision	0.9807	0.982
Recall	0.9887	0.99
AUC	0.9931	0.993
True Positives (Completed)	4213	4217
True Negatives (Dropped out)	668	672
False Positives	83	79
False Negatives	48	44

Table 2: XGBoost Performance summary (tuned vs untuned)

The confusion matrices below confirm high classification accuracy across both classes:

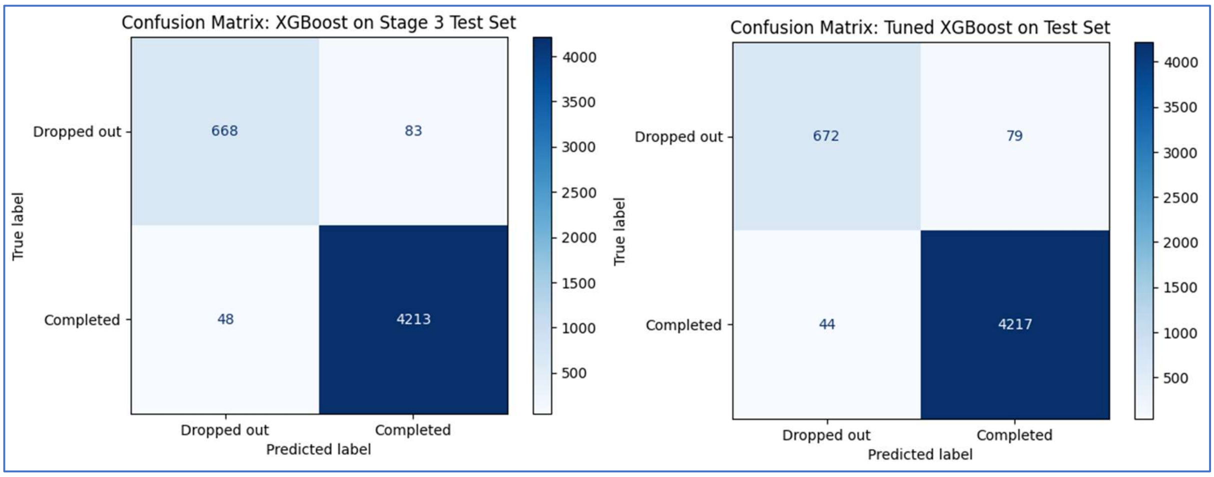

Figure 1: Confusion matrices before and after tuning the XGBoost model on Stage 3

Neural Network Performance

The neural network used a fixed architecture across all stages: two hidden layers (64 and 32 units, ReLU), a sigmoid output for binary classification, Adam optimiser (learning rate 0.001), and binary crossentropy loss. The model was trained for 20 epochs. Results are shown in Table 3 and Figure 2.

Metric	Value
Accuracy	0.9619
Precision	0.9746
Recall	0.9808
AUC	0.9745

Table 3: Performance metrics for the Neural Network model on Stage 3

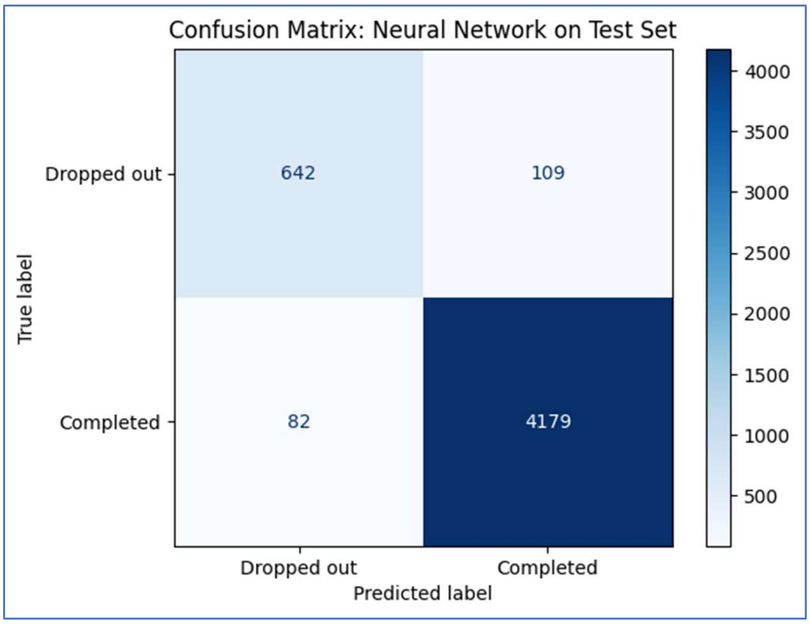

Figure 2: Confusion matrix – Neural Network (Stage 3)

Training curves, displayed in Figure 3, showed strong initial learning but early signs of overfitting:

- Loss: Training loss declined steadily, but validation loss plateaued after ~4 epochs and began to fluctuate upward — a classic sign of overfitting.
- Accuracy: Training accuracy reached ~97%, while validation accuracy peaked near 96% and then levelled off from epoch 7.
- AUC: Training AUC surpassed 0.99, while validation AUC plateaued around 0.97 from epoch 6 onward.

These patterns suggest that regularisation may improve generalisation.

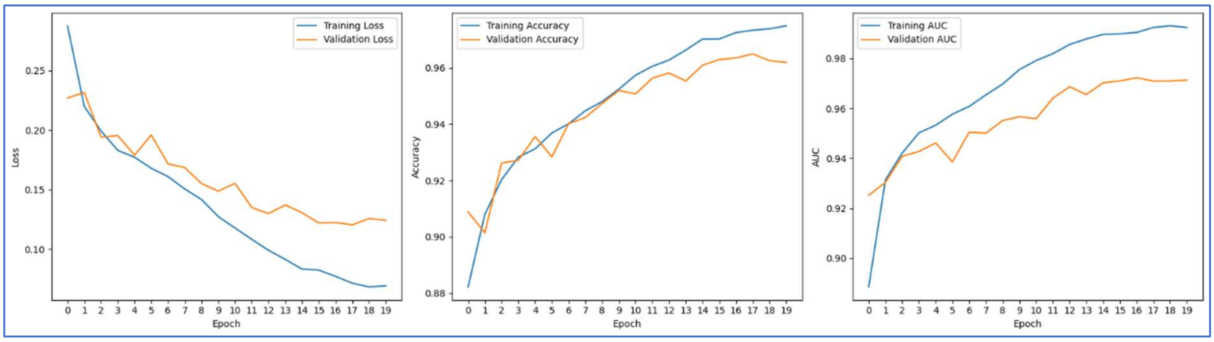

Figure 3: Training and validation curves for the Neural Network on Stage 3

Model Comparison: XGBoost vs Neural Network

Despite strong performance, Table 4 shows the Neural Network has more false positives and false negatives than XGBoost, confirming the latter as the better choice.

	XGBoost (tuned)	Neural Network
True Positives (correctly predicted completion)	4217	4179
True Negatives (correctly predicted dropout)	672	642
False Positives (predicted completed, actually dropped out)	79	109
False Negatives (predicted dropped out, actually completed)	44	82

Table 4: *Confusion matrix components for both models on the Stage 3 test set*

Stage Comparison

This section compares how model performance evolved from Stage 1 to Stage 3 and when dropout risk can be reliably detected.

Table 5 shows how XGBoost outperforms the neural network at every stage, especially on AUC. Both models demonstrate stable recall and strong classification accuracy. This suggests the presence of early predictive signals, particularly around enrolment features.

Stage 3 delivers the strongest performance, with XGBoost reaching 0.9755 accuracy and 0.9930 AUC. Richer academic data provides clearer signals of student outcomes. False positives dropped to 79 and false negatives to 44 using XGBoost (see Table 4).

Model	Stage	Accuracy	Precision	Recall	AUC
XGBoost (tuned)	1	0.8953	0.9207	0.9594	0.8802
	2	0.9038	0.9269	0.9635	0.9082
	3	0.9755	0.9816	0.9897	0.993
Neural Network	1	0.8957	0.9259	0.9535	0.8658
	2	0.899	0.9245	0.9602	0.8934
	3	0.9619	0.9746	0.9808	0.9745

Table 5: Classification metrics for both models across all stages

Performance improves at each stage, but the jump from Stage 1 to Stage 2 is modest — most gains come at Stage 3. This suggests that while early models offer a reasonable baseline, the most actionable predictions occur later in the course, once academic patterns emerge. However, by that point, intervention windows may have narrowed.

XGBoost shows particular robustness at all stages, making it the better candidate for early-stage deployment — though with caution due to weaker recall on dropouts in Stage 1. The neural network performs reliably but shows signs of overfitting and limited adaptation across feature sets.

Conclusion

This project evaluated XGBoost and a neural network across three datasets. XGBoost consistently outperformed the neural network, with higher AUC and precision at every stage.

Model performance improved steadily with richer data. Stage 3, containing academic results and attendance, produced the highest scores (e.g. XGBoost AUC: 0.9930), confirming the strength of Stage 3 features.

Earlier-stage models performed well. Stage 1 models reached AUCs near 0.88–0.89, indicating that dropout signals are present even at enrolment. However, prediction is more uncertain at this point, with a higher risk of false classifications.

Neural networks showed signs of overfitting, especially in earlier stages (based on training curves not shown here). Validation gains plateaued early, suggesting that regularisation (e.g. dropout or early stopping) could improve generalisation.

In summary, XGBoost is the more reliable model across all stages.