Predictive Modeling to Reduce Student Dropout Rates

An Analytical Approach to Enhancing Student Retention

OPORAJITA TAMANNA

August 14, 2024

Key Message

Student dropouts significantly impact educational institutions both financially and reputationally.

A predictive model has been developed to identify at-risk students early.

The model enables timely interventions, potentially reducing dropout rates.

Business Problem

Problem Statement:

- High dropout rates affect the institution's success metrics, including graduation rates, student satisfaction, and financial health.

Goal:

- To develop a predictive model that can identify students at risk of dropping out, allowing for early intervention.

Data Overview

- **Total Observation 4424, 34 variable**
- **After dropping highly correlated & irrelevant ones 4424 observations and 22 variable.**
- **Dropped target variable 'Enrolled'**

Key Features: **Academic Performance**, **Demographics and Personal Factors, Financial Situation, Course-Specific Factors**.

Curricular units 2nd sem (approved): A strong indicator of early academic success. Students struggling to pass courses early on might be at higher risk.

Previous Qualification: A student's prior academic background can influence their preparedness for higher education.

Age: Older students might face different challenges balancing academic life with other responsibilities.

Marital status: Family responsibilities could impact a student's ability to focus on studies.

Gender: There might be gender-specific factors influencing dropout rates.

Tuition fees up to date: Financial hardship is a major contributor to student dropout.

Scholarship holder: Financial support can help students stay enrolled.

Course: Certain courses might have inherently higher dropout rates due to difficulty or other factors.

Feature Engineering

Dropping Irrelevant Features: Removed several columns that were deemed irrelevant for predicting student dropout, such as 'Nationality' and 'Father's occupation',. This simplifies the model and potentially improves its performance by focusing on the most impactful features.

Encoding the Target Variable: Used **LabelEncoder** to convert the categorical 'Target' variable into numerical labels (0, 2). This is necessary for many machine learning algorithms that require numerical input.

Modeling

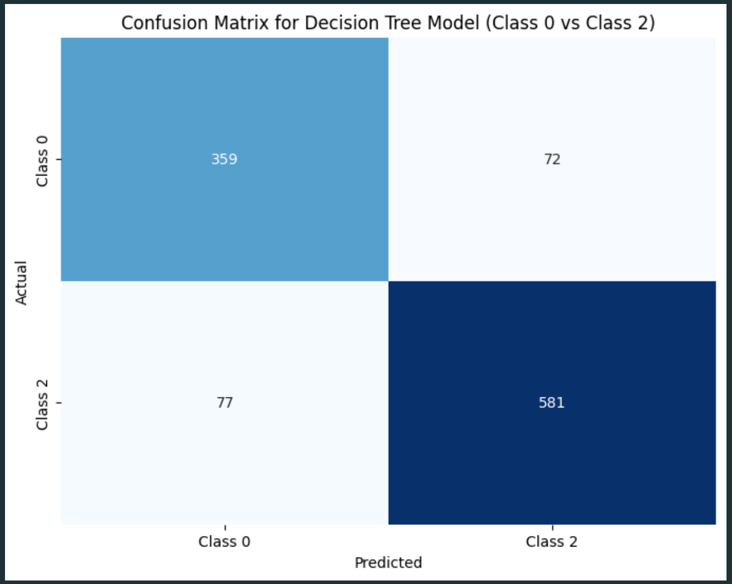
- **Techniques used** :
  - Logistic regression
  - Decision trees
  - Random forest
  - Random Forest
    with Hyperparamter Tuning

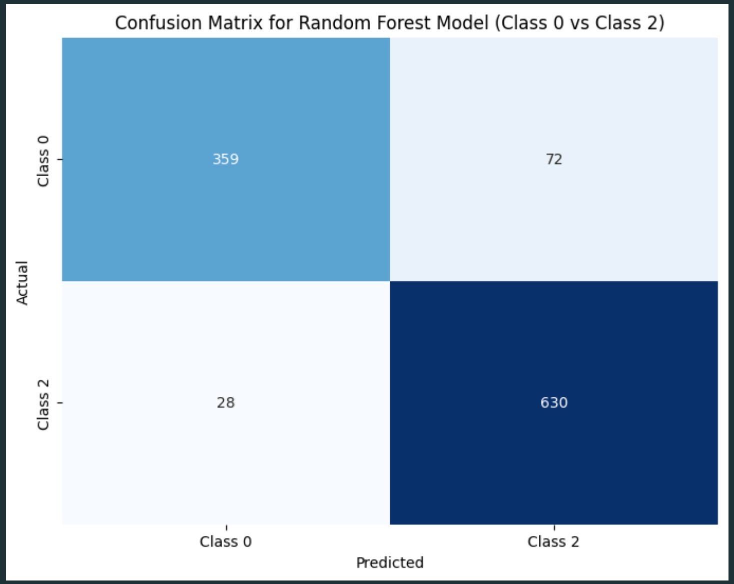
Model Performance:

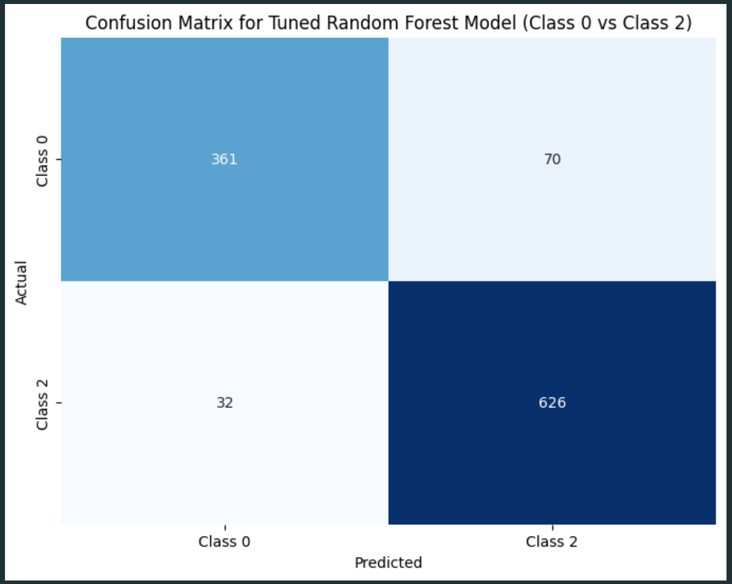
	Precision	Recall	F1-Score	Accuracy
Logistic Regression	Dropouts 89% Graduates 91%	Dropouts 86% Graduates 93%	Dropouts 87% Graduates 92%	90%
Decision Tree	Dropouts 82% Graduates 89%	Dropouts 83% Graduates 88%	Dropouts 83% Graduates 89%	86%
Random Forest	Dropouts 93% Graduates 90%	Dropouts 83% Graduates 96%	Dropouts 88% Graduates 93%	91%
Random Forest with Hyperparamter Tuning	Dropouts 92% Graduates 90%	Dropouts 84% Graduates 95%	Dropouts 88% Graduates 92%	91%

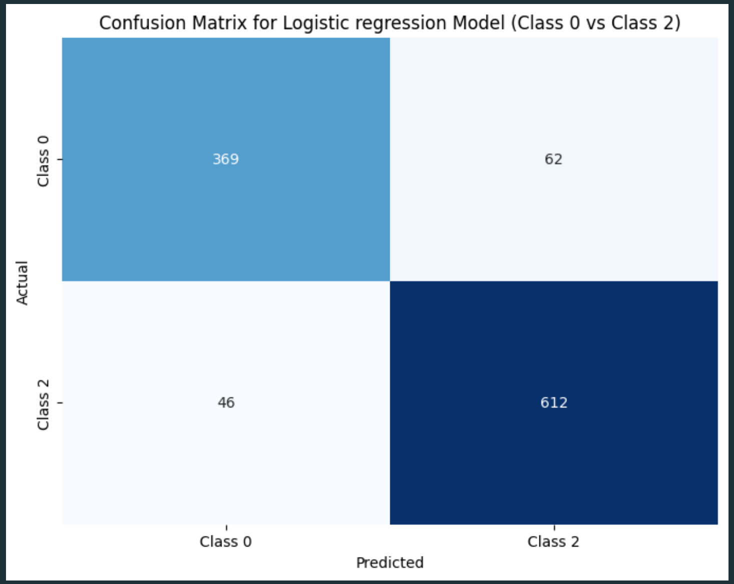
Confusion Matrix Comparison

- This confusion matrix highlights how each model performs in classifying instances into Class 0 or Class 2, with the Logistic regression, Random Forest and Tuned Random Forest models generally performing better, especially in minimizing errors for Class 2.









Best Model
Logistic Regression: For Better Readability

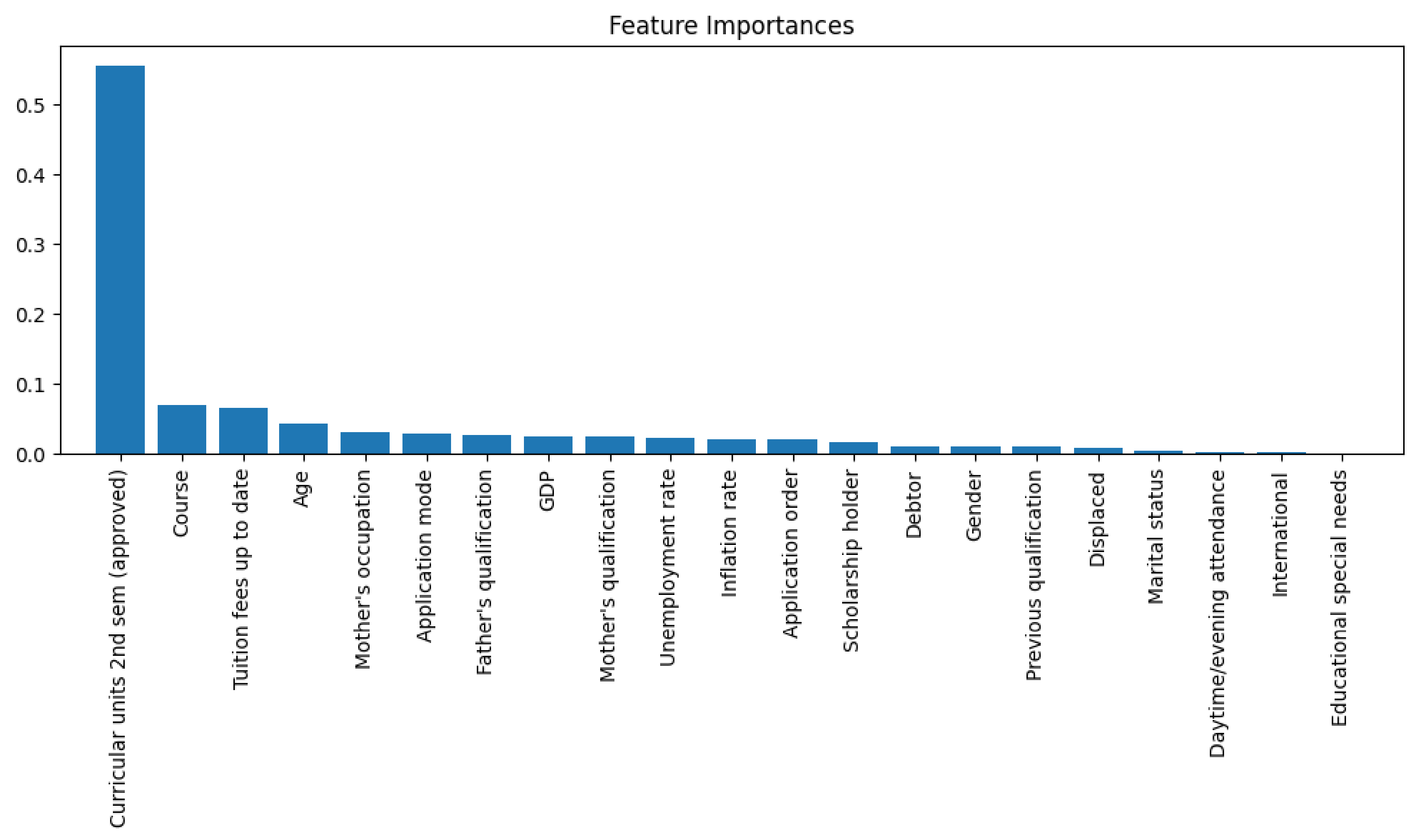
THE PRECISION FOR CLASS 0 (DROPOUTS) IS 0.89, INDICATING THAT 89% OF STUDENTS EXPECTED TO DROP OUT DID SO. THE PRECISION FOR CLASS 2 (GRADUATES) IS 0.91, INDICATING THAT 91% OF THE STUDENTS EXPECTED TO GRADUATE ACTUALLY GRADUATED.

THE RECALL FOR CLASS 0 (DROPOUTS) IS 0.86, INDICATING THAT THE MODEL ACCURATELY DETECTED 86% OF ALL REAL DROPOUT CASES. FOR CLASS 2 (GRADS), THE RECALL IS 0.93, INDICATING THAT 93% OF ALL REAL GRADUATES WERE CORRECTLY IDENTIFIED. FOR CLASS 0, THE F1-SCORE IS 0.87, SHOWING A FAIR MIX OF PRECISION AND RECALL FOR PREDICTING DROPOUTS.

FOR CLASS 2, THE F1-SCORE IS 0.92, INDICATING AN EVEN GREATER BALANCE IN PREDICTING GRADS. 431 STUDENTS DROPPED OUT (CLASS 0), WHILE 658 GRADUATED (CLASS 2).

Feature Importance

- Curricular unit's 2nd semester (approved): This attribute represents the number of courses a student successfully finished during their second semester. A substantial positive link with dropout rates would indicate that students who do not pass enough courses early on are more likely to feel disheartened and leave. This emphasizes the value of early academic help.



Actionable Insights

Recommendations:

Integrate predictive models into the institution's data systems to identify at-risk students early and accurately.

Collaborate with IT to ensure seamless integration into current systems.

Use insights from the predictive models to tailor support services such as academic advising, tutoring, and counseling.

Promote a culture of data-driven decision-making across the institution, ensuring that all staff understand the value and implications of predictive analytics.

Engage with external stakeholders, such as policymakers and educational consultants, to align predictive modeling efforts with broader educational goals.

Business Impact

Enhanced Institutional Reputation: Improved graduation rates boost the institution's rankings and reputation, making it more attractive to prospective students and faculty. Results in higher enrollment and competitive advantage in the educational market.

Increased Revenue from Higher Retention: Retaining more students results in sustained tuition revenue, reducing the need for constant recruitment efforts.
Impacting in Financial stability and improved budget forecasting.

Enhanced Student Experience: By supporting at-risk students, the institution enhances overall student satisfaction, leading to a stronger alumni network. Resulting in Long-term alumni engagement and potential future contributions.

Risk Mitigation and Funding Opportunities: Proactive management of institutional risks & improved retention rates ensure compliance with accreditation standards and enhance eligibility for grants and funding.

Conclusion

This model is not just a solution for a business problem; it's a strategic approach to nurturing the next generation of leaders and innovators.

Today's students are tomorrow's assets. Ensuring their success is crucial not just for the institutions they attend, but for the society they will go on to shape.

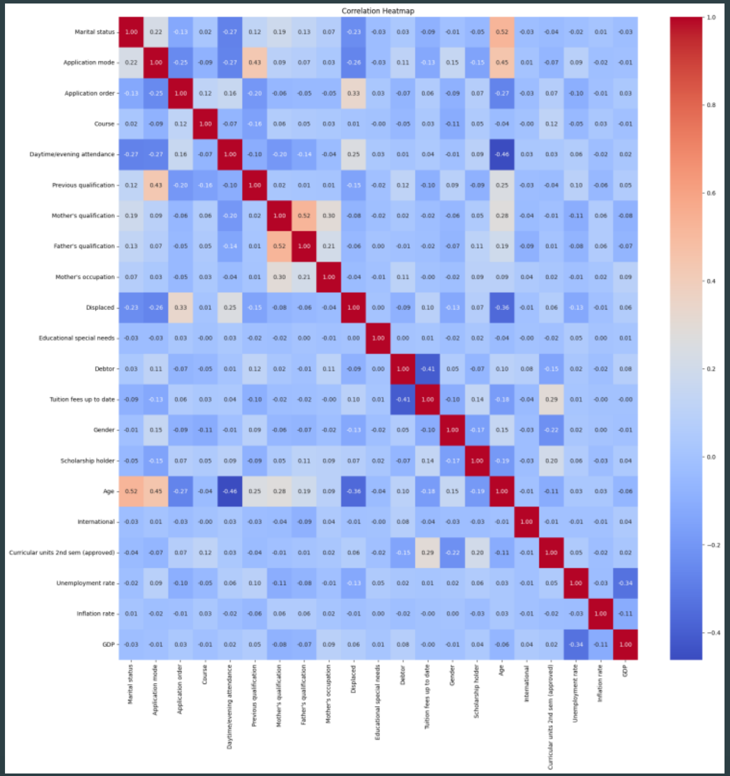
By investing in predictive modeling for student success, we're investing in a brighter, more equitable future for all.

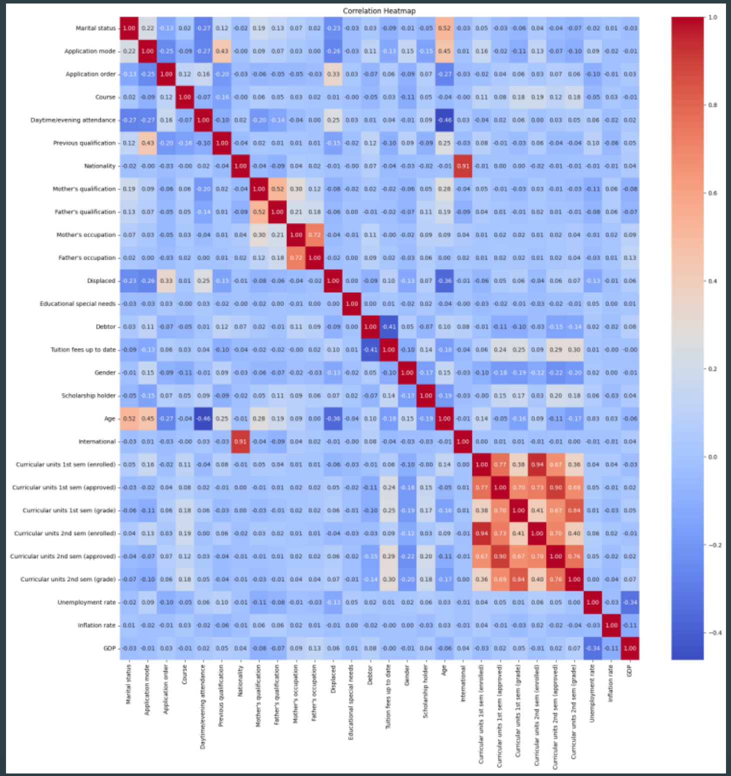
Q/A?

References

- https://www.kaggle.com/datasets/naveenkumar20bps1137/predict-students-dropout-and-academic-success/data

Appendix






```
student_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4424 entries, 0 to 4423
Data columns (total 23 columns):
 #   Column                               Non-Null Count  Dtype
---  ------                               --------------  -----
 0   Marital status                       4424 non-null   int64
 1   Application mode                     4424 non-null   int64
 2   Application order                    4424 non-null   int64
 3   Course                               4424 non-null   int64
 4   Daytime/evening attendance           4424 non-null   int64
 5   Previous qualification               4424 non-null   int64
 6   Mother's qualification               4424 non-null   int64
 7   Father's qualification               4424 non-null   int64
 8   Mother's occupation                  4424 non-null   int64
 9   Displaced                            4424 non-null   int64
 10  Educational special needs            4424 non-null   int64
 11  Debtor                               4424 non-null   int64
 12  Tuition fees up to date              4424 non-null   int64
 13  Gender                               4424 non-null   int64
 14  Scholarship holder                   4424 non-null   int64
 15  Age                                  4424 non-null   int64
 16  International                        4424 non-null   int64
 17  Curricular units 2nd sem (approved)  4424 non-null   int64
 18  Unemployment rate                    4424 non-null   float64
 19  Inflation rate                       4424 non-null   float64
 20  GDP                                  4424 non-null   float64
 21  Target                               4424 non-null   int64
 22  Target_encoded                       4424 non-null   int8
dtypes: float64(3), int64(19), int8(1)
memory usage: 764.8 KB
```