Predicting Student Dropout Through Machine Learning

Table of Contents

1. Executive Summary

Student dropout remains a challenge for higher education institutions, affecting financial stability, academic performance, and student outcomes. This project applies machine learning to predict dropout risk, allowing universities to take proactive, data-driven interventions.

1.1 Key Findings

- **Contact hours** was the strongest predictor of dropout, reinforcing the importance of direct student engagement.
- **Unauthorised absences and academic performance** were also highly predictive, ranking higher than attendance percentage.
- **Attendance percentage was still relevant** but was less influential than other engagement and academic features.
- **Machine learning models achieved over 96% accuracy**, providing a reliable foundation for early identification of at-risk students.
- **XGBoost performed best in terms of interpretability and precision**, while **neural networks provided higher recall**, making them suitable for different intervention strategies.

These insights can help universities develop targeted intervention strategies that address attendance patterns and academic performance to improve student retention.

2. Business Problem and Context

2.1 The Cost of Student Dropout

Student dropout has significant financial and reputational consequences for universities, including:

- Lost tuition revenue from students who do not complete their studies.
- Reduced government funding, as many education grants are tied to student completion rates.
- Lower university rankings, impacting future student recruitment.

For students, dropout can lead to diminished career opportunities and wasted educational investment.

2.2 Why Data-Driven Solutions Matter

Traditional dropout prevention strategies rely on manual intervention and academic records alone, often identifying at-risk students too late. A predictive model allows institutions to:

- Detect students at risk of dropping out earlier.
- Develop data-driven intervention strategies instead of reactive measures.
- Allocate student support resources more effectively.

3. Data and Methodology

3.1 Data Sources and Key Features

The dataset included three main categories of features:

- **Demographics:** Age, gender, university affiliation.
- **Academic Performance Indicators:** Credit-weighted average, course level.
- **Engagement and Attendance Metrics:** Contact hours, attendance percentage, and unauthorised absences.

3.2 Analytical Approach

1. Exploratory data analysis (EDA) to identify trends, missing values, and class imbalances.
2. Feature engineering to standardise numerical features and encode categorical variables.
3. Model selection:
    - **XGBoost** for structured data and interpretability.
    - **Neural networks** to capture complex relationships between features.

3.3 Evaluating Model Performance

Models were assessed using:

- **Accuracy** – Overall correctness of predictions.
- **Precision** – The ability to minimise false dropout alerts.
- **Recall** – The ability to correctly identify students at risk.
- **AUC (Area Under Curve)** – A measure of how well the model distinguishes between students who complete and those who drop out.

4. Key Insights

4.1 Contact Hours Was the Most Important Predictor

Students with fewer contact hours were significantly more likely to drop out.

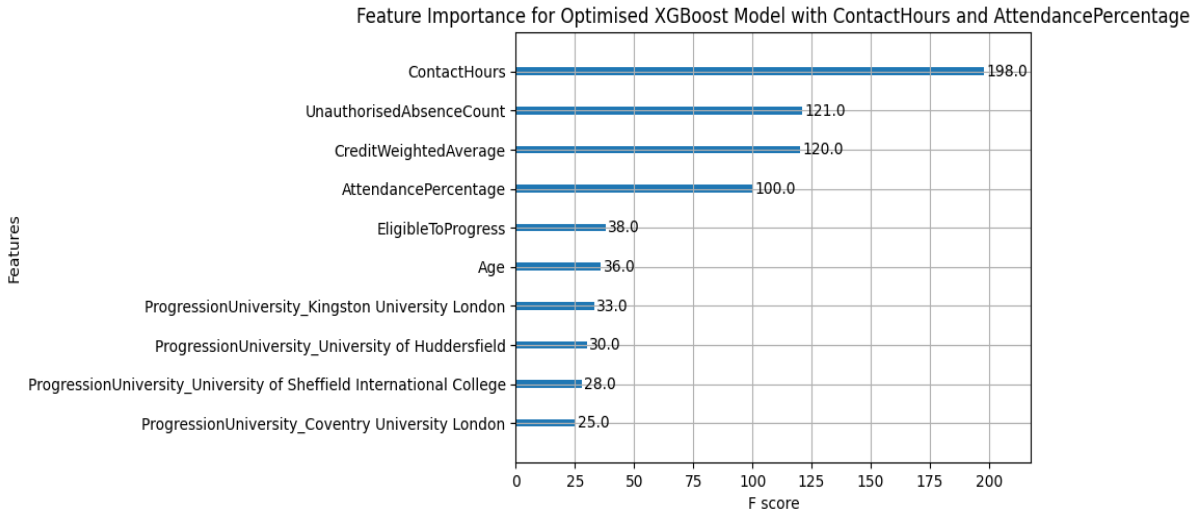

Figure 1: Feature Importance for XGBoost Model with Contact Hours and Attendance Percentage

This feature ranked highest in XGBoost's feature importance, reinforcing that direct engagement is a major factor in retention.

4.2 Unauthorised Absences and Academic Performance Were Also Key Indicators

Students with higher unauthorised absences were at greater risk of dropout.

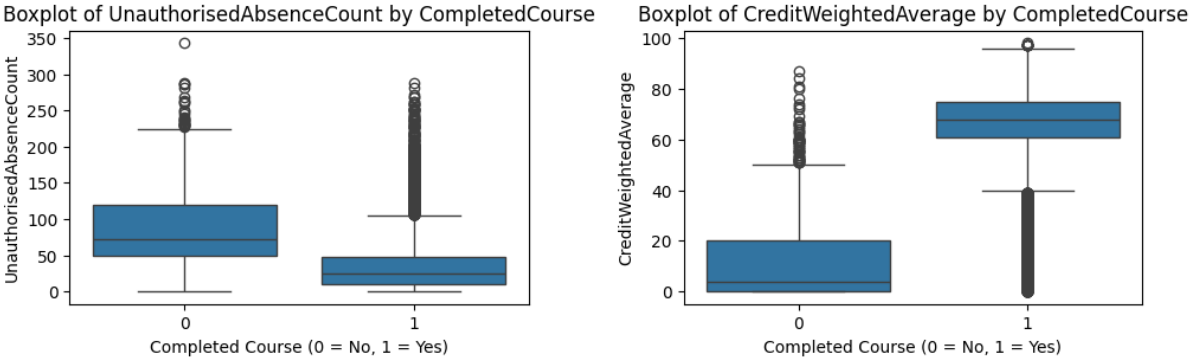

Figure 2: Boxplots of Key Features by Course Completion

Credit-weighted average was also highly predictive, indicating that academic struggles correlate strongly with dropout risk.

4.3 Attendance Percentage Was a Lower-Ranking Predictor

While still relevant, attendance percentage ranked below contact hours, unauthorised absences, and credit-weighted average.

5. Model Performance and Comparison

Model	Accuracy	Precision	Recall	AUC
XGBoost (No Attendance Features)	94.01%	99.65%	93.95%	0.9443
Neural Network (No Attendance Features)	92.95%	99.67%	92.80%	0.9404
XGBoost (With Attendance Features)	95.81%	99.78%	95.75%	0.9624
Neural Network (With Attendance Features)	96.43%	99.69%	96.50%	0.9589

Incorporating attendance features improved both models, increasing accuracy and recall, particularly for the Neural Network, which saw recall rise from 92.80% to 96.50%. XGBoost remained the best choice for interpretability and precision, while Neural Networks, with higher recall, were better at capturing at-risk students.

6. Strategic Recommendations

6.1 Prioritise Contact Hours and Unauthorised Absence Monitoring

- Develop real-time attendance tracking alerts based on contact hours.
- Identify students with high unauthorised absences for early interventions.

6.2 Address Academic Struggles in Conjunction with Attendance Issues

- Provide additional support for students with low credit-weighted averages.
- Implement mentorship programs and targeted academic assistance for struggling students.

6.3 Deployment Strategy

- Integrate the predictive model into student management systems for automated risk assessment.
- Pilot the intervention system to validate its effectiveness before full deployment.
- Expand data collection to include engagement metrics from online learning platforms.

7. Conclusion and Next Steps

7.1 Key Takeaways

- Contact hours was the strongest predictor of dropout.
- Unauthorised absences and academic performance were also highly influential, ranking above attendance percentage.
- XGBoost provided strong interpretability, while neural networks excelled in recall.

7.2 Next Steps for Implementation

- Deploy the model in student systems to provide early risk detection.
- Test intervention strategies and measure their impact on retention rates.
- Expand predictive capabilities by incorporating additional engagement metrics.

By implementing these recommendations, universities can take a proactive approach to student retention, reducing dropout rates and improving long-term academic success.