APPLYING SUPERVISED LEARNING TO PREDICT STUDENT DROPOUT

Contents

Introduction

Student retention is critical for educational institutions, impacting financial sustainability and academic success. High dropout rates can lead to revenue losses and reputational damage. Study Group, a global education provider, aims to enhance student success by identifying at-risk students early and implementing proactive interventions. This study applies supervised machine learning techniques to predict dropout risks, enabling Study Group to refine its support strategies and improve student retention.

Data and Methodology

This project follows a structured, three-stage data approach, reflecting Study Group's real-world student journey:

1. **Applicant and Course Information** – Demographic and course-related details available at enrollment.
2. **Student Engagement Data** – Attendance and engagement metrics reflecting real-time student participation.
3. **Academic Performance Data** – Assessment results, including passed and failed modules.

The results of this project will allow Study Group to develop strategic retention policies and improve student outcomes.

Stage 1: Applicant and Course Information

The dataset contained **25,059 observations** and **16 features**, with missing values in four features:

- **DiscountType (70%)**, **HomeState (64%)**, **HomeCity (14%)**, **ProgressionDegree (3%)**

- Given their limited relevance, features with over **50% missing data** or high cardinality (>200 unique values) were removed.
- The final dataset contained **11 features** with all missing values eliminated.

Neural Network Performance:

Neural Network Model: The Neural Network (NN) model was also evaluated in Stage 1, using Keras Tuner for hyperparameter optimization. Initial training used default hyperparameters, followed by tuning via **RandomizedSearchCV** to optimize the learning rate, batch size, and number of layers.

Parameter type	Accuracy	Auc	Precision	Recall
Default	0.9	0.88	0.92	0.97
Tuned	0.9	0.88	0.92	0.96

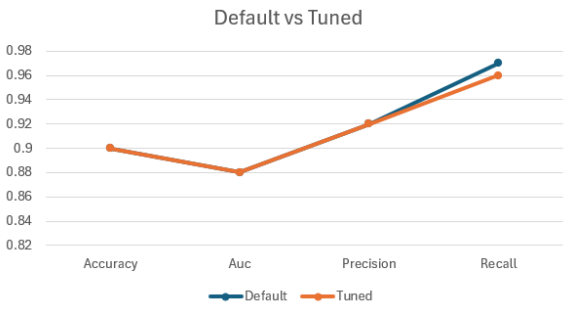

Figure 1 NN Default vs Tuned

- Accuracy: **0.90** (default and tuned models)
- AUC: **0.88**
- Precision: **0.92** (default and tuned models)
- Recall: **0.97 (default), 0.96 (tuned)**

Default			Tuned	
408	367		418	357
145	4092		152	4085

Figure 2 NN Confusion Matrix

Minimal differences between default and tuned models suggest that additional engagement or academic data is necessary to enhance predictive power. While the NN model performed well,

the results indicated that demographic features alone were not strong predictors of dropout risk.

Model Selection and Evaluation:

The **XGBoost classifier** was chosen due to its efficiency with structured data. Initial training used default hyperparameters, followed by hyperparameter tuning using **RandomizedSearchCV** (optimizing **learning_rate, max_depth, and n_estimators**).

Performance metrics were:

Parameter type	Accuracy	Auc	Precision	Recall
Default	0.9	0.89	0.93	0.96
Tuned	0.9	0.89	0.92	0.96

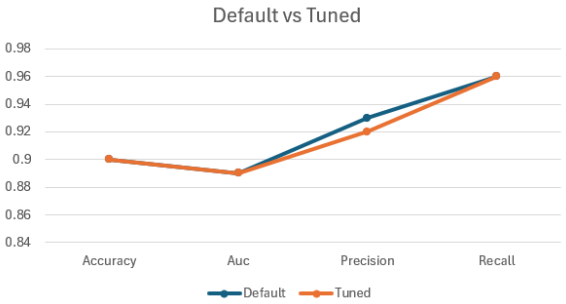

Figure 3 XGBoost Default Vs Tuned

- **Accuracy: 0.90** (default and tuned models)
- **AUC: 0.89**
- **Precision: 0.93 (default), 0.92 (tuned)**
- **Recall: 0.96**

Default			Tuned	
446	329		441	334
159	4078		157	4080

Figure 4 Confusion matrix

Minimal differences between default and tuned models suggest that the default settings were near-optimal. Feature importance analysis showed demographic factors had limited predictive power, making this data alone insufficient for dropout prediction. This indicates that institutions

like Study Group must look beyond demographic data to drive effective student retention policies.

Stage 2: Student Engagement Data

In Stage 1, dropout predictions were based solely on applicant and course information, yielding moderate predictive performance. Stage 2 enhances this analysis by incorporating student engagement data, specifically attendance records, to assess whether real-time participation serves as a stronger predictor of dropout risk. This comparison between Stage 1 and Stage 2 results demonstrates the added value of engagement-related features in improving predictive accuracy.

Two new features were introduced:

- **AuthorisedAbsenceCount** and **UnauthorisedAbsenceCount**, both with **1% missing values**, which were imputed using the median.

Comparison of Stage 1 vs. Stage 2 Results

stage 1 xgboost

Parameter type	Accuracy	Auc	Precision	Recall
Default	0.9	0.89	0.93	0.96
Tuned	0.9	0.89	0.92	0.96

stage 2 xgboost

Parameter type	Accuracy	Auc	Precision	Recall
Default	0.91	0.92	0.93	0.97
Tuned	0.91	0.92	0.93	0.97

stage 1 nn

Parameter type	Accuracy	Auc	Precision	Recall
Default	0.9	0.88	0.92	0.97
Tuned	0.9	0.88	0.92	0.96

stage 2 nn

Parameter type	Accuracy	Auc	Precision	Recall
Default	0.9	0.9	0.92	0.97
Tuned	0.91	0.91	0.93	0.97

Figure 5 Sage 1 Vs Stage 2

XGBoost Model:

- **Accuracy:** Increased from **0.90 (Stage 1)** to **0.91 (Stage 2)**
- **AUC:** Improved from **0.89** to **0.92**, demonstrating better model discrimination.
- **Precision:** Remained consistent at **0.93**.

- **Recall:** Increased slightly from **0.96** to **0.97**, suggesting better identification of at-risk students.

Neural Network Model:

- **Accuracy:** Increased from **0.90 (Stage 1)** to **0.91 (Stage 2)**
- **AUC:** Improved from **0.88** to **0.91**, showing enhanced dropout classification ability.
- **Precision:** Increased slightly from **0.92** to **0.93**.
- **Recall:** Remained stable at **0.97**, confirming the model's strong ability to detect dropouts.

The improvement in AUC across both models suggests that student engagement data provides a meaningful boost to prediction accuracy. By tracking attendance, institutions can identify at-risk students earlier and offer targeted interventions such as academic counseling or mentorship programs.

From a business perspective, Study Group benefits from this analysis by developing strategies to minimize student disengagement. Lower dropout rates directly contribute to increased course completion rates, improved student satisfaction, and enhanced institutional reputation. The findings underscore the importance of integrating attendance tracking into retention strategies, allowing Study Group to take proactive measures in reducing dropout rates while maintaining financial stability.

Stage 3: Academic Performance Data

Three new features were added:

- **AssessedModules**, **PassedModules**, **FailedModules** (each with **9% missing values**, leading to row removal).

XGBoost and Neural Network Results:

	xgboost			
Parameter type	**Accuracy**	**Auc**	**Precision**	**Recall**
Default	0.99	1	0.99	1
Tuned	0.99	1	0.99	1
	nn			
Parameter type	**Accuracy**	**Auc**	**Precision**	**Recall**
Default	0.99	1	0.99	0.99
Tuned	0.99	1	0.99	1

Figure 6 XGBoost Vs NN

- Accuracy improved to **0.99**
- AUC reached **1.00**, indicating perfect classification.
- Precision remained high at **0.99**.
- Recall improved to **1.00**, ensuring no dropout cases were missed.

These near-perfect results suggest that academic performance data is the strongest predictor of dropout risk. Study Group can use this insight to offer targeted support to struggling students before they reach a critical dropout threshold. Academic performance data provides a clear signal of students at risk, allowing institutions to allocate resources effectively. Implementing additional support measures such as tutoring, academic counseling, and mentorship programs can significantly reduce dropout rates.

Comparison of Model Performance

Across all three stages, model performance improved as additional data was incorporated:

1. **Applicant and Course Data:** Limited predictive power due to demographic constraints.
2. **Student Engagement Data:** Moderate improvement in precision and recall, highlighting attendance as a key factor.

3. **Academic Performance Data:** Significant accuracy and AUC gains, confirming that academic success is the most reliable dropout predictor.

These findings align with Study Group's objectives by identifying the best intervention points for student retention and ensuring students receive support when needed most. The insights generated from this project can inform policy changes, optimize resource allocation, and improve overall student satisfaction and institutional performance.

Conclusion

By adopting these strategies, Study Group can enhance student retention, financial stability, and institutional reputation. Machine learning models like XGBoost and neural networks offer scalable dropout prevention solutions. Future research can refine predictions by incorporating socioeconomic and mental health factors. Testing interventions based on model insights will ensure long-term improvements, reinforcing Study Group's mission of providing quality education globally.

Appendix

Business Impact and Recommendations

This study confirms that dropout risk can be predicted with increasing accuracy as more data becomes available. **Early-stage indicators (demographics) are weak predictors**, but **engagement and academic performance data significantly improve model accuracy**.

To enhance student retention, Study Group should:

1. **Use engagement data for early interventions** – Implement real-time monitoring of attendance trends and provide support to students who frequently miss classes.
2. **Leverage academic performance data for targeted support** – Identify students struggling academically and offer tutoring or mentoring services to improve their performance.
3. **Develop automated alert systems** – Integrate predictive analytics into student support services to flag at-risk students proactively and allow staff to intervene.
4. **Enhance faculty-student engagement** – Encourage faculty to identify students exhibiting early signs of disengagement and facilitate personalized support plans.
5. **Expand socio-economic data integration** – Consider incorporating external factors, such as financial difficulties or mental health indicators, to refine predictions and provide holistic support.