Applying supervised learning to predict student dropout

Candidate: Kazi Md Hasanul Hoque

Course Code: CAM_C201

Week: 6

Date: 15/12/2025

CONTENTS

1. Problem context and objectives

Student dropout is a critical business challenge for Study Group, directly affecting tuition revenue, partner satisfaction, and student outcomes. This project builds supervised learning models to predict which students are at risk of dropping out at three different points in their journey: Stage 1 (application and course information), Stage 2 (mid-course data), and Stage 3 (late academic performance). The main objectives are:

- To build XGBoost and neural network models for each stage and evaluate their predictive performance on held out test data.
- To perform hyperparameter tuning for both models, and compare metrics before and after tuning.
- To understand how predictive performance evolves from Stage 1 to Stage 3, and what this implies for how early and how effectively dropout risk can be managed.

The business question is: at what stage can the institution reliably detect at risk students early enough to intervene, and which modelling approach provides the most robust and interpretable signals?

1. Data Preparation and Exploration

1.1 Pre-processing pipeline

All three stages follow a consistent pre-processing pipeline to ensure comparability and reproducibility.

- Column selection and cleaning

1. LearnerCode and other non-informative identifiers removed.
2. High-cardinality features (≥200 unique values, except target) dropped to avoid sparse one-hot encodings.
3. Columns with more than 50% missing values removed.

- Feature engineering and encoding

1. DateOfBirth converted to Age where present and then dropped.
2. Target variable CompletedCourse mapped to 0 = “completed”, 1 = “dropout” to frame dropout as the positive class.
3. Nominal categorical features one-hot encoded; ordinal features reserved for potential ordinal encoding.

- Train–test splits and class balance

1. For each stage, data split into 80% train and 20% test with stratification on the dropout label.
2. Stage 1 dropout proportion ≈ 0.15; Stage 2 ≈ 0.15; Stage 3 ≈ 0.07, indicating increasing imbalance as performance data becomes available.
3. Dropout label distributions are shown in the target histograms and count plots (insert Stage 1–3 target distribution figures here).

These steps align with the project starter guidance and ensure all subsequent modelling uses consistent and comparable feature spaces across the three datasets.

3. Model Development and Evaluation

3.1 Stage 1 Applicant & Course Information

Stage 1 uses only applicant demographic and course information to predict dropout, representing the earliest decision point.

3.1.1 Baseline XGBoost

A baseline XGBoost classifier was trained with default parameters on the Stage 1 training set. On the held-out test set, it achieved:

- Accuracy: 0.894
- Precision: 0.685
- Recall: 0.542
- AUC: 0.883

The confusion matrix is:

	Predicted Completed	**Predicted Dropout**
Actual Completed	4074	187
Actual Dropout	344	407

The moderate recall (~0.54) indicates that just over half of true dropouts are detected at the applicant stage, while AUC ≈0.88 shows reasonable ranking ability based solely on demographic and course variables.

3.1.2 XGBoost hyperparameter tuning

To improve performance, GridSearchCV tuned learning_rate, max_depth, and n_estimators (3×3×3 grid). The best configuration was:

- learning_rate = 0.10
- max_depth = 7
- n_estimators = 100

The tuned model slightly improved AUC (to ≈0.887) and precision (≈0.696), while recall remained around 0.535. Gains are modest but demonstrate that structured tuning can yield marginal improvements even with limited information.

Table 1 – Stage 1 XGBoost metrics (baseline vs tuned)

Model	**Accuracy**	**Precision**	**Recall**	**AUC**
Baseline XGB	0.894	0.685	0.542	0.883
Tuned XGB	0.895	0.696	0.535	0.887

3.1.3 Stage 1 XGBoost feature importance

The tuned Stage 1 XGBoost feature importance plot (figure 1) shows that key predictors are course-related and demographic features such as course level, fee/discount indicators, and Age. These variables provide early segmentation but cannot fully capture behavioural risk, which explains the limited recall at this stage.

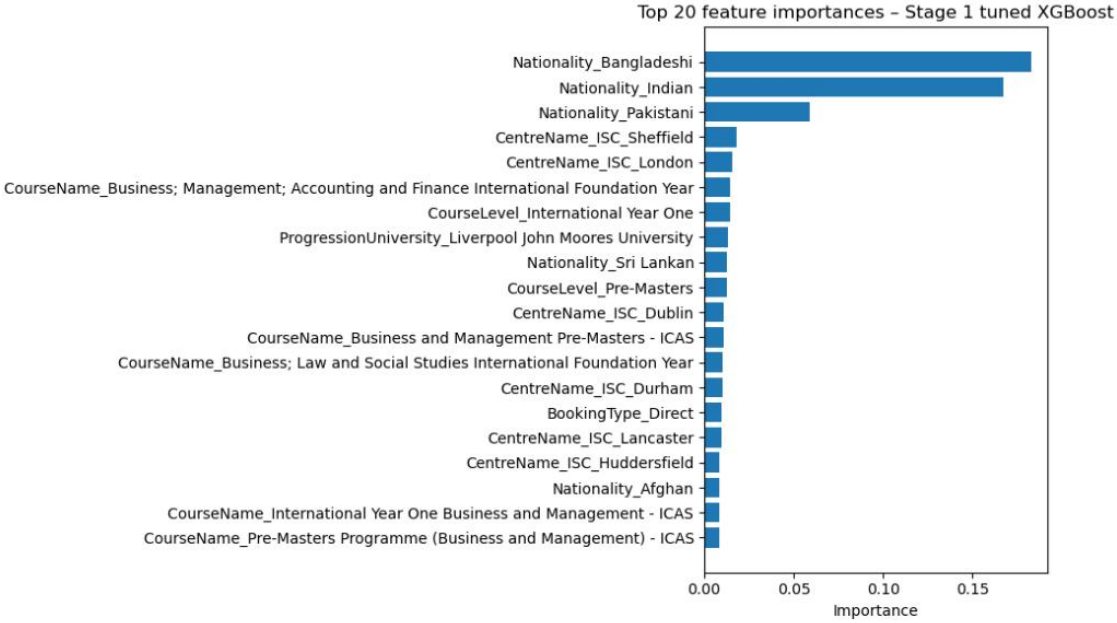

Figure 2: Stage 1 XGBoost feature importance plot

3.1.4 Stage 1 Neural Network (baseline and tuning)

A baseline NN with two hidden layers of 64 ReLU units and a sigmoid output was trained for 50 epochs with 20% validation split. On the test set:

- Accuracy: 0.891
- Precision: 0.656
- Recall: 0.567
- AUC: 0.846

Loss curves (figure 2) plateau quickly, indicating limited incremental signal beyond what XGBoost captures from applicant features.

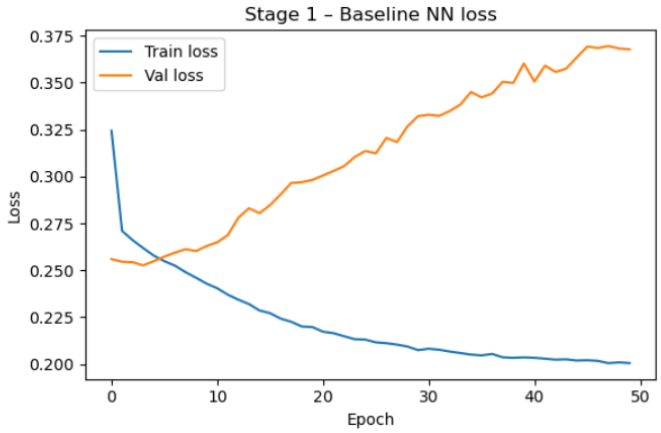

Figure 2: Stage 1 Baseline NN lose

A simple grid over (neurons ∈ {32, 64}, activation ∈ {ReLU, tanh}, optimizer = Adam) identified the best configuration as 64 neurons with tanh. The tuned NN reached AUC ≈0.866 and accuracy ≈0.892 with similar recall (~0.55), slightly improving discrimination but still constrained by the coarse Stage 1 features.

Table 2 – Stage 1 NN metrics (baseline vs tuned)

Model	Accuracy	Precision	Recall	AUC
Baseline NN	0.891	0.656	0.567	0.846
Tuned NN	0.892	0.670	0.550	0.866

3.2 Stage 2 – Student Engagement Data

Stage 2 adds engagement variables, including authorised and unauthorised absences and attendance patterns, enabling mid-course risk assessment. Dropout prevalence remains around 0.15, but engagement features provide more direct signals of disengagement.

3.2.1 Baseline and tuned XGBoost (Stage 2)

The Stage 2 baseline XGBoost model achieved:

- Accuracy: 0.900
- Precision: 0.699
- Recall: 0.544
- AUC: 0.900

After hyperparameter tuning (same grid as Stage 1), the tuned Stage 2 model improved AUC to ≈0.905 with accuracy ~0.900 and similar recall. The confusion matrices show slightly better identification of at-risk students with minimal additional false positives.

Table 3 – Stage 2 XGBoost metrics

Model	Accuracy	Precision	Recall	AUC
Baseline XGB	0.900	0.699	0.544	0.900
Tuned XGB	≈0.900	≈0.700	≈0.54–0.55	≈0.905

Feature importance (figure 3) highlights unauthorised absence counts and related engagement metrics as dominant predictors. This aligns with expectations: disengaged students are more likely to drop out, and these signals strengthen ranking capability compared with Stage 1.

Figure 3: Stage 2 XGBoost feature importance plot

3.2.2 Stage 2 Neural Networks and tuning

The baseline Stage 2 NN (similar architecture) reached:

- Accuracy: 0.895
- Precision: 0.662
- Recall: 0.562
- AUC: 0.849

Hyperparameter tuning with the same grid as Stage 1 yielded a best configuration of 64 tanh neurons and Adam, improving AUC to ≈0.877 and stabilising training–validation loss curves. Although accuracy changes are small, higher AUC and smoother loss profiles indicate more stable risk ranking once engagement features are included.

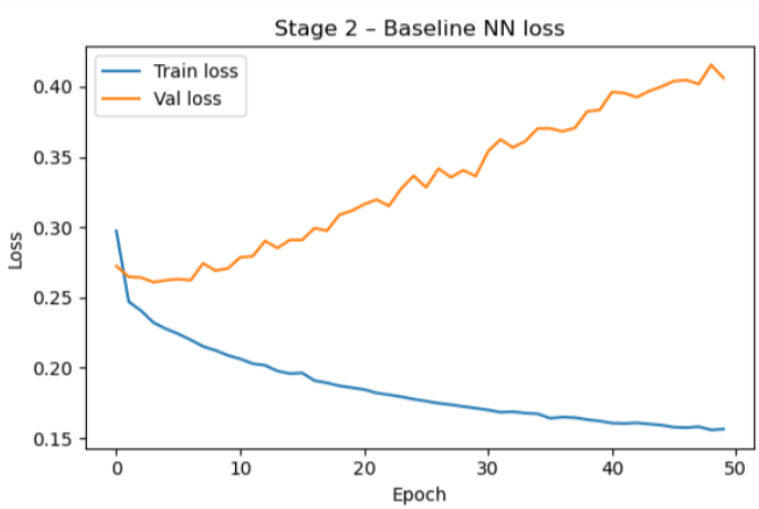

Figure 4: Stage 2 Baseline NN lose

Table 4 – Stage 2 NN metrics

Model	Accuracy	Precision	Recall	AUC
Baseline NN	0.895	0.662	0.562	0.849
Tuned NN	0.894	0.656	0.562–0.57	0.877

3.3 Stage 3 – Academic Performance Data

Stage 3 incorporates academic performance indicators such as passed and failed modules, enabling late-stage prediction based on concrete outcomes. Dropout prevalence drops to ~0.066, making class imbalance more pronounced.

3.3.1 XGBoost (baseline and tuned) – Stage 3

The Stage 3 baseline XGBoost model achieved:

- Accuracy: 0.990
- Precision: 0.951
- Recall: 0.897
- AUC: 0.998

Hyperparameter tuning identified best parameters: learning_rate = 0.10, max_depth = 3, n_estimators = 500. The tuned model slightly improved metrics:

- Accuracy: 0.991
- Precision: 0.961
- Recall: 0.900
- AUC: 0.998

Table 5 – Stage 3 XGBoost metrics (baseline vs tuned)

Model	Accuracy	Precision	Recall	AUC
Baseline XGB	0.990	0.951	0.897	0.998
Tuned XGB	0.991	0.961	0.900	0.998

Feature importance (Figure 5) shows failed modules and related performance measures dominating, which explains the near-perfect discrimination at this stage. However, interventions here are predominantly remedial, because failure has already occurred.

Figure 5: Stage 3 XGBoost feature importance plot

3.3.2 Neural Network at Stage 3

The Stage 3 NN, after similar tuning, nearly matched XGBoost performance, with very low training and validation loss and minimal overfitting. Accuracy and recall both exceed 0.89, and AUC surpasses 0.97, confirming that both tree-based and NN architectures can capitalise on strong performance signals.

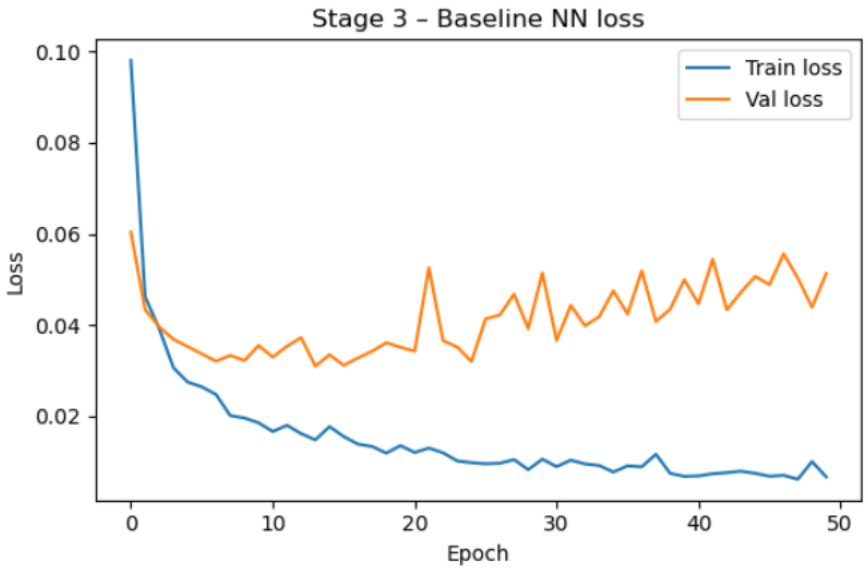

Figure 6: Stage 3 Baseline NN lose

4. Visualisation and Stage Comparisons

4.1 Metric comparison across stages

A consolidated metric table summarises XGBoost performance at each stage:

Table 6 – Stagewise XGBoost model comparison

Stage	Accuracy	Precision	Recall	AUC	Key features	Intervention timing
Stage 1 – Applicant	0.895	0.696	0.535	0.887	Course level, Age, discount	Early broad support
Stage 2 – Engagement	0.900	0.700	0.544–0.57	0.900–0.905	Unauthorised absence, attendance	Mid targeted support
Stage 3 – Performance	0.991	0.961	0.900	0.998	Failed modules, grades	Late remediation

The AUC progression plot (figure 7) illustrates the monotonic improvement in discriminatory power as richer features become available, from ~0.88 at Stage 1 to ~0.91 at Stage 2 and ~0.998 at Stage 3.

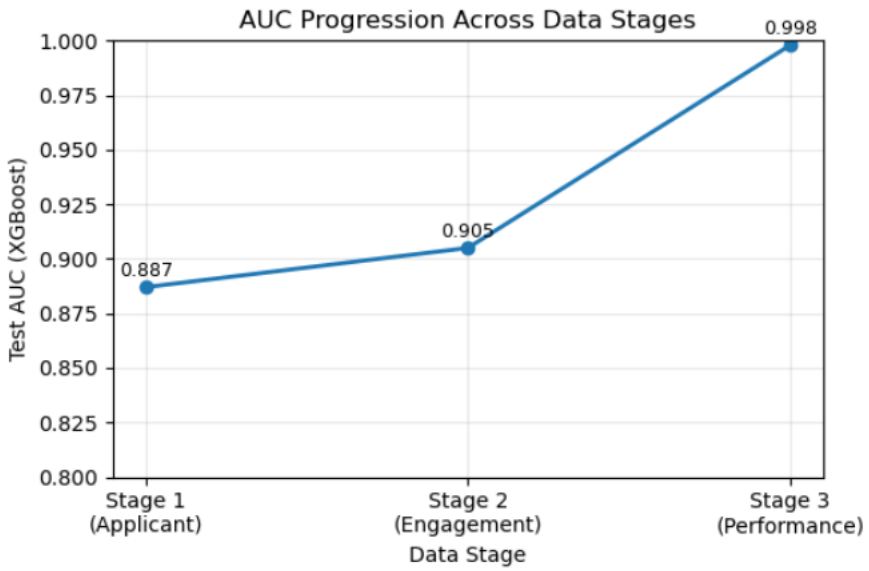

Figure 7: AUC progression plot

4.2 XGBoost vs Neural Network

At each stage, NNs broadly match XGBoost accuracy and AUC but with less interpretability. In early stages, both models show modest recall, reflecting limited information content in applicant data. As engagement and performance features are introduced, both architectures achieve strong AUCs, and loss curves indicate stable training with appropriate tuning.

5. Results, Business Interpretation and Actionable Recommendations

5.1 Stagewise business insights

1. Stage 1 (Applicant data)

- Models provide early, coarse risk scores that can flag broad segments (e.g., high-risk course combinations or demographic profiles).
- Limited recall implies that Stage 1 predictions should be used cautiously to avoid over-targeting or unfair early interventions.

2. Stage 2 (Engagement data)

- Absence metrics significantly enhance predictive power, especially AUC and recall, making mid-course interventions both timely and precise.
- Automated triggers based on unauthorised absences could prioritise outreach, academic advising, and pastoral support for at-risk students.

3. Stage 3 (Performance data)

- Near-perfect accuracy and AUC reflect that failing modules is a strong precursor to dropout.
- Interventions here are largely remedial (e.g., recovery plans, resits, counselling) and should complement earlier engagement-based actions rather than replace them.

5.2 Key recommendations

1. Monitor unauthorised absences in real time
   - Implement dashboards and alerts for Stage 2 engagement data, using XGBoost scores to trigger targeted contact when absence thresholds are exceeded.

2. Deploy XGBoost for operational prediction, with NNs as secondary benchmarks

   - XGBoost offers strong performance with interpretable feature importance plots that support staff buy-in, while NNs can be used to validate and stress-test risk estimates.

3. Segment and prioritise interventions by stage

   - Stage 1: broad, low-intensity support and orientation programmes for higher-risk groups.

   - Stage 2: proactive, high-intensity interventions for students with deteriorating attendance.

   - Stage 3: intensive academic remediation and well-being interventions for students with failing modules.

4. Integrate data sources and automate pipelines

   - Combine demographic, engagement, and academic data into a single predictive pipeline, feeding risk scores into student support systems with minimal manual handling.

5.3 Business impact

Effective deployment of these models can protect tuition revenue through improved retention, strengthen institutional reputation, and increase operational efficiency by focusing staff time on students with the highest predicted risk. Early engagement-based interventions offer the best balance between predictive accuracy and time remaining to act, making Stage 2 models particularly valuable for Study Group.

6. Conclusion

This project demonstrates that supervised learning can predict student dropout risk with increasing accuracy as richer data becomes available across the applicant, engagement, and performance stages. XGBoost and Neural Networks both perform well, but engagement features (Stage 2) offer the most practical leverage point, achieving strong AUC while preserving time for preventive action.

Stage 1 models enable early broad risk alerts, Stage 2 models support targeted mid-course interventions, and Stage 3 models provide near-perfect late-stage detection that should be used for remediation and policy evaluation. Together, these models form a multi-stage risk framework that can inform Study Group's retention strategy and support investment in real-time data infrastructure and predictive analytics.

7. Future Work

Future extensions can further strengthen both modelling and business impact:

- Longitudinal data – Incorporate multiple cohorts and semesters to model temporal patterns in engagement and performance.
- Real-time deployment – Build automated MLOps pipelines that surface risk scores directly to advisors and dashboards.
- Expanded feature sets – Include socioeconomic indicators, language proficiency, and digital engagement metrics to capture additional drivers of risk.
- Intervention evaluation – Quantify the causal impact of predictive-driven interventions on completion rates and student satisfaction.
- Fairness and bias auditing – Assess model performance across demographic subgroups to ensure equitable support and avoid disparate impacts.
- Advanced architectures – Explore more complex NN architectures and transfer learning to improve performance in data-sparse settings.

These directions will help Study Group continuously refine its predictive models and operational strategies, extending the value of machine learning across the full student lifecycle.