PREDICTING STUDENT RETENTION IN HIGHER EDUCATION

Jane Zulu

August 1st, 2024
Source: https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention

EXECUTIVE SUMMARY

OBJECTIVE

Forecast student dropout rates and pinpoint crucial factors that impact retention and academic achievement.

APPROACH

Exploratory analysis and predictive modeling (Logistic Regression, KNN, XGBoost).

KEY FINDINGS

- Logistic Regression: 91% accuracy, high interpretability.
- KNN: 89% accuracy, effective in pattern detection.
- XGBoost: 89% accuracy, potential with further tuning.

CONCLUSION

- Prioritize Logistic Regression and KNN for immediate deployment.
- Consider refining XGBoost for complex interactions.
- Implement strategies based on model insights to improve retention.

Project Background

Three categories were initially included in the dataset (dropout, enrolled, and graduate) at the end of the normal duration of the course. The dataset was filtered to include only students who “dropout” or “graduate.”

39% DROPOUT

61% GRADUATE

The filtered dataset comprises 4,424 students enrolled in a higher education institution, gathered from different databases. These students are studying undergraduate programs in fields like agronomy, design, education, nursing, journalism, management, social service, and technologies.

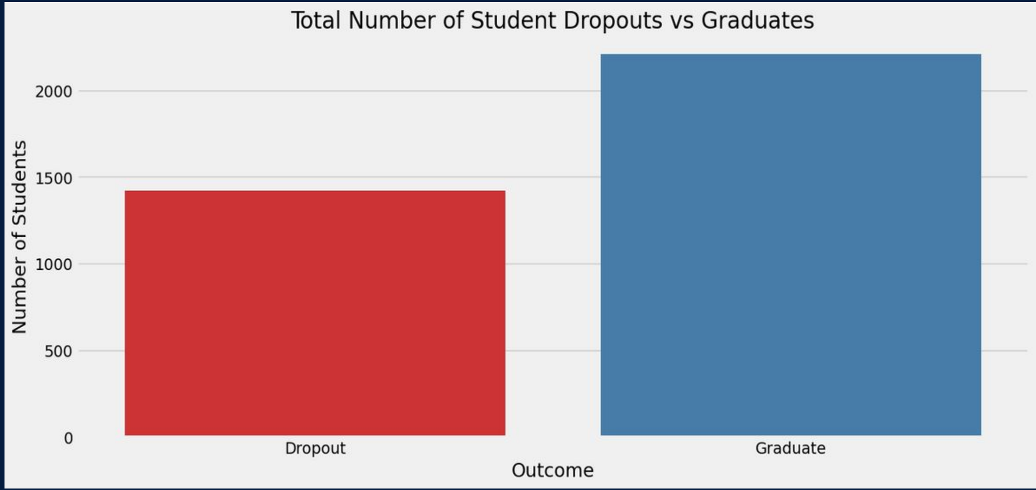

Figure 1.1 - Barchart of students that dropout vs students that graduate

Objectives & Data Exploration

To forecast student outcomes, various factors affecting student retention were considered.

52.8% — **Academic Performances**
- Curricular Units
- Admission Grade

30.6% — **Demographics**
- Age at enrollment (17 – 70)
- Gender
- Nationality

16.7% **Socio -Economics**
- Unemployment rate
- Tuition fees up to date
- Debtor

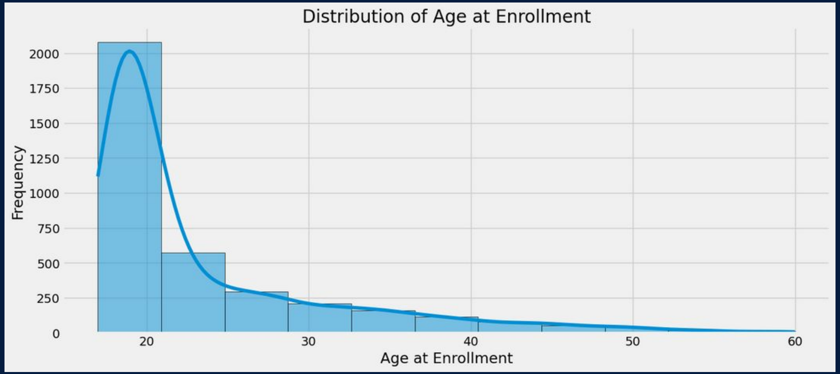

Figure 1.2 - Distribution of students age at enrollment into their Undergraduate program

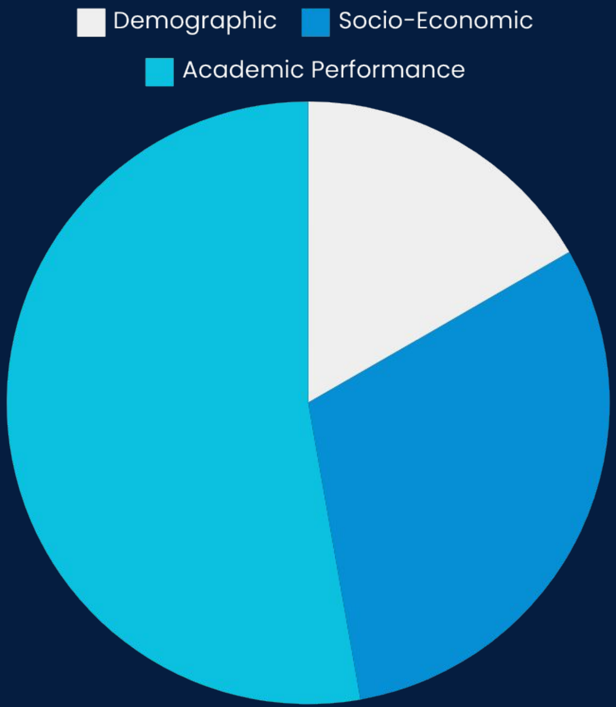

Figure 1.3 - Pie Chart categorizing variables included in the dataset

Methodology and Models

The dataset was divided into training and testing sets to accurately assess the models' performance. The training set aids the model in learning from the data, while the testing set verifies the model's precision with new, unseen data.

Logistic Regression

- Predicted outcomes based on the relationship between variables.
- Offered a well-balanced accuracy with simpler interpretation

K - Nearest Neighbor

- Predicted a student's status by looking at the status of the closest students with similar characteristics
- Optimal 'k' value: 7
- Handles complex patterns but sensitive to outliers and computationally expensive for large datasets..

XGBoost Classifier

- Improved predictions by learning from previous mistakes.
- Provided the most detailed insights into feature importance.

Model Comparisons

All models had high accuracy scores with only slight differences between each model

	Precision	Recall	Accuracy
Logistic Regression	90.88%	89.31%	90.54%
K-Nearest Neighbor	90.49%	88.57%	89.99%
XGBoost	90.05%	88.93%	89.99%

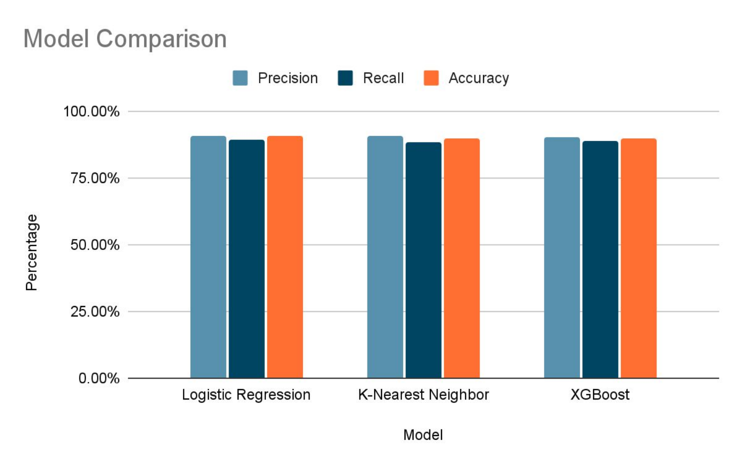

Figure 1.4 - Results testing the different models precision, accuracy, and recall ability

Logistic Regression

- **Precision:** About **90.88%** of the instances predicted as Graduates were actually Graduates.
- **Recall:** About **89.31%** of the actual Graduates were correctly predicted.
- **Accuracy:** The model correctly predicted about **90.54%** of all instances.
- **91%** accuracy on the training data and **90%** on the test data, showing it could make reliable predictions across new datasets.

K – Nearest Neighbor

- **Precision:** About **90.49%** of the instances predicted as Graduates were actually Graduates.
- **Recall:** About **88.57%** of the actual Graduates were correctly predicted.
- **Accuracy:** The model correctly predicted about **89.99%** of all instances.
- Training accuracy of **91%** and a testing accuracy of **89%**, showing good generalization.

XGBoost Classifier

- **Precision:** About **90.05%** of the instances predicted as Graduates were actually Graduates.
- **Recall:** About **88.93%** of the actual Graduates were correctly predicted.
- **Accuracy:** The model correctly predicted about **89.99%** of all instances
- Achieved the **highest accuracy** during training **(95%)** but showed signs of overfitting, as its test accuracy was **89%**

Model Recommendation

Logistic Regression outperformed the other model with the highest accuracy

- Correctly predicted Graduates (True Positives): 360
- Incorrectly predicted Dropouts as Graduates (False Positives): 72
- Incorrectly predicted Graduates as Dropouts (False Negatives): 31
- Correctly predicted Dropouts (True Negatives): 626

- The predictions mostly fall into the true positive (Graduate) and true negative (Dropout) categories, showing strong performance.
- Although there are a few false positives and false negatives, they are minimal, hinting at potential areas for enhancement.

- The AUC (Area Under Curve) is recorded at 0.89, indicating a 89% capability of the model in distinguishing between students who dropout and those who successfully graduate.

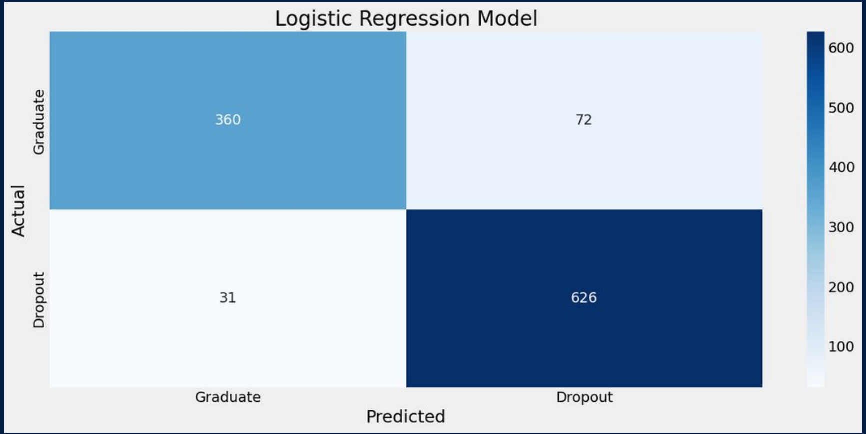

Figure 1.5 - Heatmap of the prediction accuracy of the Logistic Regression model

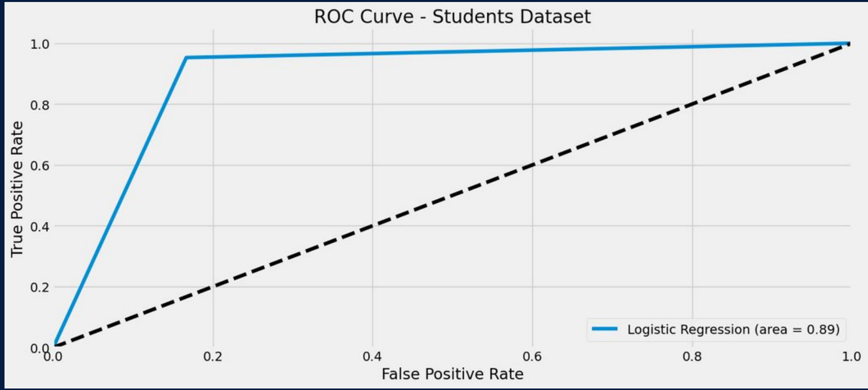

Figure 1.6 - Measurement of the Area Under Curve (AUC)

Key Findings and Recommendations

Financial and academic factors emerge as crucial predictors of student retention in feature selection analysis.

01 The number of units a student was enrolled in and approved for played a significant role.

02 The influence of the second semester on student retention surpasses that of the first semester.

03 Students' academic outcomes are significantly impacted by their financial circumstances.

Recommendations:

- Implement logistic regression in predictive monitoring tools for early intervention strategies.
- Improve predictive abilities by investigating additional data sources.
- XGBoost, with further adjustments, may prove to be a more efficient model in the future.

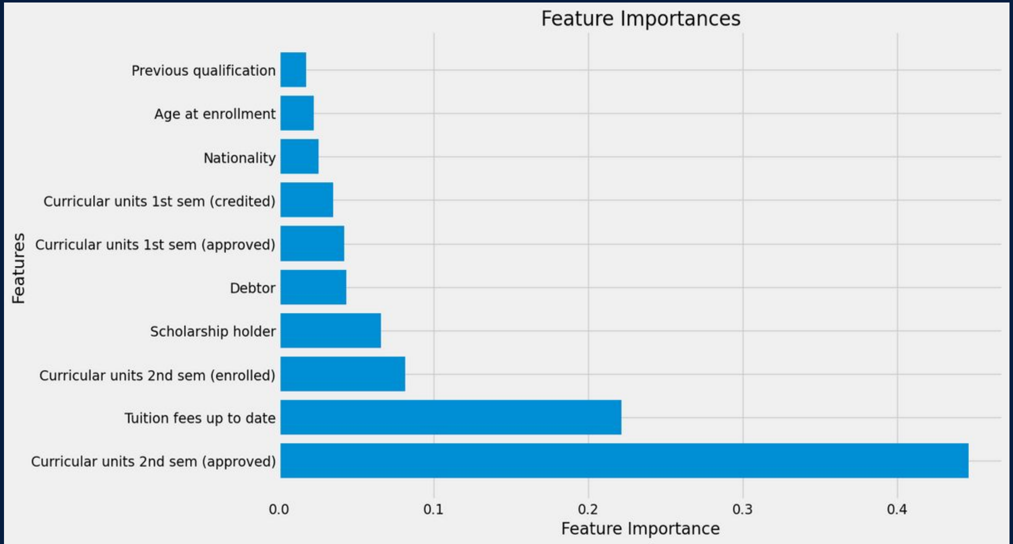

Figure 1.7 - Variables with the most influence on student retention