MAY, 2025

Student Dropout Predictions

PRESENTED BY

Onyeka Muoka

Table of Content

Introduction

Keeping students enrolled is a key challenge for education providers like Study Group. Dropout happens when many students begin a course but don't finish it. This not only affects the provider's finances and reputation but also impacts student success. Early dropouts can lead to long-term setbacks for both the students and the wider community. If providers can identify students at risk of dropping out before it happens, they can take action to reduce financial losses and give students the support they need to stay on track.

This project focuses on developing a student dropout prediction model for Study Group using supervised machine learning techniques. It uses demographic, engagement, and academic performance data provided at different stages of the student journey to identify patterns linked to dropout risk. By uncovering key risk factors, the model helps Study Group take early, targeted actions to support at-risk students and improve retention. The results offer data-driven insights to guide educators and administrators in making informed decisions that strengthen student support systems. This project highlights the practical use of machine learning in education, offering scalable tools for improving outcomes across institutions.

Problem Statement

Student dropout is a major challenge for Study Group, with many students starting courses but failing to complete them. This affects the organization's financial performance, damages its reputation, and limits student success. The goal of this project is to build a predictive model using supervised machine learning to identify students at risk of dropping out early, enabling timely interventions to improve retention and outcomes.

Data Description

The dataset consists of three distinct stages, each progressively incorporating additional features to improve predictive modeling of student dropout risk. Stage 1 includes demographic and course-related information, such as nationality, gender, course name, and academic level. This data allows institutions to identify early indicators of dropout risk before enrollment. Stage 2 expands upon Stage 1 by introducing student engagement metrics, including authorized and unauthorized absence counts, enabling analysis of how attendance correlates with academic success. Stage 3 incorporates academic performance variables, such as the number of assessed modules, passed modules, and failed modules, providing the most direct indicators of student progress.

To prepare the data for modeling, preprocessing steps were applied at each stage. Irrelevant identifier columns were removed, categorical variables were encoded using ordinal and one-hot encoding, and columns with excessive missing values (over 50%) were dropped. For Stage 2, rows with missing values were removed due to their minimal impact on the dataset (less than 2% of total entries). In Stage 3, missing numerical values were imputed using the median instead of being discarded to preserve data integrity. These preprocessing techniques ensured a clean and well-structured dataset, optimizing model accuracy.

Methodology

This study adopted a data driven methodology to predict student dropout rates at different academic stages using supervised machine learning techniques. Two primary models were employed: XGBoost (Extreme Gradient Boosting) and Neural Networks, each selected for their distinct strengths in handling classification tasks. The methodology encompassed data preprocessing, model training, hyperparameter optimization, and rigorous evaluation to ensure reliable predictions.

The XGBoost algorithm was implemented due to its efficiency in handling structured data and providing interpretable feature importance scores. Key hyperparameters, including learning rate, maximum tree depth, and number of estimators, were systematically tuned to enhance predictive performance. Additionally, the model's feature importance rankings were visualized to identify the most influential factors affecting student dropout at the different stages.

To complement this approach, We also built a neural network with one input layer, two hidden layers, and one output layer, and added dropout regularization to prevent overfitting. The architecture was optimized through experimentation with different neuron configurations, activation functions, and optimizers. The learning rate was carefully adjusted to balance training speed and model convergence.

Performance was assessed using classification metrics: accuracy, precision, recall, F1-score, and AUC-ROC. Confusion matrices and classification reports provided detailed insights into prediction errors, while ROC curves compared the models' ability to distinguish between dropout and non-dropout cases.

Feature engineering played a crucial role in model performance. The XGBoost model's built-in feature importance analysis helped identify key predictors, ensuring optimal input variables. To validate model robustness, the dataset was split into training (80%) and testing (20%) sets. For the Neural Network, the training data was further split to include a validation data for the model. Training and validation loss curves were monitored to detect overfitting, with early stopping implemented if performance plateaued. This comprehensive methodology ensured that both models were rigorously trained, evaluated, and optimized for accurate student dropout prediction while maintaining interpretability and generalizability.
Limitations in the study included limited time constraints and limited computing power.

Results and Analysis

The predictive models were evaluated across the three stages of data availability to assess how different data types impact performance.

Stage 1

In the first stage, both the XGBoost and Neural Network models achieved relatively low AUC scores. This was primarily due to their heavy reliance on demographic features such as nationality and center name, which offered limited predictive power.

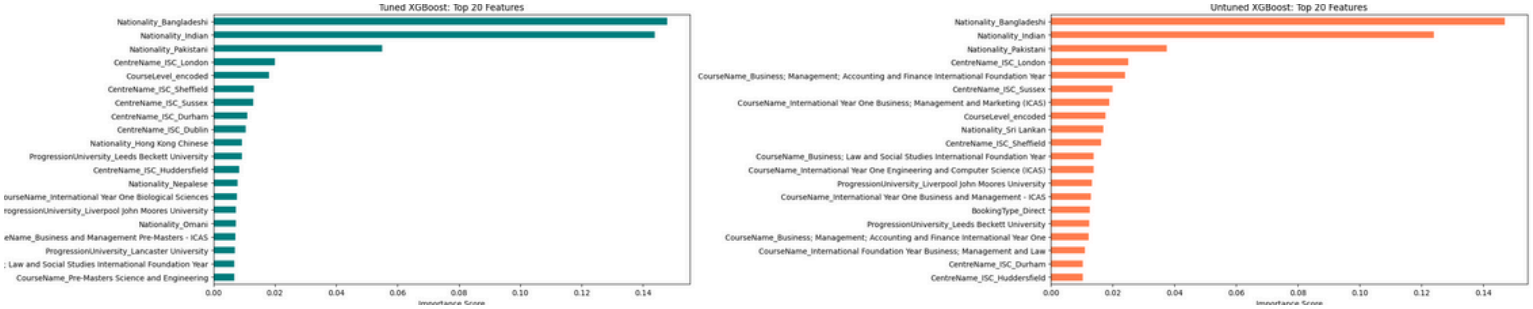


Figure 1. Feature Importance Untuned and Tuned XGBoost (Stage 1)

Additionally, the models faced challenges in accurately identifying dropout cases because of class imbalance in the dataset. While hyperparameter tuning improved predictions for the dropout class, it came at a cost of reducing the models' overall accuracy, particularly less accuracy classifying students who successfully completed their courses.

Metric	XGboost		Neural Network	
	Untuned	Tuned	Untuned	Tuned
Accuracy	0.89	0.85	0.88	0.86
Precision	0.69	0.50	0.67	0.54
Recall	0.53	0.77	0.49	0.72
F1-Score	0.60	0.61	0.56	0.62
AUC-ROC	0.74	0.81	0.72	0.80

Table 1. Classification metrics for all Stage 1 models

Based on the ROC curves, the tuned XGBoost model delivers the best performance with an AUC of 0.8886, showing the most ability to distinguish between classes amongst all the stage one models.

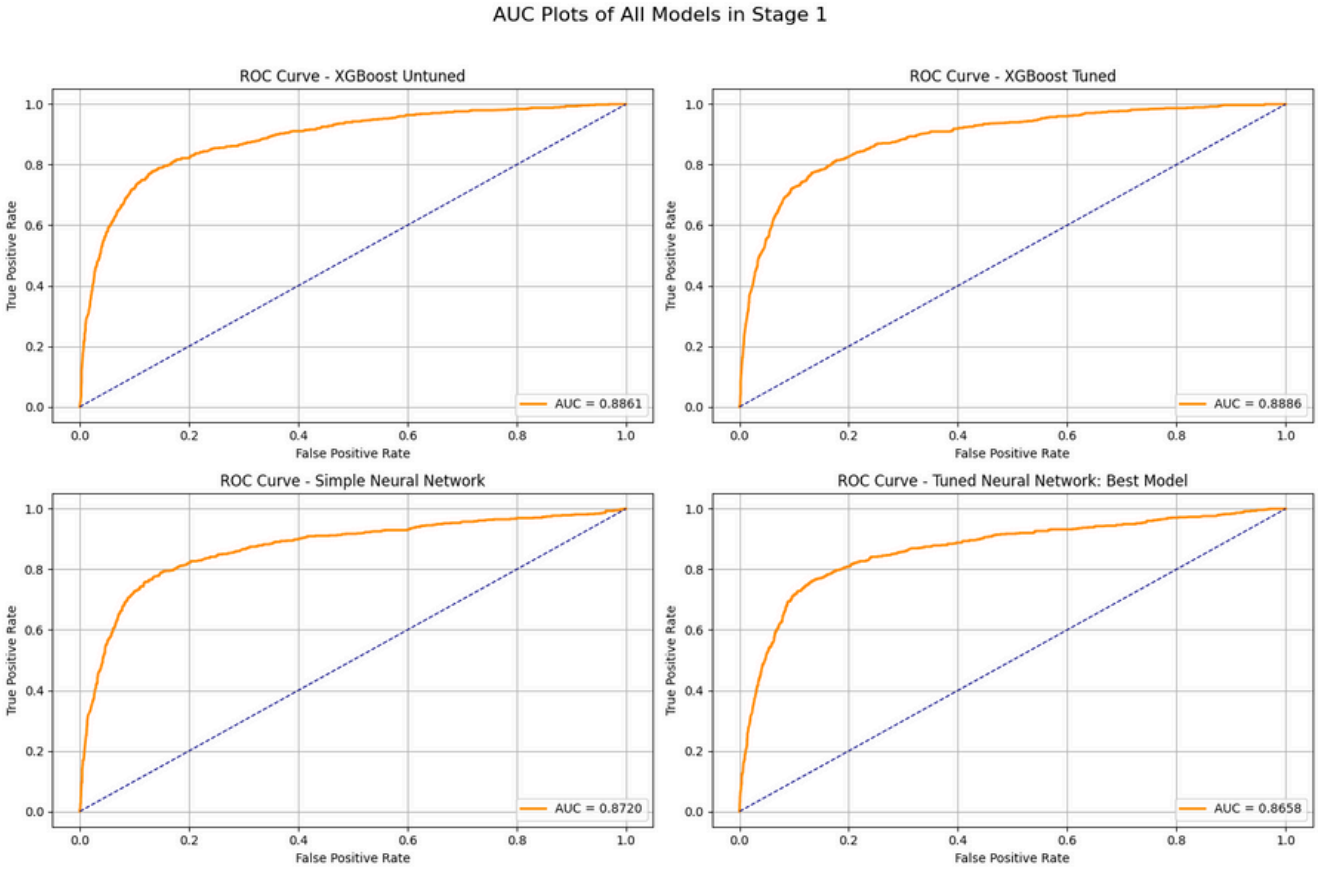


Figure 2. ROC Curve for all Stage 1 models

Stage 2

Engagement metrics, played a crucial role in improving the stage 2 models' ability to identify at-risk students compared to Stage 1 models. Notably, unauthorized absence count ranked among the top 20 most influential features in the performance of the Stage 2 model.

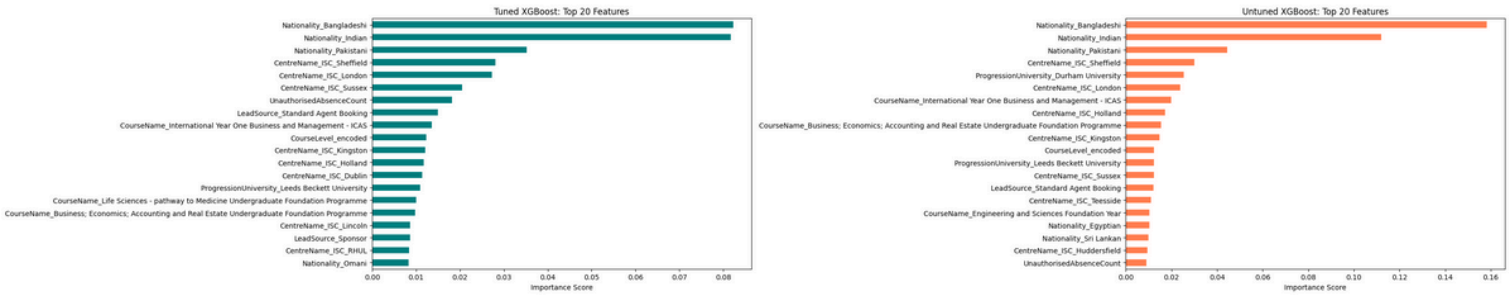


Figure 3. Feature Importance Untuned and Tuned XGBoost (Stage 2)

Although the class imbalance still reflected in the performance of the models, there was improved recall due to the added engagement metrics

Metric	XGboost		Neural Network	
	Untuned	Tuned	Untuned	Tuned
Accuracy	0.90	0.87	0.89	0.86
Precision	0.73	0.54	0.66	0.52
Recall	0.58	0.80	0.56	0.74
F1-Score	0.65	0.64	0.61	0.61
AUC-ROC	0.77	0.84	0.75	0.81

Table 2. Classification metrics for all Stage 2 models

The untuned XGBoost model achieved the highest ROC curve AUC at 0.9090, showing the most classification strength amongst all stage 2 models even without tuning. Surprisingly, the tuned XGBoost model slightly underperformed with an ROC curve AUC of 0.9088, suggesting the original settings were already close to optimal. The untuned neural network followed with an ROC curve AUC of 0.8809, performing well but lagging behind XGBoost. The tuned neural network, despite optimization, recorded the lowest ROC curve AUC at 0.8688.

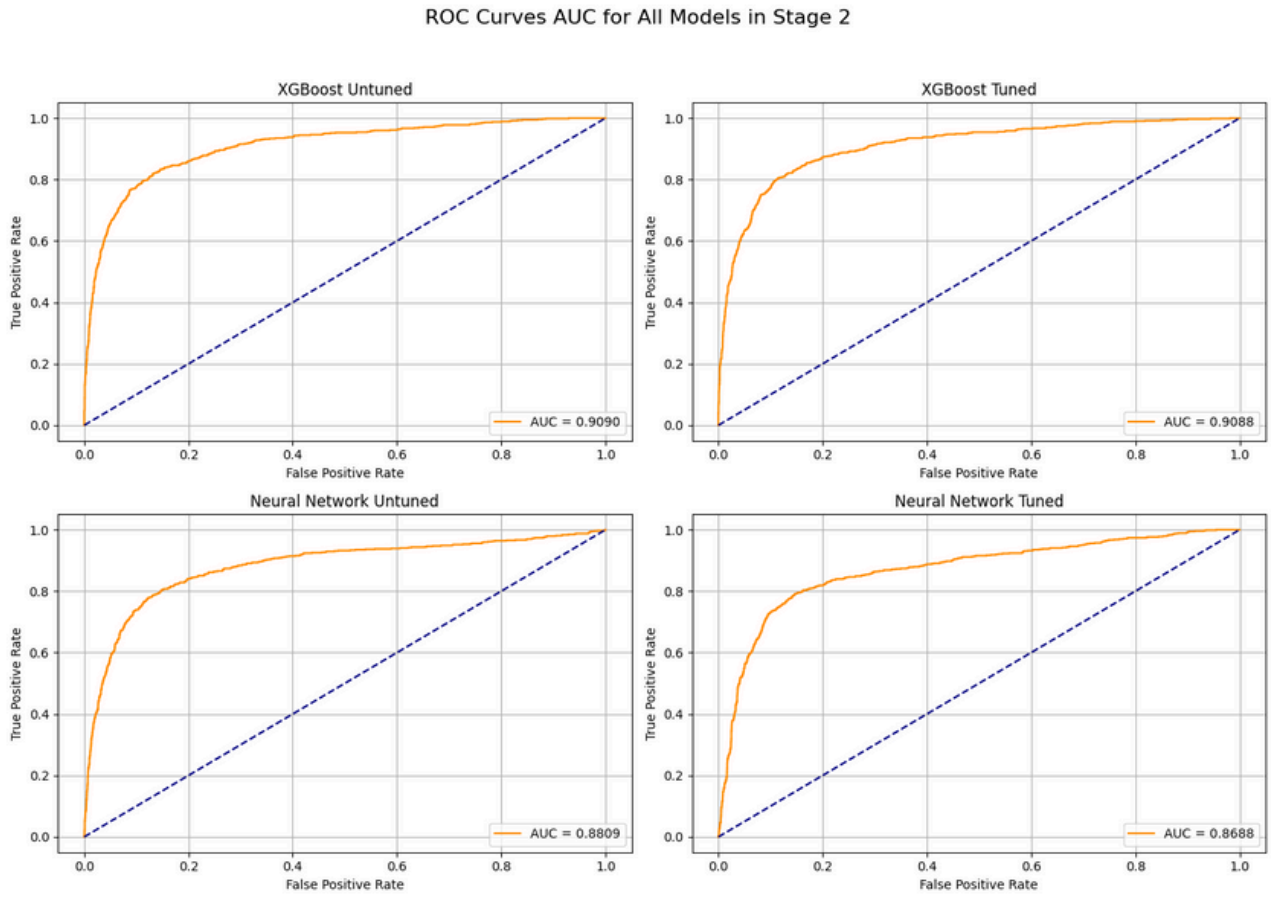


Figure 4. ROC Curve for all Stage 2 models

Stage 3

In Stage 3, the model achieved the highest performance even though they where untuned, largely driven by the strong predictive power of academic metrics. By incorporating student demographics, engagement levels, and academic performance, the model was able to generate the most accurate predictions across all stages.

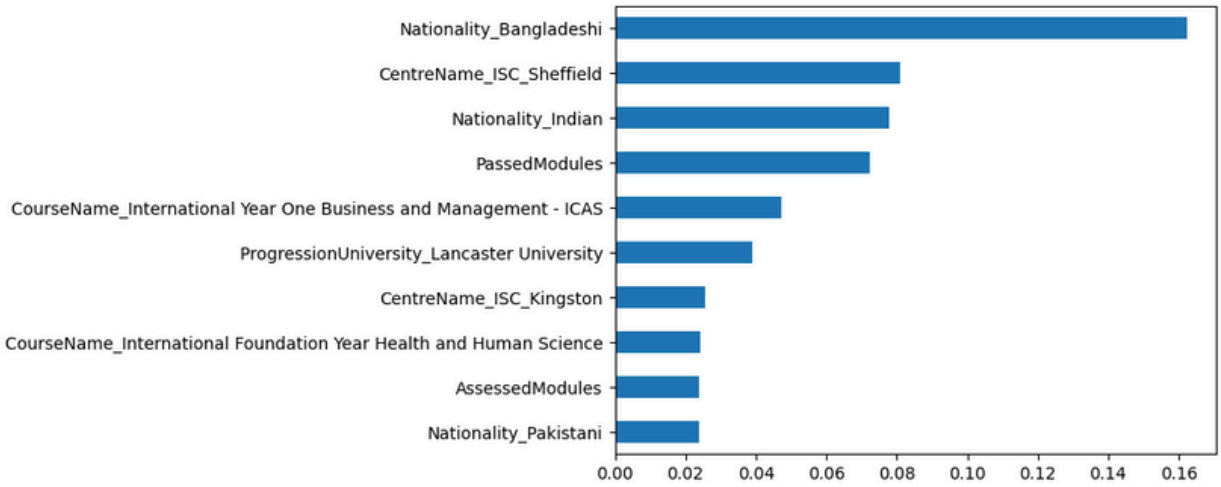


Figure 5. Feature Importance Untuned XGBoost (Stage 3)

Although both models achieved high classification accuracy, the AUC-ROC scores at 93% for XGBoost and 91% for the neural network highlighted the presence of class imbalance in the dataset. This imbalance also affected the recall, with XGBoost reaching 88% and the neural network 83%, which in turn influenced their F1 scores of 90% and 88% respectively.

Metric	XGboost	Neural Network
	Untuned	Untuned
Accuracy	0.97	0.96
Precision	0.92	0.93
Recall	0.88	0.83
F1-Score	0.90	0.88
AUC-ROC	0.93	0.91

Table 3. Classification metrics for all Stage 2 models

The ROC curves show that the untuned XGBoost model performs the best, achieving an AUC of 0.9918. This indicates it has the strongest ability to separate the classes compared to all other models across the three stages.

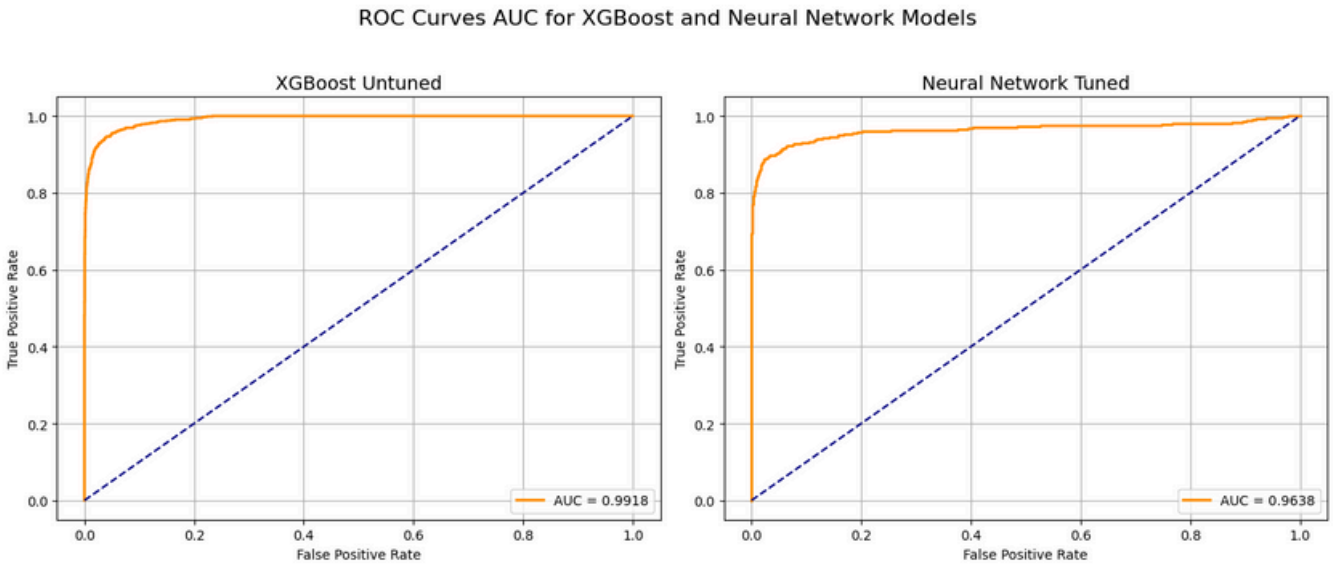


Figure 6. ROC Curve for all Stage 3 models

XGBoost models vs Neural Network

Difference in Performance

1. **Structured Data Compatibility:** XGBoost performed well with the project's tabular data, where feature relationships were clearly structured. Neural networks are better suited for large, complex datasets with less structure. Since the data in this project was relatively straightforward, the neural network model was less effective and showed its limitations.
2. **Training Speed and Efficiency:** XGBoost demonstrated faster training times compared to the neural network, particularly when fine-tuning the latter's hyperparameters. Its sequential boosting approach efficiently refocused learning on previous errors, enhancing time efficiency.
3. **Interpretability:** XGBoost's built-in feature importance scoring offered clear insights into which factors most influenced predictions, enhancing model interpretability. In contrast, neural networks provided less transparency, functioning more opaquely and making it harder to explain their decision-making process.
4. **Regularization and Overfitting Control**: XGBoost's built-in regularization helped prevent overfitting effectively. For the neural network, dropout layers and regularization techniques were also applied to control overfitting. However, due to limited time and computing power, only 10 randomly selected hyperparameter combinations were tested. This means the model may not have reached its optimal performance.

Key Inferences for the Project

The analysis shows that academic performance indicators are the strongest predictors of student dropout. This suggests that Study Group should closely monitor student performance to identify those at risk. However, these metrics are only available later in the academic journey. As a result, they offer limited value for early detection, making it harder to intervene before students begin to disengage.

Early-stage data, such as demographics alone, offer limited predictive power. However, the XGBoost model highlighted certain country demographics course name and center names as influential in its decision-making. These flagged features should be further examined using tools like SHAP to understand whether they contribute positively or negatively to the likelihood of dropout. This could reveal underlying patterns that may inform targeted support strategies.

XGBoost consistently outperformed the neural network as it is better suited for structured, tabular data and effectively captures feature interactions. While the neural network showed potential, it was more sensitive to hyperparameter tuning and required more extensive optimisation to match XGBoost's performance. Due to time and computational constraints, this level of tuning wasn’t fully explored. With greater investment in tuning and computing resources, the neural network may yield stronger results.

Conclusions and Recommendations

This project demonstrates that machine learning, particularly using XGBoost can effectively predict student dropout risk when leveraging academic performance data (Stage 3: AUC 0.99). However, since these metrics become available late, early intervention requires robust engagement tracking (Stage 2: AUC 0.91).

For real world use Study Group should:

1. **Ensure Data Prioritization:**

-Study Group should expand early data collection (e.g., attendance, participation) to enable proactive interventions.

- Academic metrics remain critical but are retrospective; combining stages 2 and 3 data optimizes predictive power.

2. **Ensure Appropriate Model Selection**:

- XGBoost is recommended for deployment due to its superior performance (higher AUC, faster training) and interpretability (clear feature importance).

- Neural networks showed potential but require more tuning and computational resources to compete with XGBoost.

3. **Actionable Interventions**:

- Flag high-risk students early using engagement trends (e.g., unauthorized absences).

- Investigate demographic biases (e.g., nationality, center) via SHAP analysis to ensure equitable support.

4. **Explore Future Enhancements**:

- Hybrid models (e.g., XGBoost + NN ensemble) can be tested as datasets grow.
- Address class imbalance with advanced sampling techniques to improve recall.

Engaging with an industry expert can help validate these findings and ensure that the proposed strategies reflect practical, real-world conditions.

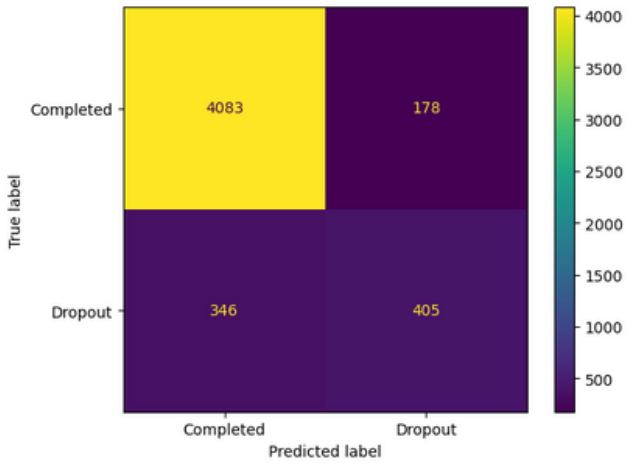
By embedding these models into an early-warning system, Study Group can minimize financial losses, boost student retention, and deliver targeted support to those at risk—turning insights into concrete actions that improve educational outcomes.

Appendix

Stage 1

Untuned XGboost Model

```
XGBoost Classifier Performance
Accuracy: 0.8954509177972865
Precision: 0.6946826758147513
Recall: 0.5392809587217043
F1 Score: 0.6071964017991005
AUC-ROC: 0.7487533636603124
              precision    recall  f1-score   support

           0       0.92      0.96      0.94      4261
           1       0.69      0.54      0.61       751

    accuracy                           0.90      5012
   macro avg       0.81      0.75      0.77      5012
weighted avg       0.89      0.90      0.89      5012
```



Tuned XGboost Model

```
Best Parameters: {'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 200}
XGBoost Classifier Performance
Accuracy: 0.8531524341580208
Precision: 0.5065616797900262
Recall: 0.7709720372836218
F1 Score: 0.6114044350580782
AUC-ROC: 0.8193043711412241
              precision    recall  f1-score   support

           0       0.96      0.87      0.91      4261
           1       0.51      0.77      0.61       751

    accuracy                           0.85      5012
   macro avg       0.73      0.82      0.76      5012
weighted avg       0.89      0.85      0.86      5012
```

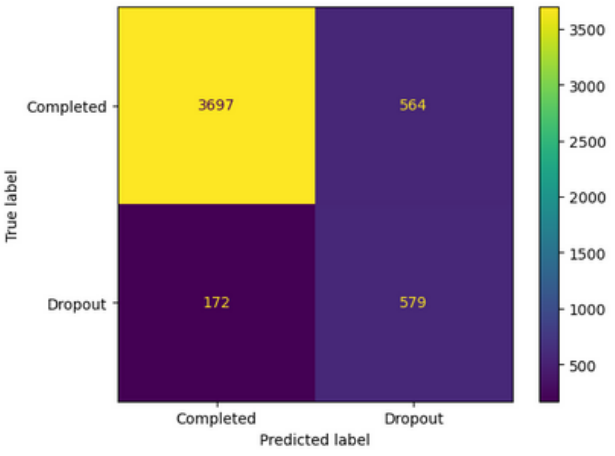
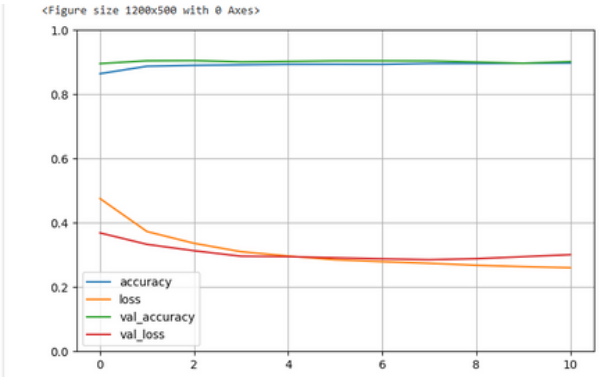
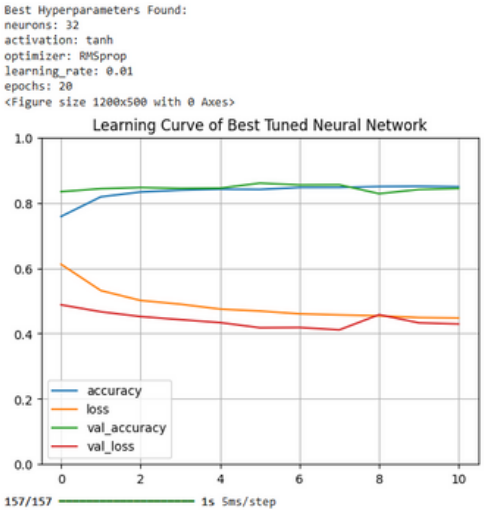


Figure 7. Tuned and Untuned XGBoost Classifier Performance Metrics and Test Data Confusion Metrics (Stage 1)

Simple Neural Network

Tuned Neural Network





Classification Report Untuned Neural Net Model:
Accuracy: 0.8884676775738228
Precision: 0.6758241758241759
Recall: 0.49134487350199735
F1 Score: 0.569005397070162
AUC-ROC: 0.7249026643970912

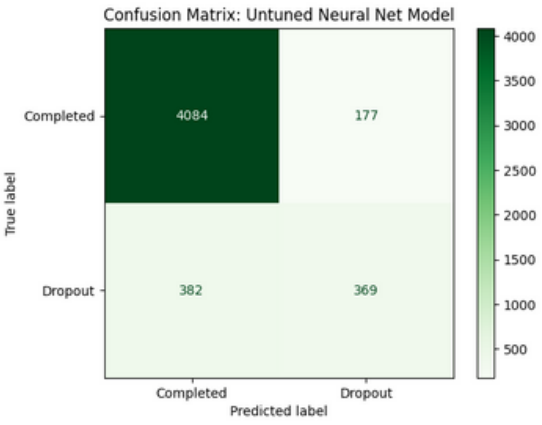
	precision	recall	f1-score	support
0	0.91	0.96	0.94	4261
1	0.68	0.49	0.57	751
accuracy			0.89	5012
macro avg	0.80	0.72	0.75	5012
weighted avg	0.88	0.89	0.88	5012

<Figure size 600x600 with 0 Axes>

Classification Report Tuned Neural Net Model:
Accuracy: 0.8683160415003991
Precision: 0.5455455455455456
Recall: 0.725699067909454
F1 Score: 0.6228571428571429
AUC-ROC: 0.8095756545836874

	precision	recall	f1-score	support
0	0.95	0.89	0.92	4261
1	0.55	0.73	0.62	751
accuracy			0.87	5012
macro avg	0.75	0.81	0.77	5012
weighted avg	0.89	0.87	0.88	5012

<Figure size 600x600 with 0 Axes>



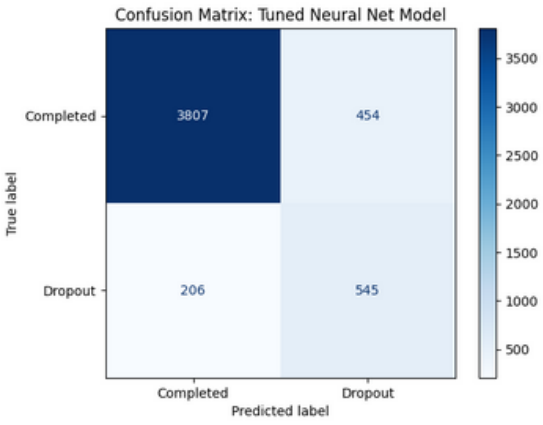
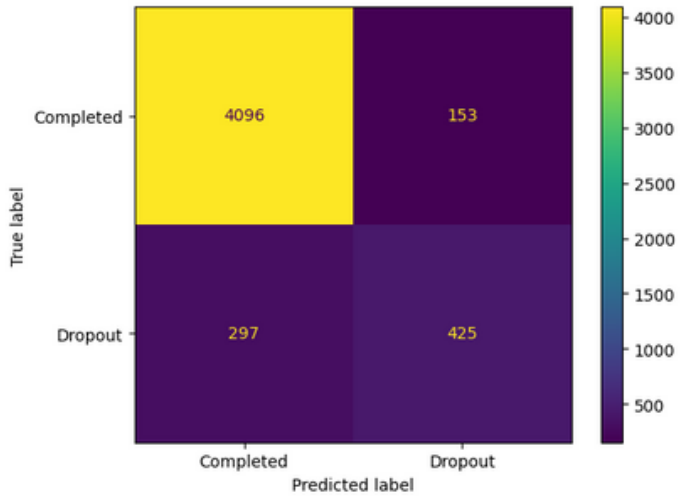


Figure 8. Tuned and Untuned Neural Network Classifier Performance Metrics and Test Data Confusion Metrics (Stage 1)

Stage 2

Untuned XGBoost

```
XGBoost Classifier Performance
Accuracy: 0.9094749547374774
Precision: 0.7352941176470589
Recall: 0.5886426592797784
F1 Score: 0.6538461538461539
AUC-ROC: 0.7763170933489972
              precision    recall  f1-score   support

           0       0.93      0.96      0.95      4249
           1       0.74      0.59      0.65       722

    accuracy                           0.91      4971
   macro avg       0.83      0.78      0.80      4971
weighted avg       0.90      0.91      0.91      4971
```



Tuned XGBoost

```
Best Parameters: {'learning_rate': 0.2, 'max_depth': 5, 'n_estimators': 200}
XGBoost Classifier Performance
Accuracy: 0.8736672701669684
Precision: 0.5440900562851783
Recall: 0.8033240997229917
F1 Score: 0.6487695749440716
AUC-ROC: 0.844472122819839
              precision    recall  f1-score   support

           0       0.96      0.89      0.92      4249
           1       0.54      0.80      0.65       722

    accuracy                           0.87      4971
   macro avg       0.75      0.84      0.79      4971
weighted avg       0.90      0.87      0.88      4971
```

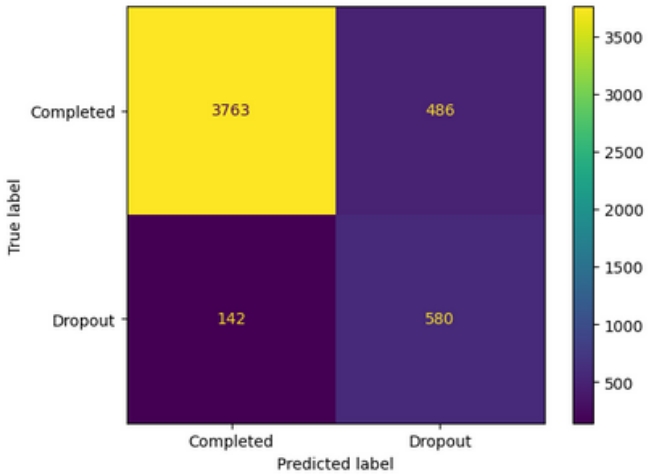
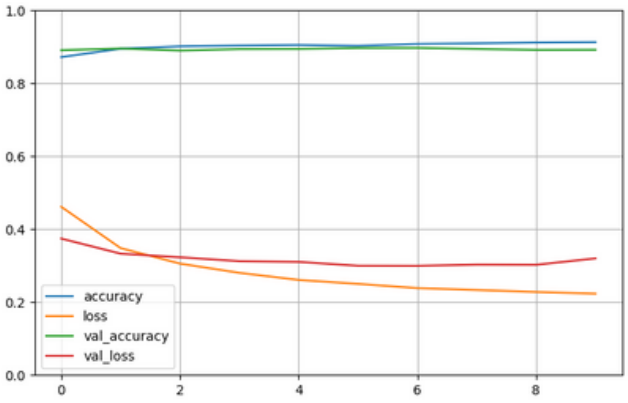
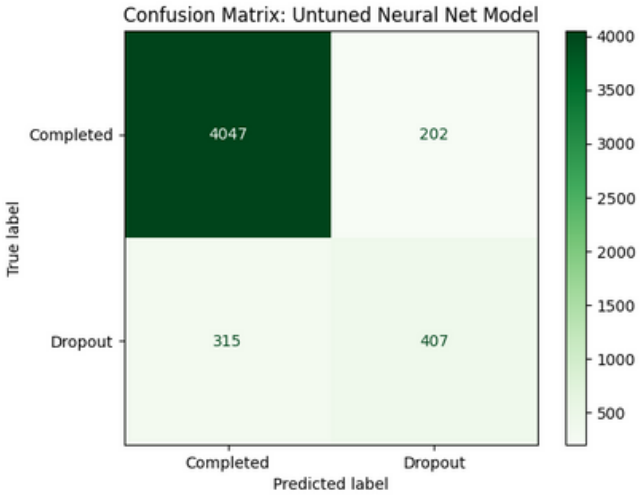


Figure 9. Tuned and Untuned XGBoost Classifier Performance Metrics and Test Data Confusion Metrics (Stage 2)

Simple Neural Network

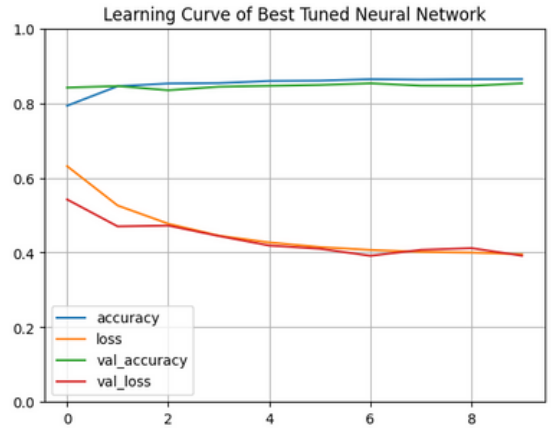



```
Classification Report Untuned Neural Net Model:
Accuracy: 0.895996781331724
Precision: 0.6683087027914614
Recall: 0.5637119113573407
F1 Score: 0.6115702479338843
AUC-ROC: 0.758085656784813
              precision    recall  f1-score   support

           0       0.93      0.95      0.94      4249
           1       0.67      0.56      0.61       722

    accuracy                           0.90      4971
   macro avg       0.80      0.76      0.78      4971
weighted avg       0.89      0.90      0.89      4971
```



Tuned Neural Network

```
Best Hyperparameters Found:
neurons: 64
activation: tanh
optimizer: RMSprop
learning_rate: 0.01
epochs: 20
<Figure size 1200x500 with 0 Axes>
```




```
Classification Report Tuned Neural Net Model:
Accuracy: 0.8666264333132166
Precision: 0.5290068829891839
Recall: 0.7451523545706371
F1 Score: 0.6187464059804485
AUC-ROC: 0.8162099734726568
              precision    recall  f1-score   support

           0       0.95      0.89      0.92      4249
           1       0.53      0.75      0.62       722

    accuracy                           0.87      4971
   macro avg       0.74      0.82      0.77      4971
weighted avg       0.89      0.87      0.88      4971
```

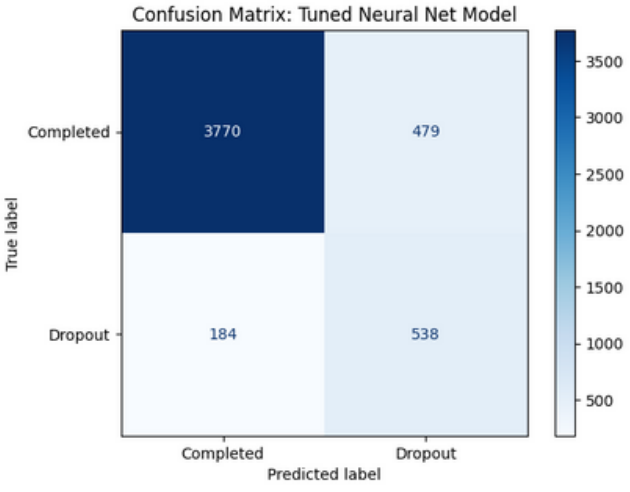


Figure 10. Tuned and Untuned Neural Network Classifier Performance Metrics and Test Data Confusion Metrics (Stage 2)

Stage 3

Untuned XGBoost

```
XGBoost Classifier Performance
Accuracy: 0.9714684756584198
Precision: 0.9208333333333333
Recall: 0.8851802403204272
F1 Score: 0.9026548672566371
AUC-ROC: 0.9359046873664064
              precision    recall  f1-score   support

           0       0.98      0.99      0.98      4263
           1       0.92      0.89      0.90       749

    accuracy                           0.97      5012
   macro avg       0.95      0.94      0.94      5012
weighted avg       0.97      0.97      0.97      5012
```

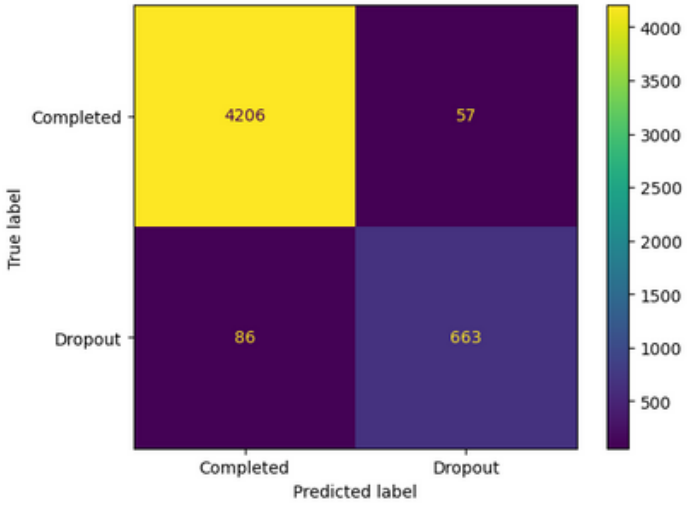
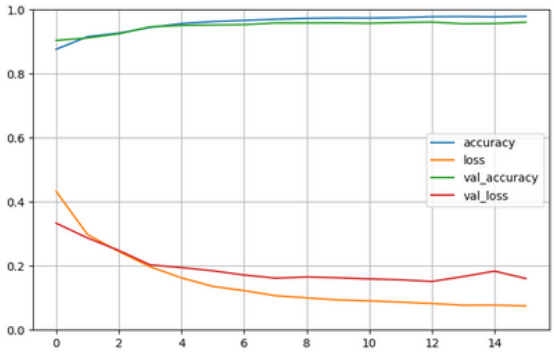


Figure 11. Untuned XGBoost Classifier Performance Metric and Test Data Confusion Metrics (Stage 3)

Simple Neural Network




```
Classification Report Untuned Neural Net Model:
Accuracy: 0.9660814046288907
Precision: 0.9314456035767511
Recall: 0.8344459279038718
F1 Score: 0.8802816901408451
AUC-ROC: 0.9118277023990389
              precision    recall  f1-score   support

           0       0.97      0.99      0.98      4263
           1       0.93      0.83      0.88       749

    accuracy                           0.97      5012
   macro avg       0.95      0.91      0.93      5012
weighted avg       0.97      0.97      0.97      5012
```

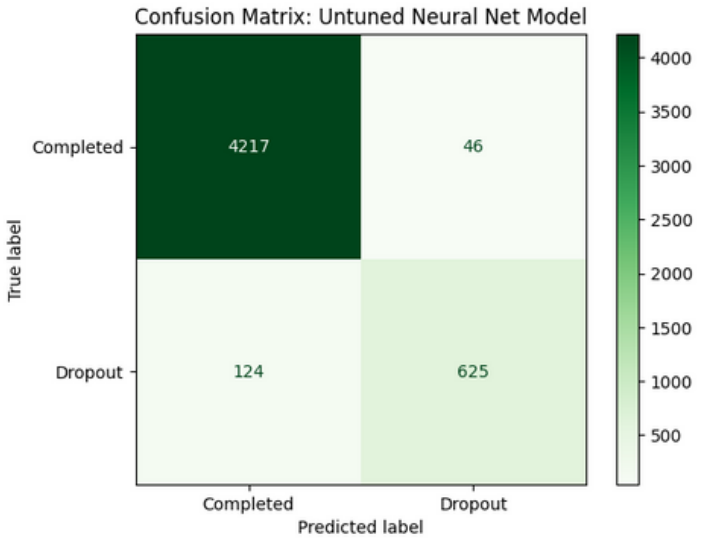


Figure 13. Untuned Neural Network Classifier Performance Metrics and Test Data Confusion Metrics (Stage 3)