LEVERAGING MACHINE LEARNING TO PREDICT DROPOUT LIKELIHOOD IN HIGHER EDUCATION

RAQUEL ANA BUSH (8), BRIAN KADE BETTERTON (7)

Introduction

Student dropout is a persistent challenge in higher education institutions worldwide. The decision to leave school prematurely can have profound academic, emotional, and financial consequences for students, while also impacting the institutions themselves through lost tuition revenue, reduced graduation rates, and diminished reputation. Early identification of at-risk students enables timely intervention and academic support, improving graduation rates and educational outcomes [1]. As universities seek to improve student retention and support services, data-driven approaches have emerged as powerful tools for identifying students at risk of disengaging before it's too late.

In this project, we analyze the factors that most strongly influence student persistence and explore the use of machine learning (ML) to predict student dropout risk based on a rich dataset of demographic, academic, and enrollment information from a Portuguese higher education institution [2] [3]. By training four classification models, including XGBoost, Decision Tree, Random Forest, and Logistic Regression, we aim to determine whether reliable predictions can be made. We consider the sensitivity, specificity, precision, and accuracy of each model, intending to maximize early intervention opportunities for at-risk students.

With appropriate implementation, we believe that such predictive models could be integrated into academic advising systems, enabling tailored outreach and resource allocation that promotes student success.

Problem Statement

Higher education institutions face significant challenges in retaining students through to graduation. Dropout rates remain a concern across universities and colleges globally. These early departures not only hinder students' academic and career trajectories but also result in financial losses and reputational impacts for the institutions themselves.

One of the most effective strategies to address this issue is early identification and support of students who are at risk of dropping out. Traditional methods, such as academic advising and manual flagging systems, often lack the scalability and precision required to proactively intervene for large student populations. Machine learning offers a promising alternative by automating the identification process using patterns found in existing data.

We seek to develop and evaluate machine learning models that can predict whether a student is at risk of dropping out based on features including demographics, early academic performance, enrollment behavior, and familial background. Additionally, we aim to understand which features most significantly contribute to

a student's likelihood of dropout. By highlighting key predictive indicators, our analysis can guide future interventions and help institutions design more effective, data-informed retention strategies.

Dataset Description

The dataset that we used in this project was obtained from a Portuguese higher education institution and contains 4,424 student records with 36 feature variables. These features span multiple domains, including:

- **Demographic characteristics**: age at enrollment, gender, nationality, and parental education and occupation.
- **Academic background**: admission grade, previous qualification, grades in the first and second semesters, course enrollment information.
- **Socioeconomic indicators**: scholarship holder status, whether tuition fees are up to date, displacement status, and macroeconomic variables such as GDP and unemployment rate.
- **Enrollment behavior**: application mode and order, attendance type (daytime/evening), and whether the student is an international or displaced student.

The target variable originally consisted of three categories:

- 0 = Dropout
- 1 = Enrolled
- 2 = Graduate

For binary classification purposes, the categories were regrouped as follows:

- 1 = Dropout
- 0 = Enrolled/Graduated

This transformation allowed us to train models focused specifically on predicting dropout risk.

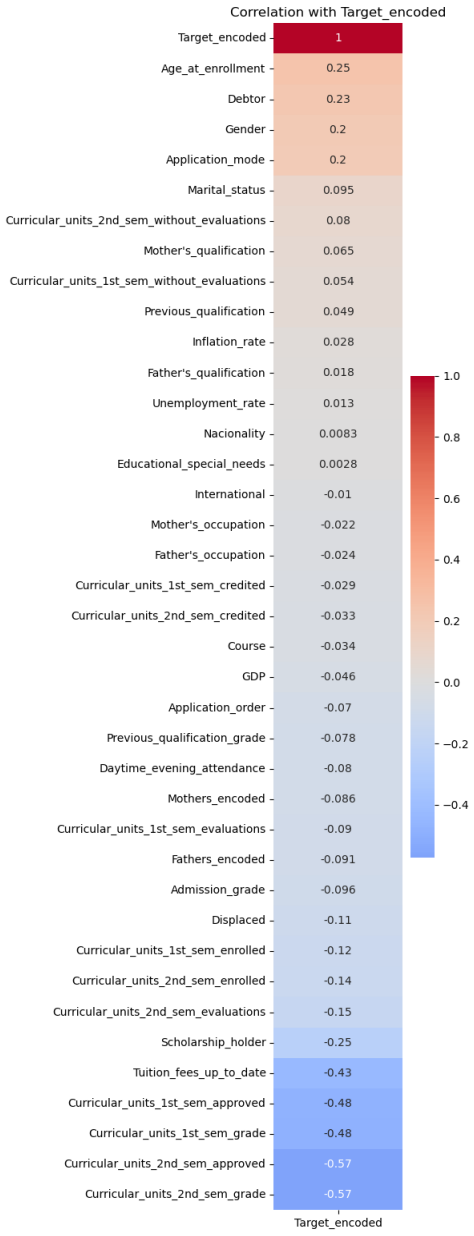


FIGURE 1. Feature correlation with target label (Dropout = 1). Strongest predictors are grades during the first two semesters and tuition payment status.

Figure 1 displays the correlation of each feature with the encoded target variable. Notably, first and second semester performance and tuition payment status show the strongest negative correlations with dropout, suggesting high predictive power.

To further visualize the separation between dropout and non-dropout students, we analyzed the distributions of grades and enrollment age using boxplots:

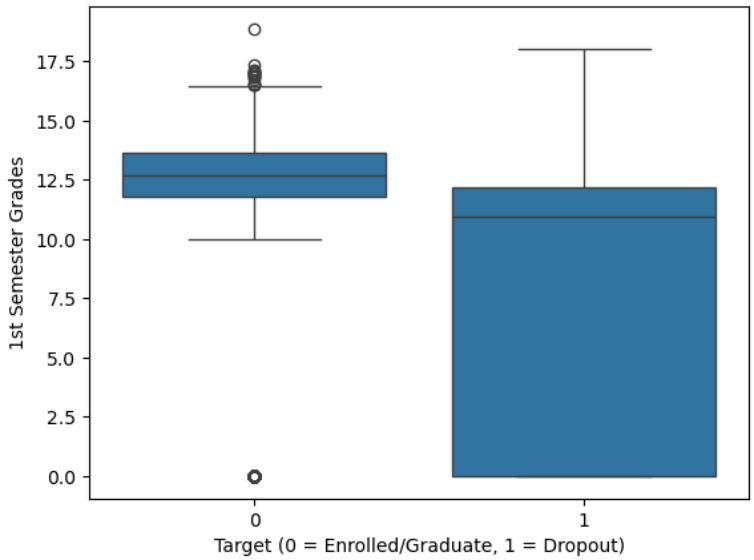


FIGURE 2. Distribution of 1st Semester Grades by Target Outcome

In Figure 2, we can see that students who were still enrolled or graduated by the end of the typical course duration had consistently higher first semester grades, with most scores tightly grouped around a median of around 12.5. Dropout students showed a wider spread of grades, including many failing grades, which strongly suggests that poor performance or disengagement during the first semester is a strong predictor of dropout.

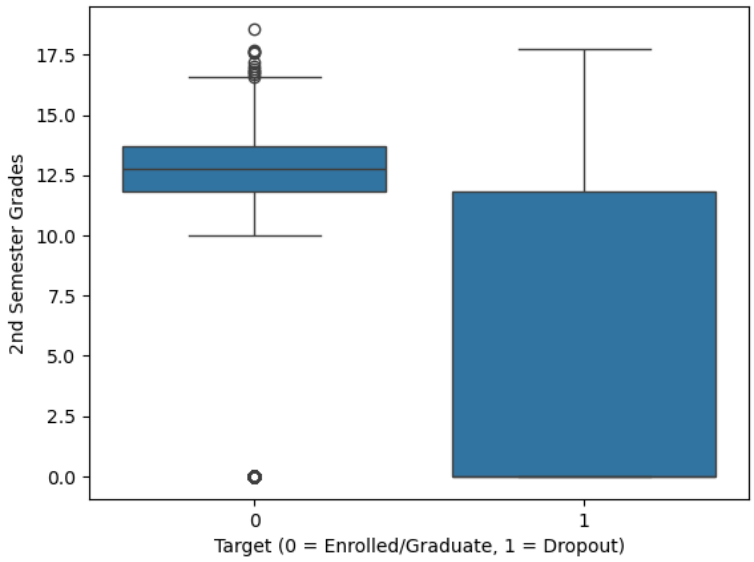


FIGURE 3. Distribution of 2nd Semester Grades by Target Outcome

Like the first semester, students who stayed enrolled had consistently higher second semester grades, as seen in Figure 3. Dropouts again had lower and more varied grades, with many students failing to achieve passing marks. Together, these two grade plots reinforce that academic performance, particularly early on, is a strong signal of retention.

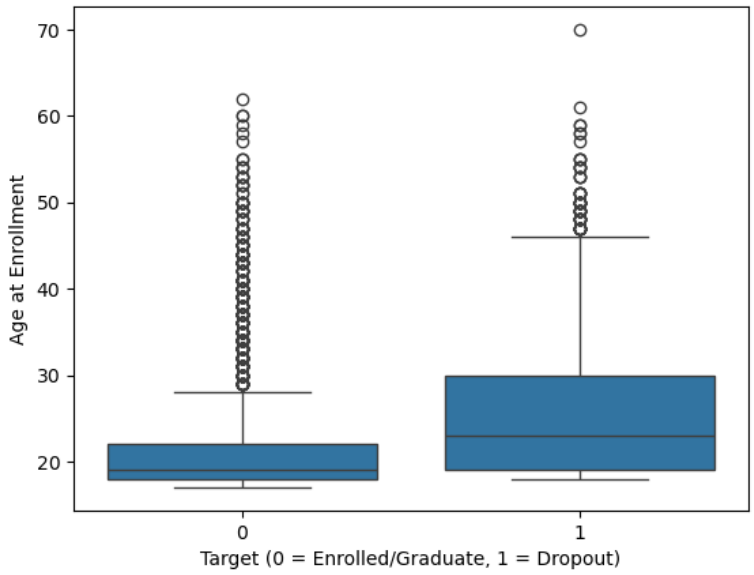


FIGURE 4. Distribution of Age at Enrollment by Target Outcome

Figure 4 also reveals that dropout students were, on average, slightly older at the time of enrollment. This may potentially be due to increased outside responsibilities older students tend to have in their lives.

Data Preprocessing

Prior to model training, we performed several preprocessing steps to ensure the dataset was clean, structured, and suitable for supervised learning. These steps included the following:

- **Missing Values:** The dataset did not contain any missing values, so we did not need to handle that.
- **Target Label Encoding:** Like mentioned above, the original target variable had three classes: `0 = Dropout`, `1 = Enrolled`, and `2 = Graduate`. We recoded these into a binary classification task where `1 = Dropout` and `0 = Enrolled/Graduated`. This simplified the problem into identifying whether or not a student dropped out.
- **Categorical Feature Encoding:** We encoded certain categorical features to facilitate model compatibility. For example, parental occupations were grouped into broader categories such as `white collar`, `blue collar`, `military`, `services`, etc. All string-based categorical variables were either label encoded or one-hot encoded depending on the model requirements.
- **Feature Selection:** After our exploratory correlation analysis, we dropped no features.
- **Normalization and Scaling:** We scaled numerical features, particularly grades and age, to standardize value ranges. This step is especially important for algorithms sensitive to magnitude, such as Logistic Regression.
- **Train-Test Split:** The dataset was split into training and test sets to evaluate model generalization. A typical 80/20 split was used.

These preprocessing steps helped ensure the dataset was suitable for training a range of machine learning classifiers.

Machine Learning Pipeline

Our machine learning pipeline followed a structured and iterative approach to model development, training, and evaluation to address the business problem of identifying students in need of support to prevent dropout. The key stages are outlined below:

(1) **Problem Formulation:** The initial step involved clearly defining the objective: to predict whether a student will drop out based on their demographic, academic, and socioeconomic data. We framed this as a binary classification problem.
(2) **Data Collection:** We obtained the dataset from the UCI Machine Learning Repository, and it is from a public higher education institution in Portugal and consists of 4,424 instances and 36 features.
(3) **Feature Engineering:** We grouped some categorical variables (e.g., parental occupation) into broader categories to reduce dimensionality.
(4) **Data Evaluation:** We evaluated our data, looking at various feature column names and their categories or values, as well as the shape of the dataset. We looked for missing values we would have needed to handle but

there were none. We also generated density plots, boxplots, and correlation values.

(5) **Model Selection and Training:** Four supervised learning algorithms were selected for comparison:
   - **XGBoost Classifier (XGBClassifier):** Chosen due to the finding in a comparative study by Villar and de Andrade that it has high potential as a valuable tool for "enhancing educational research and improving student outcomes," due to what they found was consistently superior performance in predicting student dropout. They stated that it outperformed traditional classification methods consistently [4].
   - **Decision Tree Classifier:** Chosen as a baseline tree-based model.
   - **Random Forest Classifier:** Another ensemble method, like XGBoost, however trees are grown independently in parallel rather than sequentially.
   - **Logistic Regression (via Statsmodels):** A simple yet effective linear model, chosen to analyze the impact of changing decision thresholds.

(6) **Evaluation:** We evaluated the models using sensitivity, specificity, precision, and accuracy metrics. Additionally, we visualized performance using confusion matrices and AUC-ROC curves.

(7) **Model Tuning:** Hyperparameters for the three former models were tuned using random search tuning. We found no significant improvement in comparison from grid search tuning.

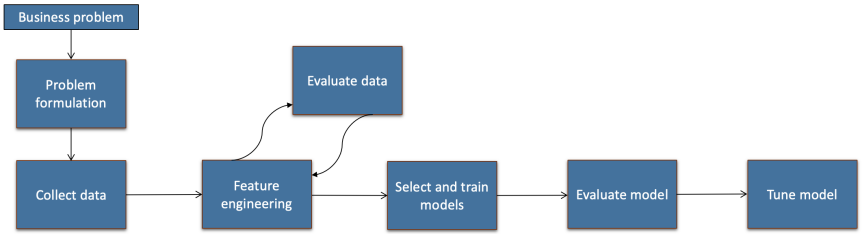


FIGURE 5. Overview of the Machine Learning Pipeline

Figure 5 summarizes the end-to-end process that we followed throughout this project.

Model Training and Evaluation

Each model was initially evaluated using default hyperparameters, followed by tuning with the aim of improving performance, particularly sensitivity, which we prioritized to maximize the identification of at-risk students.

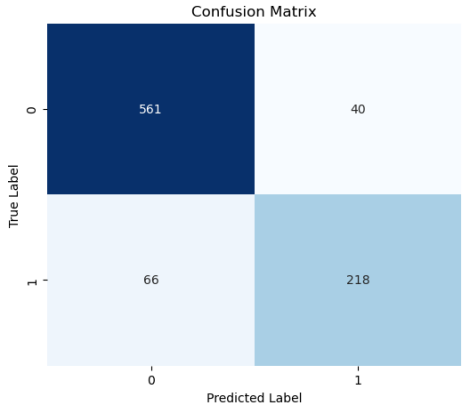
TABLE 1. Model Metrics Before Hyperparameter Tuning (all metrics in %)

Model	**Sensitivity**	**Specificity**	**Precision**	**Accuracy**
XGBoost	76.76	93.34	84.50	88.02
Decision Tree	73.59	94.51	86.36	87.80
Random Forest	72.18	95.51	88.36	88.02
Logistic Regression	73.59	96.01	89.70	88.81

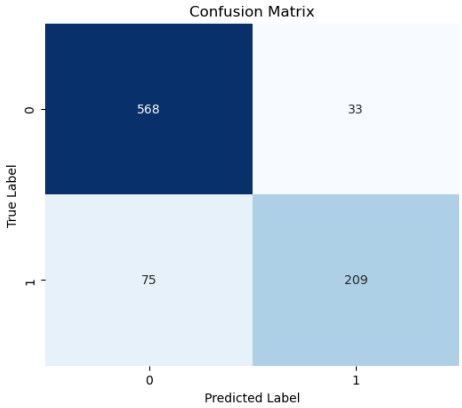
TABLE 2. Model Metrics After Hyperparameter Tuning (all metrics in %)

Model	**Sensitivity**	**Specificity**	**Precision**	**Accuracy**
XGBoost	73.59	94.18	85.66	87.57
Decision Tree	63.38	96.01	88.24	85.54
Random Forest	73.59	94.51	86.36	87.80
Logistic Regression	89.08	81.53	69.51	83.95

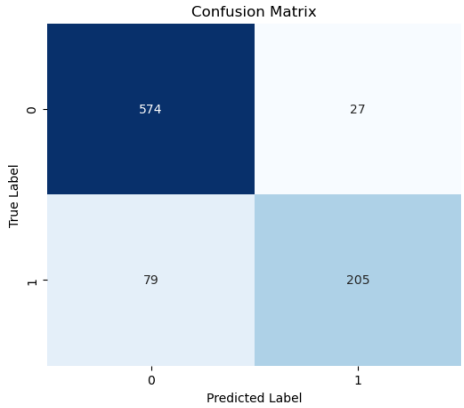
* Logistic Regression used a lowered decision threshold of 0.2 instead of the default 0.5 to increase sensitivity.



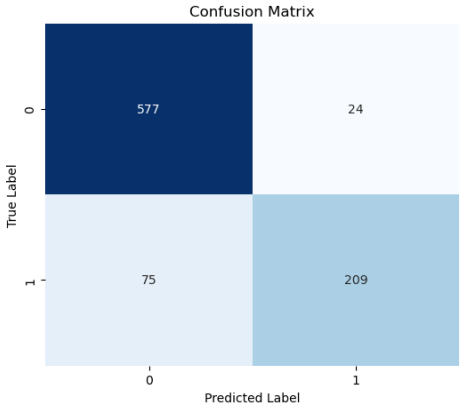
(A) XGBoost



(B) Decision Tree

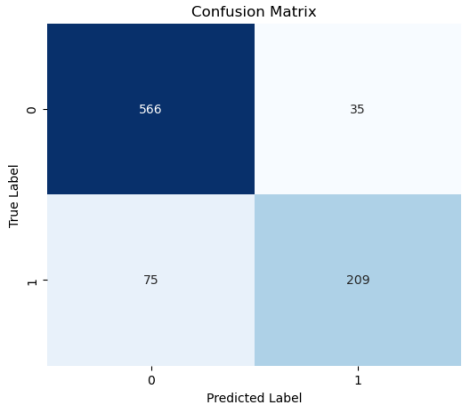


(C) Random Forest

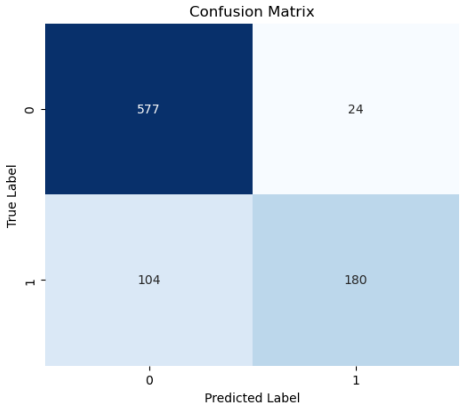


(D) Logistic Regression

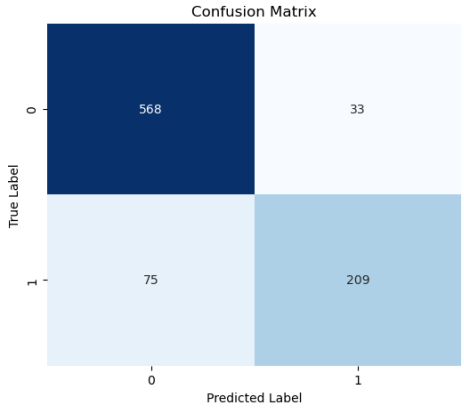
FIGURE 6. Confusion Matrices Before Tuning



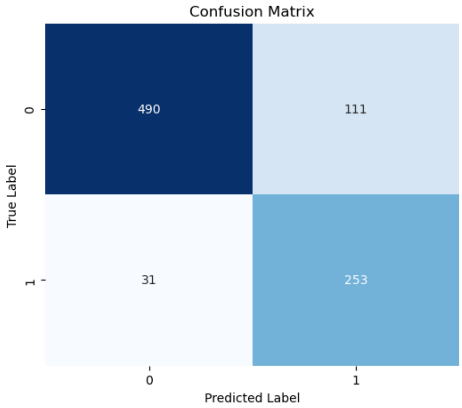
(A) XGBoost



(B) Decision Tree

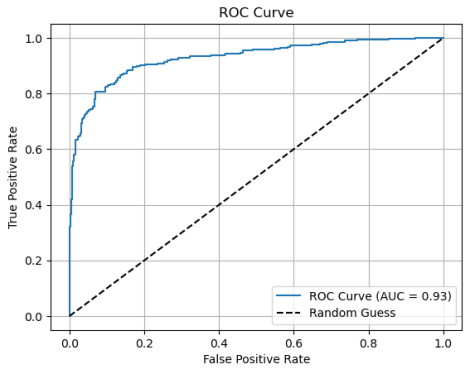


(C) Random Forest

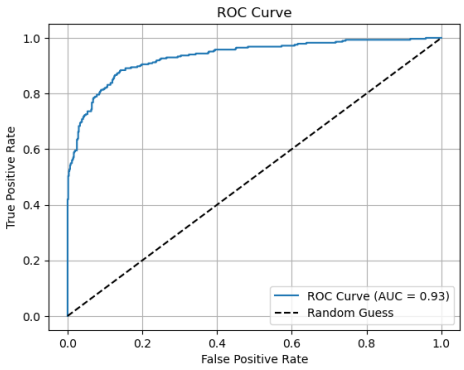


(D) Logistic Regression

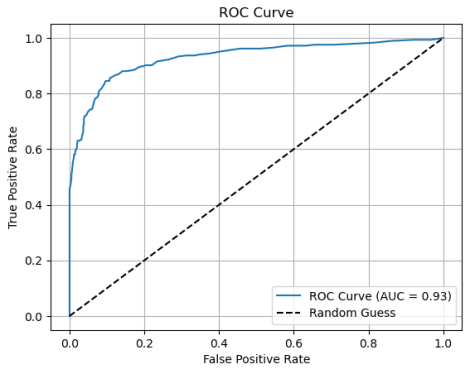
FIGURE 7. Confusion Matrices After Tuning



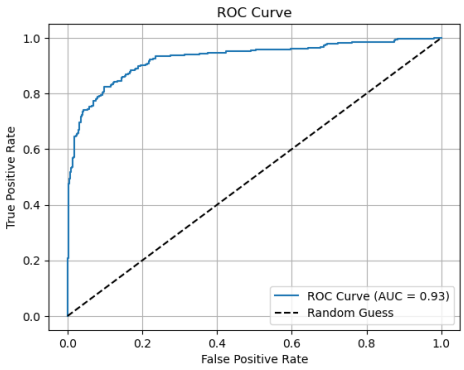
(A) XGBoost



(B) Decision Tree

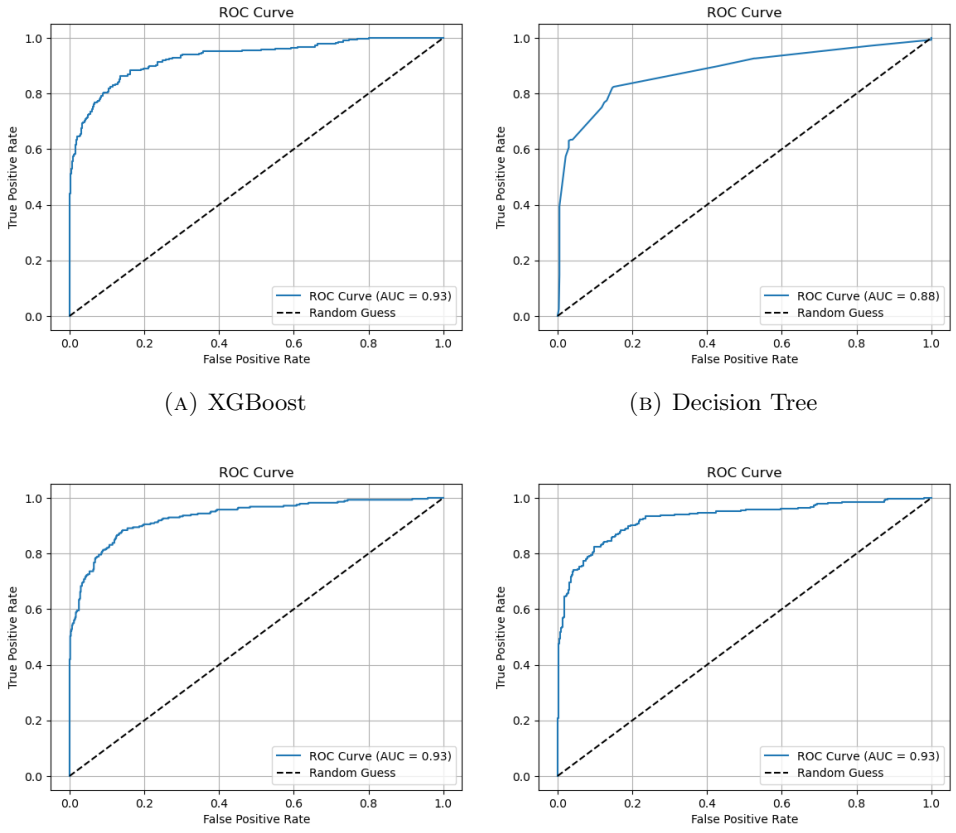


(C) Random Forest



(D) Logistic Regression

FIGURE 8. AUC-ROC Curves Before Tuning



(A) XGBoost (B) Decision Tree

(C) Random Forest (D) Logistic Regression

FIGURE 9. AUC-ROC Curves After Tuning

Interpretation. We found that tuning did not significantly improve performance for the XGBoost, Decision Tree, or Random Forest models. In fact, these tended to slightly diminish in performance post-tuning. Logistic Regression, however, benefited from lowering the classification threshold to increase sensitivity, which is a critical consideration when the cost of missing an at-risk student outweighs that of a false positive. Random Forest and XGBoost demonstrated consistent balance between sensitivity and specificity, making them strong candidates for deployment in real-world academic support systems. The simple Decision Tree was unsurprisingly the weakest performer.

RESULTS INTERPRETATION

The performance metrics and visualizations reveal valuable insights into the behavior and strengths of each machine learning model. While all four models achieved reasonably high accuracy, they varied in their ability to detect dropout cases (sensitivity) versus students who remained enrolled or graduated (specificity).

XGBoost and Random Forest. XGBoost and Random Forest delivered the most balanced performance across all metrics. Both models maintained high specificity (above 94%) while achieving competitive sensitivity (around 73%). This balance suggests these ensemble methods are well-suited for identifying at-risk students without producing excessive false positives.

Decision Tree. The Decision Tree model exhibited a noticeable decline in sensitivity after tuning. While specificity increased slightly, the model's ability to correctly identify dropouts fell to 63.38%, the lowest among all models. This suggests that simple tree structures may struggle to generalize effectively in the presence of subtle or overlapping feature distributions.

Logistic Regression. Logistic Regression performed surprisingly well, particularly after we adjusted the decision threshold from the default value of 0.5 to 0.2. This shift significantly increased sensitivity to 89.08%, making it the most effective model for detecting dropouts. However, this improvement came at the cost of specificity, which dropped to 81.53%. This trade-off is meaningful: for institutional settings that prioritize catching as many at-risk students as possible, even if some false positives occur, this model may be the most appropriate.

Discussion

Overall, the results indicate that ensemble methods like XGBoost and Random Forest are strong baseline models for dropout prediction due to their consistent performance. However, when the goal is maximum dropout detection, Logistic Regression offers a compelling alternative.

Beyond predictive performance, understanding which features most significantly influence model decisions is crucial for actionable insights. By analyzing feature importance across models, we identified several consistent predictors of student dropout, which could inform early intervention strategies.

The most influential features across the ensemble models (XGBoost and Random Forest) included:

- **Curricular Units 1st and 2nd Semester Grade**: Strongly correlated with student retention. Low grades in either semester increased the likelihood of dropout; however, the second semester grades seemed to have a larger impact.
- **Tuition Fees Up to Date and Debtor**: Students with overdue tuition or debt were significantly more likely to drop out, suggesting financial difficulties as a critical risk factor.
- **Age at Enrollment**: Older students were slightly more likely to drop out, possibly due to competing responsibilities.

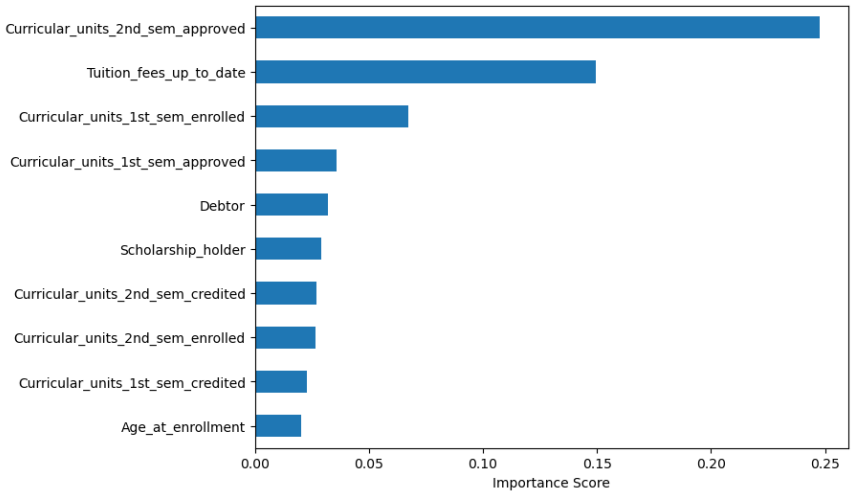


FIGURE 10. Top 10 Most Important Features from XGBoost

Figure 10 presents the top 10 features used by the XGBoost model. Academic performance indicators dominate the list, reaffirming that grades and course completion metrics are strong signals of student risk. However, tuition fees up to date was the second-highest on the list, showing that financial stability has a large impact.

By incorporating predictive analytics into existing student success infrastructure, institutions can better allocate resources and improve graduation rates.

We should note that some features, particularly those related to parental education, nationality, or socioeconomic status, may correlate with systemic inequality. While they improve predictive power, care must be taken to avoid reinforcing biases. Future models should undergo fairness audits to ensure that interventions guided by predictions do not disadvantage vulnerable student populations.

FUTURE WORK

While our current machine learning models show strong potential in predicting student dropout risk, there are several avenues for improvement and expansion in the future.

Validating our models on datasets from other institutions from various world regions and institution types (e.g., public versus private) would test their generalizability. If consistent, these models could be integrated into a real-time academic advising system, providing automated alerts and risk scores to inform proactive student support.

Future work could involve generating new features that capture deeper behavioral trends, such as changes in performance between semesters and interaction between financial and academic features.

Lastly, our dataset reflects student outcomes at a fixed endpoint. A more realistic modeling approach would involve tracking students over time and predicting dropout risk dynamically. Time-series models or recurrent neural networks (RNNs) could be explored for this purpose.

Conclusion

Ensemble methods XGBoost and Random Forest offered reliable performance consistent with the findings of Villar and de Andrade [4]. However, Logistic Regression, when the threshold was adjusted, delivered the highest sensitivity. These insights are crucial for real-world applications, where the priority is to catch as many at-risk students as possible, even at the expense of a few extra false positives.

Through analysis of feature importance, we identified key predictors such as first and second semester grades and tuition payment status.

Machine learning offers promising tools for improving student retention, provided they are used thoughtfully and responsibly. With continued refinement, such models could become a valuable component of academic support systems, helping institutions support their students more effectively and equitably.

References

[1] A.-M. Faria, N. Sorensen, J. B. Heppen, J. Bowdon, and R. Eisner, *Getting students on track for graduation: Impact of the Early Warning Intervention and Monitoring System after one year*, Society for Research on Educational Effectiveness, 2018. [Online]. Available: `https://api.semanticscholar.org/CorpusID:150549484`.

[2] M. V. Martins, D. Tolledo, J. Machado, L. M. T. Baptista, and V. Realinho, *Early prediction of student's performance in higher education: a case study*, in *Trends and Applications in Information Systems and Technologies*, vol. 1, Advances in Intelligent Systems and Computing, Springer, 2021. doi: 10.1007/978-3-030-72657-7_16.

[3] V. Realinho, M. Vieira Martins, J. Machado, and L. Baptista, *Predict Students' Dropout and Academic Success*, UCI Machine Learning Repository, 2021. [Online]. Available: `https://doi.org/10.24432/C5MC89`.

[4] A. Villar and C. R. V. de Andrade, *Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study*, Discover Artificial Intelligence, vol. 4, no. 2, 2024. [Online]. Available: `https://link.springer.com/article/10.1007/s44163-023-00079-z`.

Appendix

All code used in this analysis is provided in the attached Jupyter notebook titled `RaquelAna_Bush_ProjectCode.ipynb`.

Department of Computer Science, University of Massachusetts Dartmouth, North Dartmouth, Massachusetts 02747