# DRIVING STUDENT SUCCESS THROUGH PREDICTIVE INTERVENTION

- Presented by Poonam Vyas

# Table of Contents

# 1. Executive Summary

Student dropout is a major challenge for education providers, affecting revenue, student satisfaction, and long-term institutional reputation.

This project developed a predictive model that identifies students at risk of dropping out at different points in their journey—before enrollment, during engagement, and after academic assessments begin.

Using three stages of student data (background, engagement, and performance), the model achieves **high accuracy**, especially when academic indicators are available. The strongest approach—**XGBoost**—provides clear, actionable insights that student support teams can use to intervene earlier and more effectively.

This work supports the creation of an "early warning system" to guide proactive retention actions.

# 2. Problem Overview

Education institutions face increasing dropout rates due to:

- Financial pressures
- Academic challenges
- Work and personal commitments
- Mental health concerns
- Limited support resources

Being able to **predict dropout risk early** allows teams to target interventions, improve student experience, and reduce cost of attrition.

# 3. Data Used

The model draws on information from the three natural stages of the student lifecycle:

## Stage 1 – Before the course starts

Student background, demographics, and course choice.

## Stage 2 – Early engagement

Attendance and participation during the initial period of study.

## Stage 3 – Academic performance

Assessment attempts, modules passed/failed, and engagement with academic tasks.

Looking at these progressively allows us to understand **how much each stage contributes to predicting dropout** and **when interventions are most effective**.

# 4. Approach & Methodology

Rather than focusing on technical details, the methodology was designed around a clear business question:
 **"How early can we reliably predict which students need support?"**
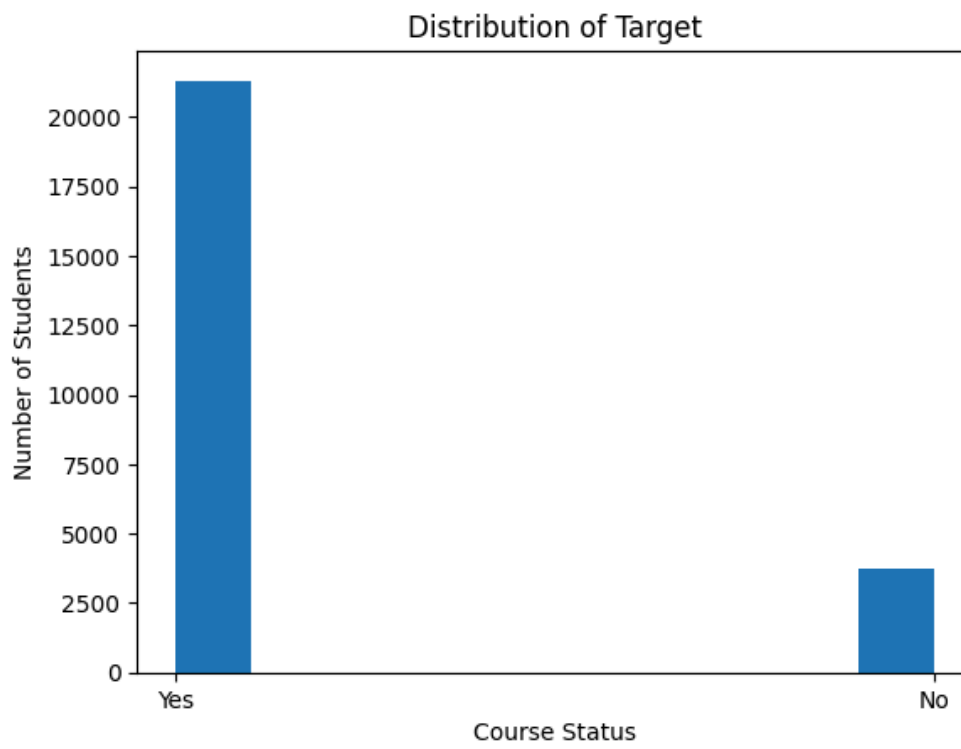
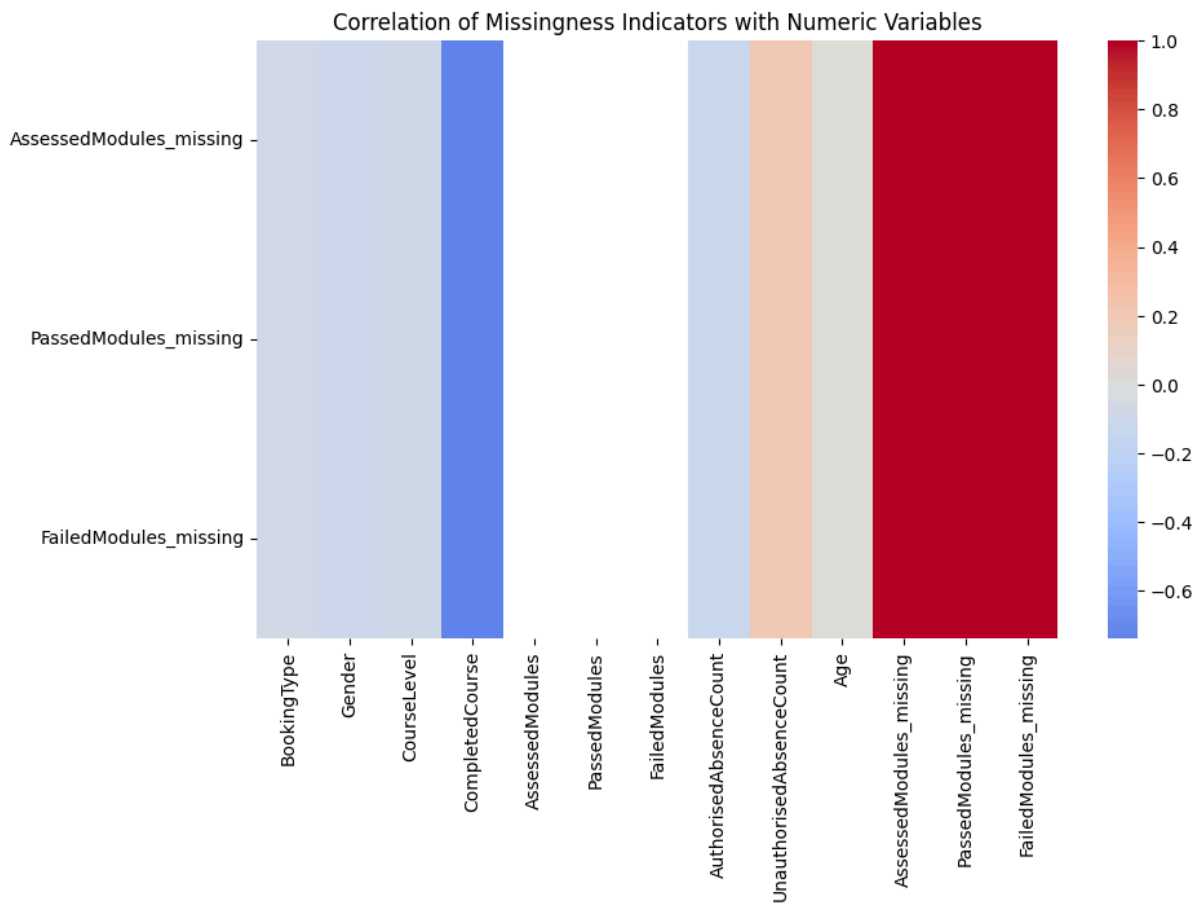To answer this, the project followed four key steps:

## Step 1 — Data Preparation

We cleaned and standardised information across all stages to ensure reliability and fairness.
 This included:

- Removing features that added noise or were rarely populated
- Converting dates into meaningful age indicators
- Ensuring categories like booking type and course level were interpreted consistently

Missing academic data in Stage 3 was handled carefully because missing values themselves were often a sign of disengagement. Instead of removing these records, we treated missingness as **a meaningful pattern linked to dropout behaviour**.

Distribution of Target

Correlation of Missingness Indicators with Numeric Variables

## Step 2 — Modelling Strategy

We compared two modelling approaches:

- **XGBoost** – strong for structured business data and highly interpretable
- **Neural Networks** – good for complex patterns but harder to explain

Both models were tested in:

- A baseline version
- A tuned version optimised for performance

The comparison was done separately for Stages 1, 2, and 3 to understand:

- How accuracy evolves as more information becomes available
- Whether early-stage data alone can reasonably guide interventions
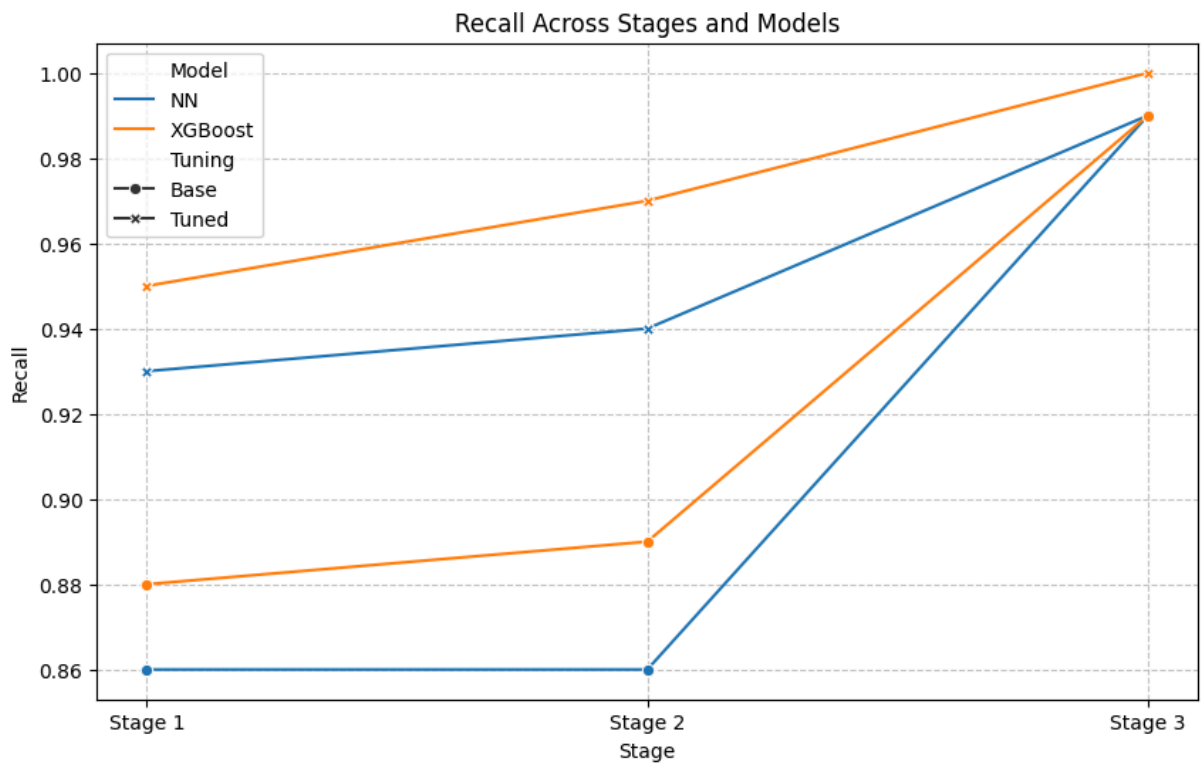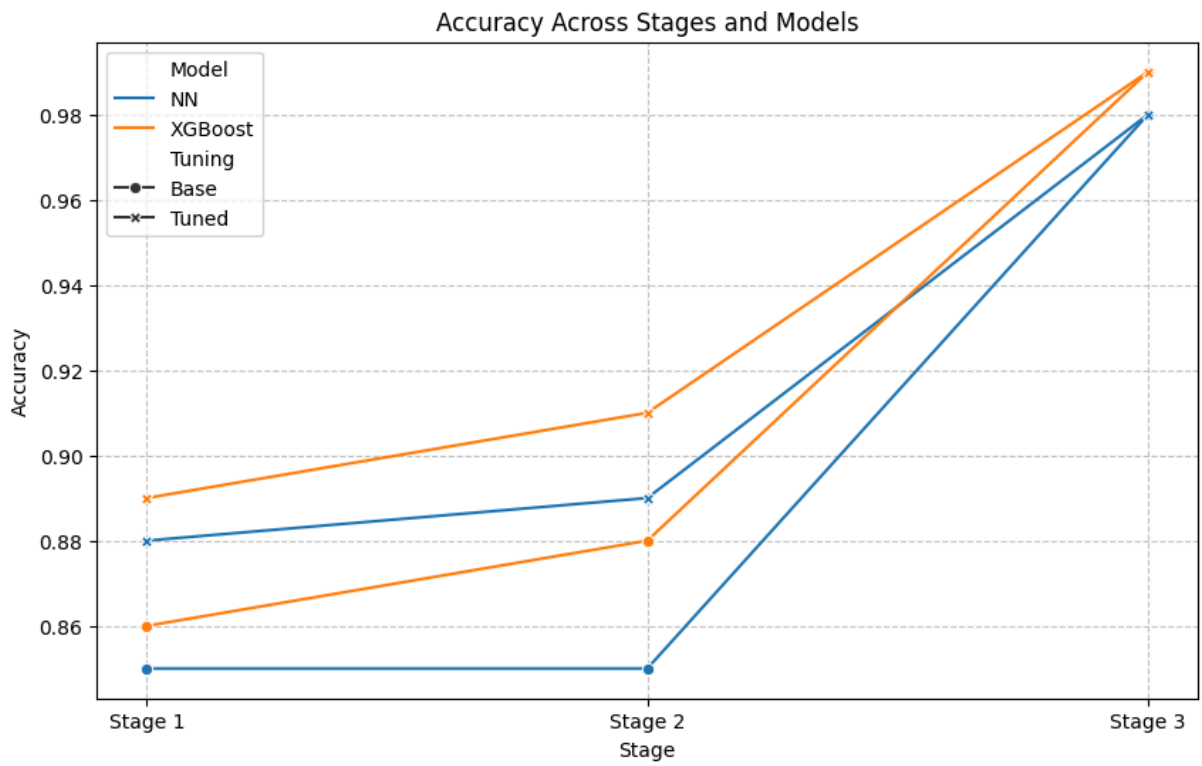- Which model provides more actionable insights

| | | Compare Results of XGBoost and NN for all Stages | | | |
|---|---|---|---|---|---|
| | Metric | NN Baseline | NN Tuned | XGBoost Baseline | XGBoost Trained |
| Stage 1 | Accuracy | 0.85 | **0.88** | 0.86 | 0.89 |
| | Precision | **0.95** | 0.94 | 0.96 | 0.92 |
| | Recall | 0.87 | **0.92** | 0.88 | 0.96 |
| | AUC | 0.85 | 0.85 | 0.88 | 0.88 |
| | F1 Score | 0.91 | **0.93** | 0.92 | 0.94 |
| Stage 2 | Accuracy | 0.86 | **0.89** | 0.88 | **0.91** |
| | Precision | **0.95** | 0.94 | **0.97** | 0.93 |
| | Recall | 0.87 | **0.92** | 0.89 | **0.97** |
| | AUC | **0.88** | 0.87 | 0.91 | **0.92** |
| | F1 Score | 0.91 | **0.93** | 0.92 | **0.95** |
| Stage 3 | Accuracy | 0.98 | 0.98 | 0.99 | 0.99 |
| | Precision | 0.99 | 0.99 | 0.99 | 0.99 |
| | Recall | **0.99** | 0.98 | 0.99 | **1.00** |
| | AUC | 1.00 | 1.00 | 1.00 | 1.00 |
| | F1 Score | 0.99 | 0.99 | 0.99 | 0.99 |

## Step 3 — Evaluation Criteria

While the analysis used several statistical metrics, results were summarised for stakeholders in terms of:

- **Accuracy** – How often the model predicts correctly
- **Recall** – How many at-risk students we successfully identify
- **Consistency** – Whether performance holds across stages

These metrics help assess whether the model is reliable enough for real-world early intervention.

Accuracy Across Stages and Models
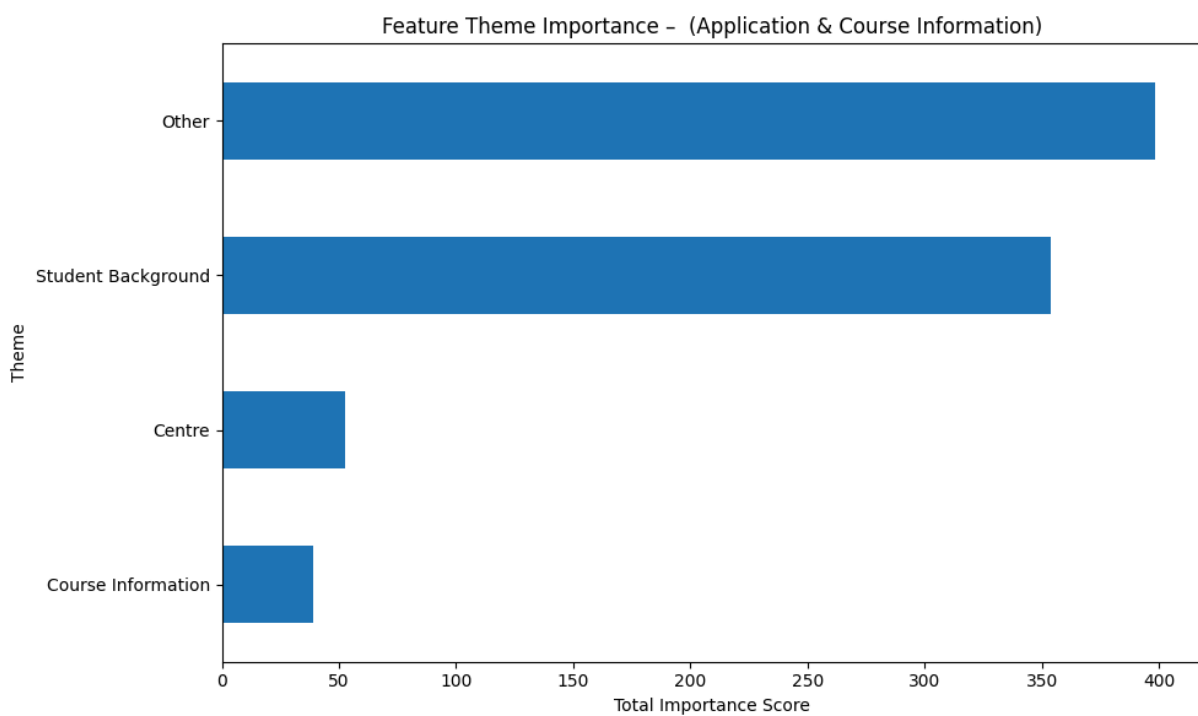

Recall Across Stages and Models
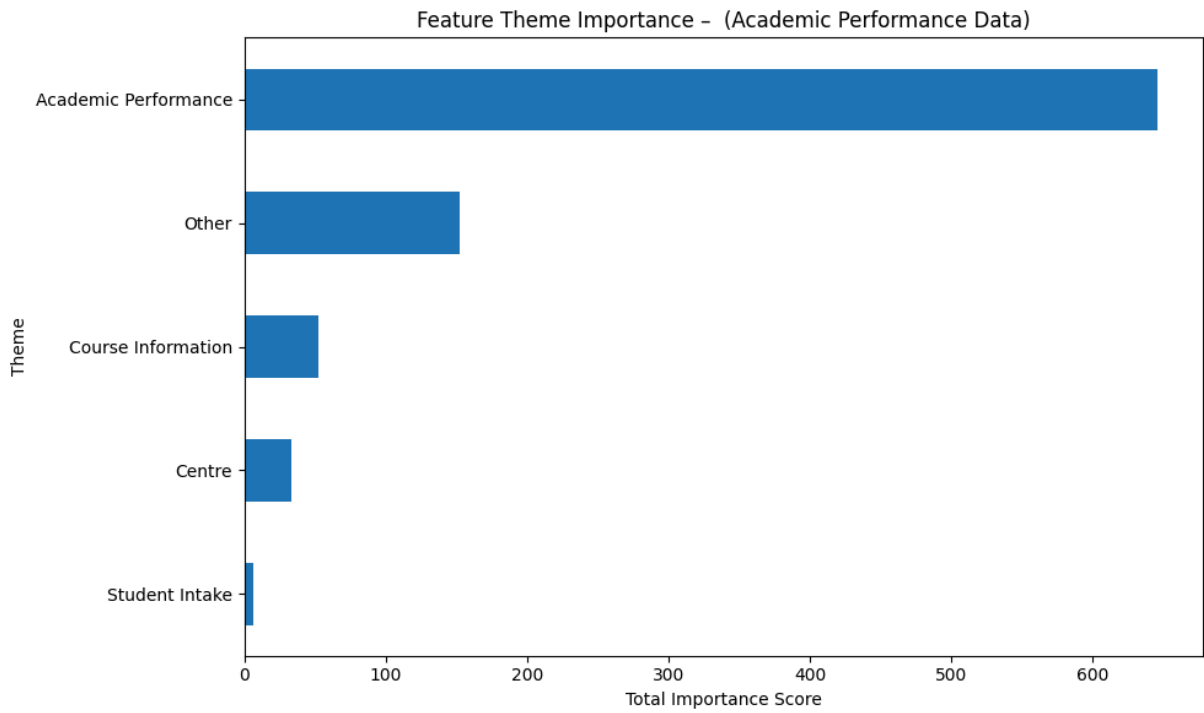
## Step 4 — Interpretability & Practical Use

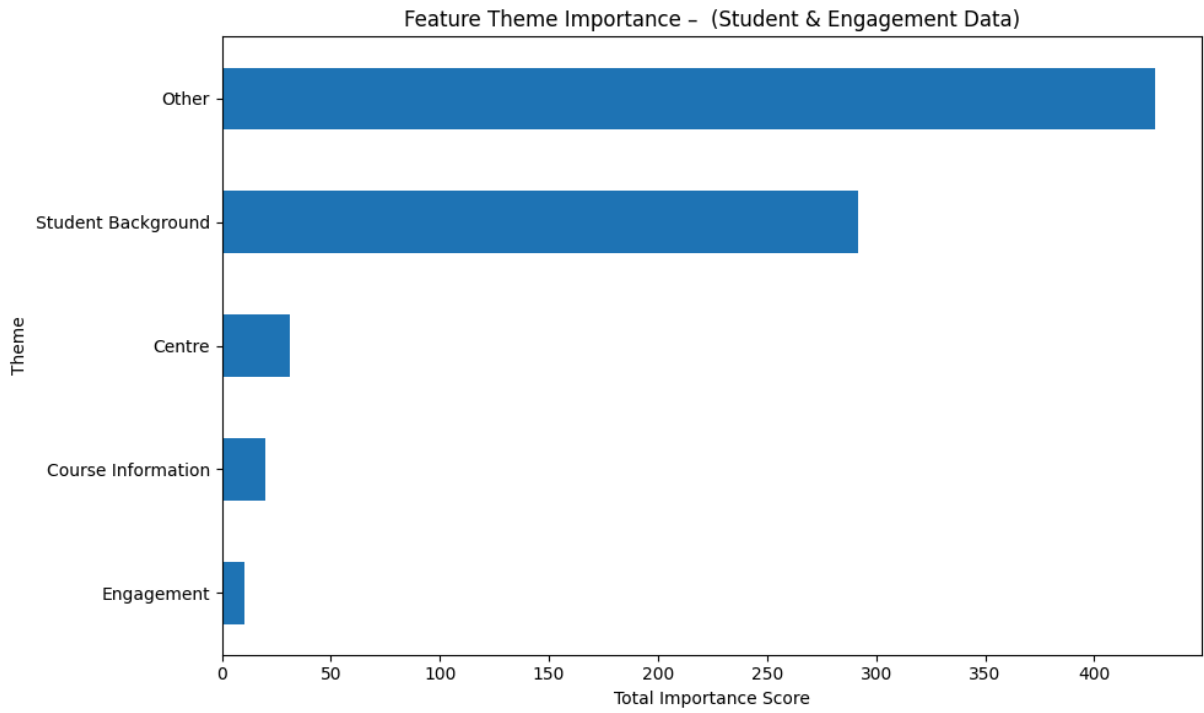For a model to be useful operationally, staff must understand what drives the predictions.

XGBoost performed best here, highlighting:

- Attendance patterns
- Academic engagement
- Module outcomes
- Certain demographic and course-choice factors

Neural networks also surfaced similar patterns, but explanations were less transparent.



Feature Theme Importance – (Application & Course Information)

## Feature Theme Importance – (Student & Engagement Data)



## Feature Theme Importance – (Academic Performance Data)

# 5. Key Findings

### 1. Stage 1 Data Provides Early Risk Signals

Even before a course begins, certain background and course-selection factors correlate with higher dropout likelihood.
 This stage gives a **baseline risk profile**.

### 2. Stage 2 Engagement Is the Most Actionable Window

Attendance—especially unauthorised absence—emerged as one of the strongest predictors.
 This is the **ideal point for early intervention**, as students are still in the early weeks of study.

### 3. Stage 3 Academic Performance Makes Dropout Highly Predictable

Once assessment outcomes are available, the model becomes extremely accurate (≈99%).
 However, by this stage some students may already be close to withdrawing, making earlier stages more important operationally.

### 4. XGBoost Is the Most Suitable Model for the Business

It consistently outperformed neural networks in:

- Accuracy
- Ability to detect at-risk students
- Interpretability

Its clarity makes it ideal for implementation in dashboards or early-warning systems.

# 6. What this Means for Student Support Teams

**Earlier, More Targeted Interventions**

Support staff can focus outreach on students flagged at:

- **Stage 1** — onboarding and pre-course communication
- **Stage 2** — attendance follow-ups and engagement support
- **Stage 3** — academic support, tutoring, reassessment preparation

**A Data-Driven Dropout Prevention Strategy**

The model can be embedded into:

- CRM systems
- Attendance dashboards
- Student support workflows

This ensures high-risk students are automatically identified and prioritised.

# 7. Recommendations

**Short-Term**

- Build a dashboard showing real-time risk scores per student
- Trigger automated alerts based on attendance patterns
- Implement early checks for high-risk Stage 1 profiles

**Medium-Term**

- Integrate predicted risk into academic advisor workflows
- Develop tailored intervention plans based on key predictors

**Long-Term**

- Expand dataset with wellbeing, support usage, and tutor feedback
- Continuously retrain the model with new student cohorts
- Use insights to redesign programmes or onboarding processes