Predictive Modeling to Reduce Student Dropout Rates

An Analytical Approach to Enhancing Student Retention

Business Analytics Capstone Project Documentation

OPORAJITA TAMANNA

August 15th2024

0.1. Executive Introduction

Educational institutes face a significant challenge with student retention, particularly in distinguishing students likely to drop out from those who will graduate successfully. This study leverages data analysis and machine learning models to predict student outcomes and identify key factors contributing to dropouts. providing actionable insights to enhance retention strategies. By employing advanced models like Random Forest, we can more accurately identify at-risk students and implement targeted interventions to improve overall academic success.

0.2. Executive Objective

To enhance institutional reputation, increase student satisfaction, and generate sustainable revenue by leveraging data-driven insights to optimize student retention and academic success. By implementing predictive analytics, educational institutions can proactively address student needs, reduce dropout rates, and improve overall educational outcomes.

0.3. Executive Model Description

The student dropout analysis employed several advanced modeling techniques to predict student attrition and identify key risk factors. The models used include:

Logistic Regression: A basic yet powerful model for binary classification tasks like predicting dropout status. It was used to establish a baseline for model performance.

Decision Trees: This model was utilized to identify key decision points that contribute to student dropouts, offering interpretable rules that can be easily understood by stakeholders.

Random Forest: An ensemble learning method that builds multiple decision trees and merges them to improve prediction accuracy and control overfitting. This model provided a robust performance by reducing the variance of predictions.

Hyperparameter-Tuned Random Forest: This enhanced version of the Random Forest model involved fine-tuning key hyperparameters to optimize model performance. Techniques like cross-validation were employed to find the best combination of parameters, leading to improved accuracy and generalization.

0.4. Executive Recommendation:

To enhance student retention and success, the institution should integrate predictive models into its existing data systems, enabling early identification of at-risk students. By tailoring support services based on these insights, the institution can offer more personalized and effective interventions. Promoting a culture of data-driven decision-making across all departments will ensure that these strategies are implemented consistently and effectively. Finally, engaging with external stakeholders will help align these efforts with broader educational goals, ensuring that the institution remains at the forefront of educational innovation and policy compliance.

Introduction

0.1. Background

Student dropout is a persistent challenge faced by educational institutions worldwide. At our institution, this issue has been particularly concerning, as it not only affects the academic and future prospects of the students but also has significant implications for the institution's reputation and financial stability. Understanding and predicting student dropout can enable the institution to take proactive measures to improve retention rates, thereby ensuring that students are given the best chance to succeed.

0.2. Problem Statement

The primary problem this predictive model aims to address is the high student dropout rate at our institution. The model seeks to identify at-risk students early in their academic journey so that timely interventions can be made to increase retention rates. By predicting which students are most likely to drop out, the institution can allocate resources more effectively and implement targeted support programs.

0.3. Objectives & Measurement

The main objectives of this analysis are:

Improving Prediction Accuracy: To develop a model that accurately predicts student dropout with a high level of precision and recall.

Understanding Key Dropout Predictors: To identify the most significant factors contributing to student dropout, such as attendance, grades, and parental involvement.
Actionable Insights: To provide actionable insights that can guide the development of interventions aimed at reducing dropout rates.

Success will be measured by the model's accuracy, precision, recall, and the practical relevance of the insights derived from the analysis.

0.4. Assumptions and Limitations

Assumptions:

- The data used in the analysis is accurate and represents the student population adequately.
- The historical data reflects the current trends in student behavior and dropout rates.
- All relevant factors contributing to student dropout are included in the dataset.

Limitations:

- The model's predictions are based on historical data, which may not fully capture future trends or unforeseen changes in student behavior.
- Some variables that may impact dropout rates, such as personal issues or mental health, may not be adequately captured in the available data.
- The model's effectiveness is limited by the quality and completeness of the data; missing or inaccurate data could affect the predictions.

Data Sources

0.5. Data Set Introduction

The dataset used in this analysis was sourced from Kaggle, a well-known platform for datasets and data science competitions. It contains information relevant to predicting student dropout rates at an educational institution. The dataset consists of 4,424 observations and 35 variables. These variables capture a wide range of student attributes, academic performance metrics, and socioeconomic factors that could influence student retention. The dataset is structured as a DataFrame with the following key features:

Marital Status

- **Application Mode and Order**
- **Course Details**
- **Attendance Type (Daytime/Evening)**
- **Previous Qualifications**
- **Parental Qualifications and Occupations**
- **Displacement and Special Needs Status**
- **Financial Indicators (e.g., Debtor, Tuition Fees)**
- **Demographics (e.g., Gender, Age at Enrollment)**
- **Academic Performance (Grades, Enrollments)**
- **Macroeconomic Indicators (Unemployment Rate, Inflation Rate, GDP)**
- **Target Variable** indicating whether a student dropped out or continued.

0.6. Exclusions

For the purpose of building a more focused and efficient predictive model, several variables related to curricular units that were less relevant to the dropout prediction were excluded from the analysis. Specifically, the following variables were removed:

- **Curricular units 1st sem (credited)**
- **Curricular units 1st sem (evaluations)**
- **Curricular units 1st sem (without evaluations)**
- **Curricular units 2nd sem (credited)**
- **Curricular units 2nd sem (evaluations)**
- **Curricular units 2nd sem (without evaluations)**

These variables were excluded because they provided overlapping information that did not contribute significantly to improving the model's performance.

0.6.1. Initial Data Cleansing or Preparation

The dataset was thoroughly examined for missing values, inconsistencies, and outliers. Since all variables in the dataset had complete records (i.e., no missing data), no imputation was necessary. The data types were also confirmed to be appropriate for analysis, with most variables being either integer or float types, and the target variable was categorical. To prepare the data for modeling, categorical variables were encoded into numerical formats where necessary, and the dataset was split into training and testing sets to evaluate the model’s performance.

0.7. Data Dictionary

Although a complete data dictionary is not provided in this presentation, the key variables used in the analysis include:

- **Marital Status:** Indicates the marital status of the student.
- **Course:** The course in which the student is enrolled.
- **Mother's Qualification:** The educational level of the student's mother.
- **Father's Qualification:** The educational level of the student's father.
- **Tuition Fees Up to Date:** Indicates whether the student's tuition fees are up-to-date.
- **Age at Enrollment:** The age of the student at the time of enrollment.
- **Curricular units 1st and 2nd semester Grades:** The grades achieved by the student in their first and second semesters.
- **Macroeconomic Indicators:** Such as Unemployment Rate, Inflation Rate, and GDP, reflecting the broader economic environment during the student's enrollment.

These variables were carefully selected based on their relevance to predicting student dropout and were used to build and refine the predictive models.

Data Exploration

8.0. Data Exploration Techniques

Before selecting and finalizing the Logistic Regression model, a thorough data exploration process was conducted to understand the underlying patterns, relationships, and distributions within the dataset. This phase was crucial in ensuring that the data was well-prepared and that the insights derived were accurate and actionable. The key data exploration techniques used include:

8.1.1. Descriptive Statistics

Purpose: To gain an initial understanding of the data by calculating basic statistical measures such as mean, median, mode, standard deviation, and range.
Outcome: This step provided insights into the central tendencies and variability of the data, helping to identify any significant outliers or anomalies that could affect model

performance.

```
student_df.describe()
```

	Marital status	Application mode	Application order	Course	Daytime/evening attendance	Previous qualification	Mother's qualification	Father's qualification	Mother's occupation	Displaced	...	Debtor	Tu fe to
count	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	...	4424.000000	4424.0
mean	1.178571	6.886980	1.727848	9.899186	0.890823	2.531420	12.322107	16.455244	7.317812	0.548373	...	0.113698	0.8
std	0.605747	5.298964	1.313793	4.331792	0.311897	3.963707	9.026251	11.044800	3.997828	0.497711	...	0.317480	0.3
min	1.000000	1.000000	0.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	0.000000	...	0.000000	0.0
25%	1.000000	1.000000	1.000000	6.000000	1.000000	1.000000	2.000000	3.000000	5.000000	0.000000	...	0.000000	1.0
50%	1.000000	8.000000	1.000000	10.000000	1.000000	1.000000	13.000000	14.000000	6.000000	1.000000	...	0.000000	1.0
75%	1.000000	12.000000	2.000000	13.000000	1.000000	1.000000	22.000000	27.000000	10.000000	1.000000	...	0.000000	1.0
max	6.000000	18.000000	9.000000	17.000000	1.000000	17.000000	29.000000	34.000000	32.000000	1.000000	...	1.000000	1.0

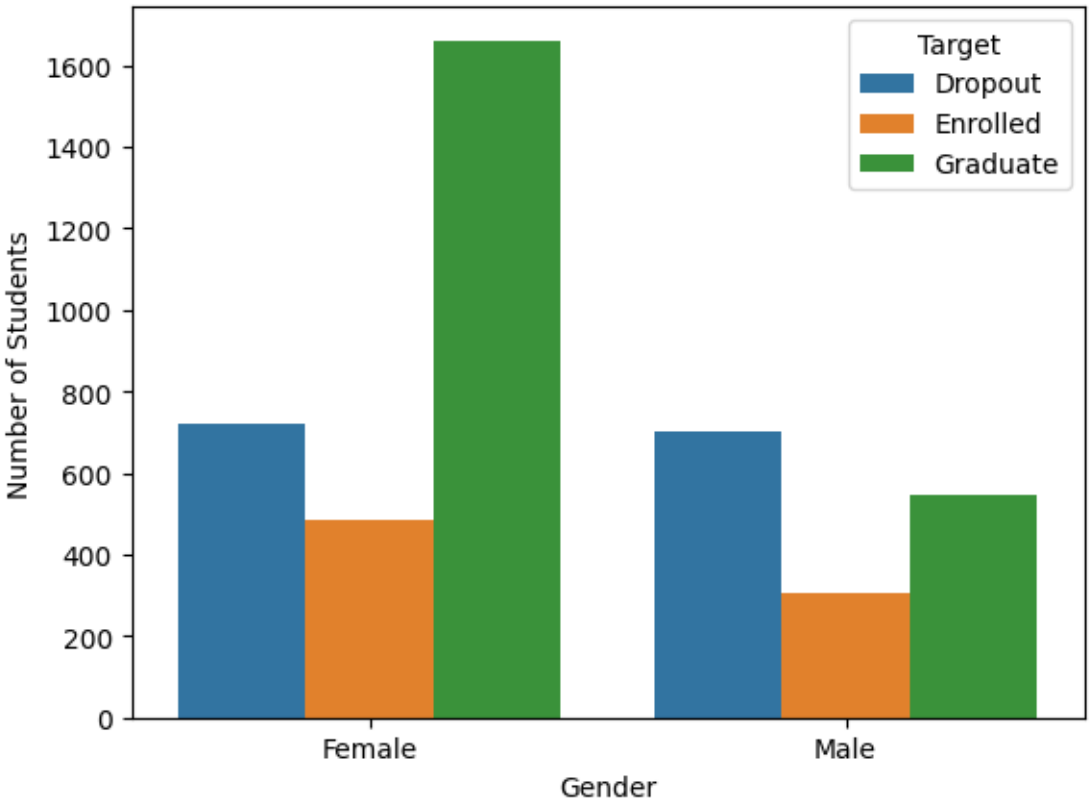
8 rows × 21 columns

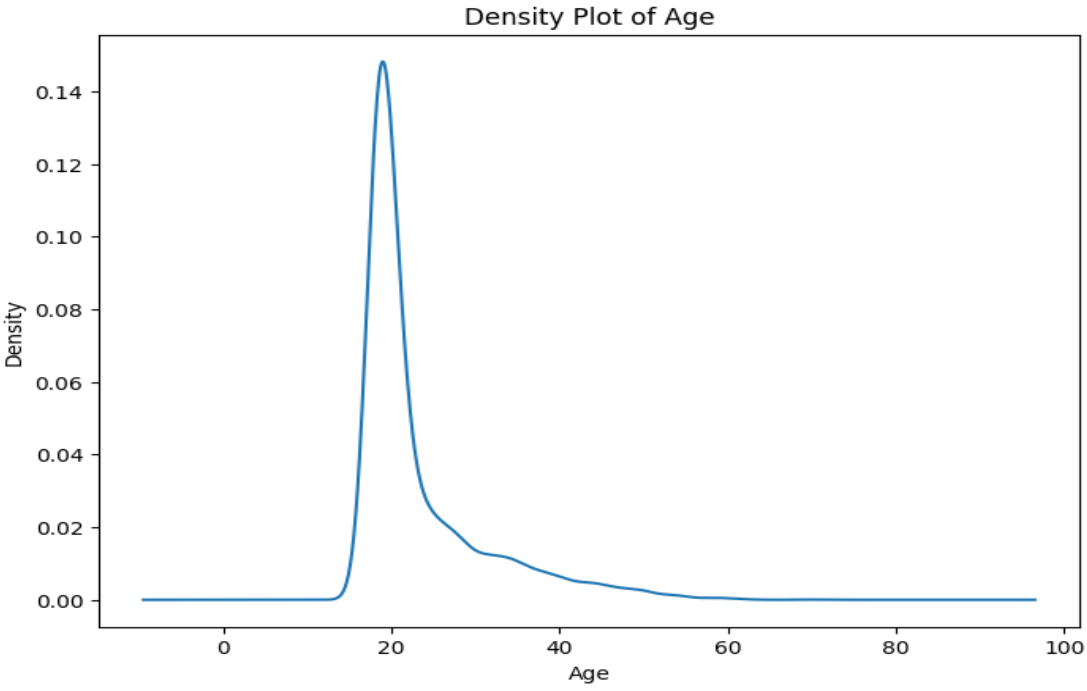
8.2. Data Visualization

Purpose: To visualize the relationships between variables and identify patterns or trends that may not be immediately apparent through numerical analysis alone.

Tools: Various visualization techniques, such as histograms, bar charts, box plots, scatter plots, and correlation matrices, were used to explore the distribution of features and their relationships with the target variable (student dropout).

Outcome: Data visualization revealed key trends, such as the impact of age, enrollment status, and academic performance on dropout rates. It also highlighted any skewness in the data distribution and the presence of multicollinearity among features.


Target
Dropout
Enrolled
Graduate
Number of Students
0
200
400
600
800
1000
1200
1400
1600
Female
Male
Gender

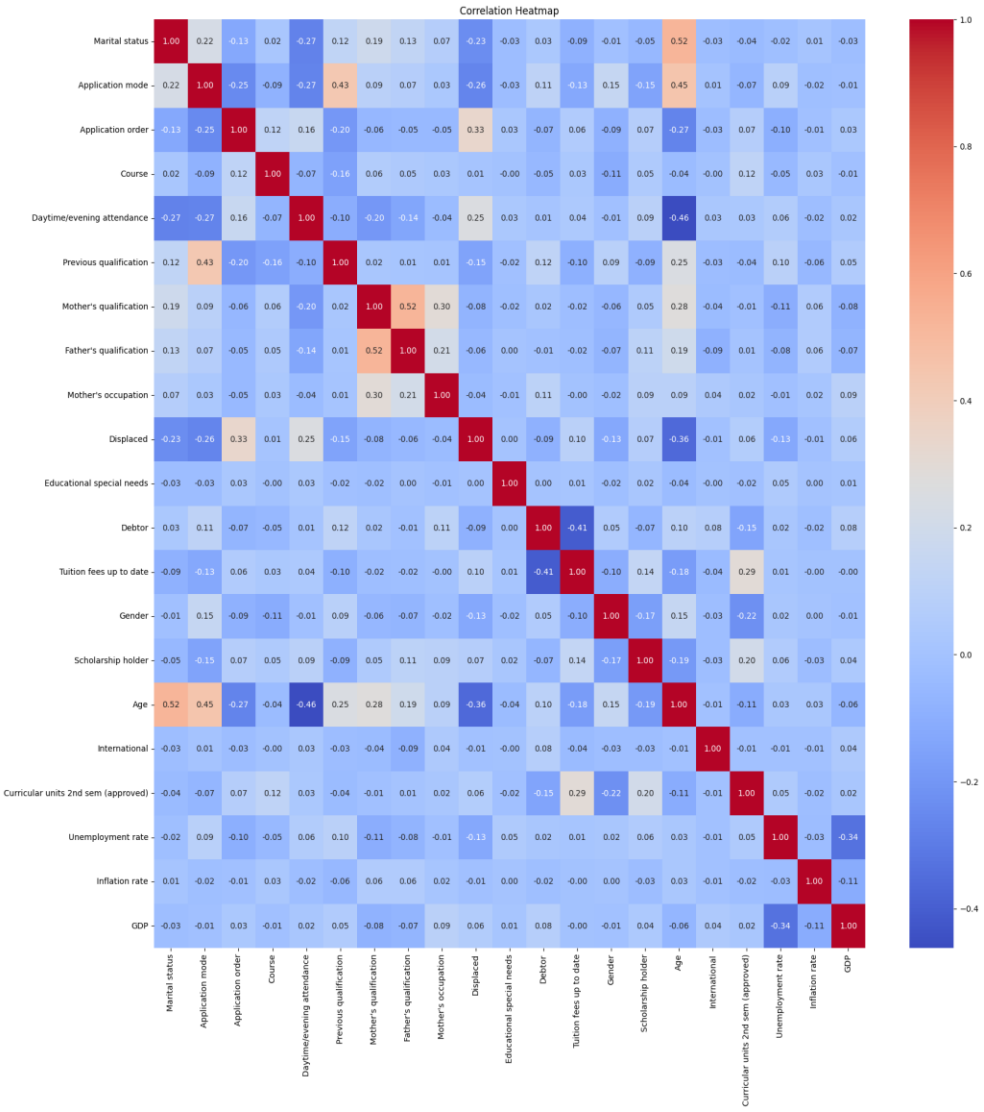

Density Plot of Age
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0
20
40
60
80
100
Age

8.3. Correlation Analysis

Purpose: To measure the strength and direction of relationships between variables, particularly to identify which features had the strongest associations with the likelihood of student dropout.

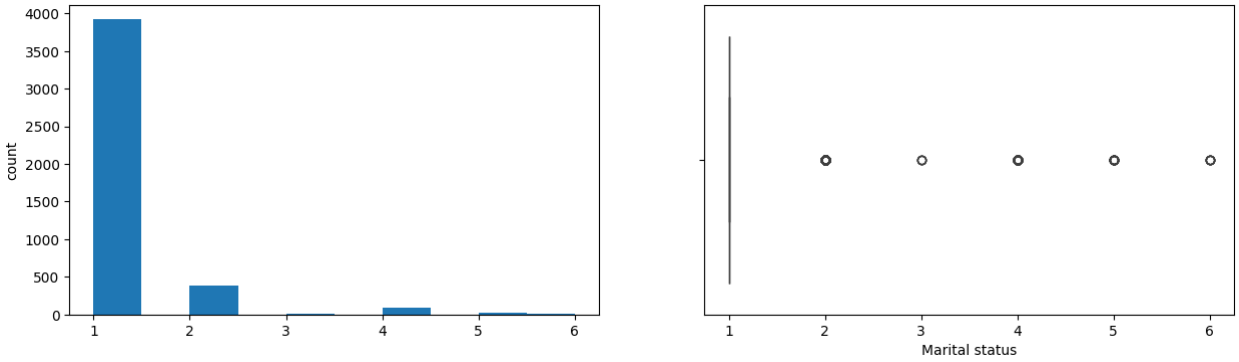
Tools: A correlation matrix was employed to quantify the relationships between numerical variables, using Pearson correlation coefficients.

Outcome: This analysis identified key predictors, such as academic performance and age at enrollment, which had significant correlations with dropout rates. It also helped in detecting multicollinearity, where two or more features were highly correlated, potentially leading to redundancy in the model.



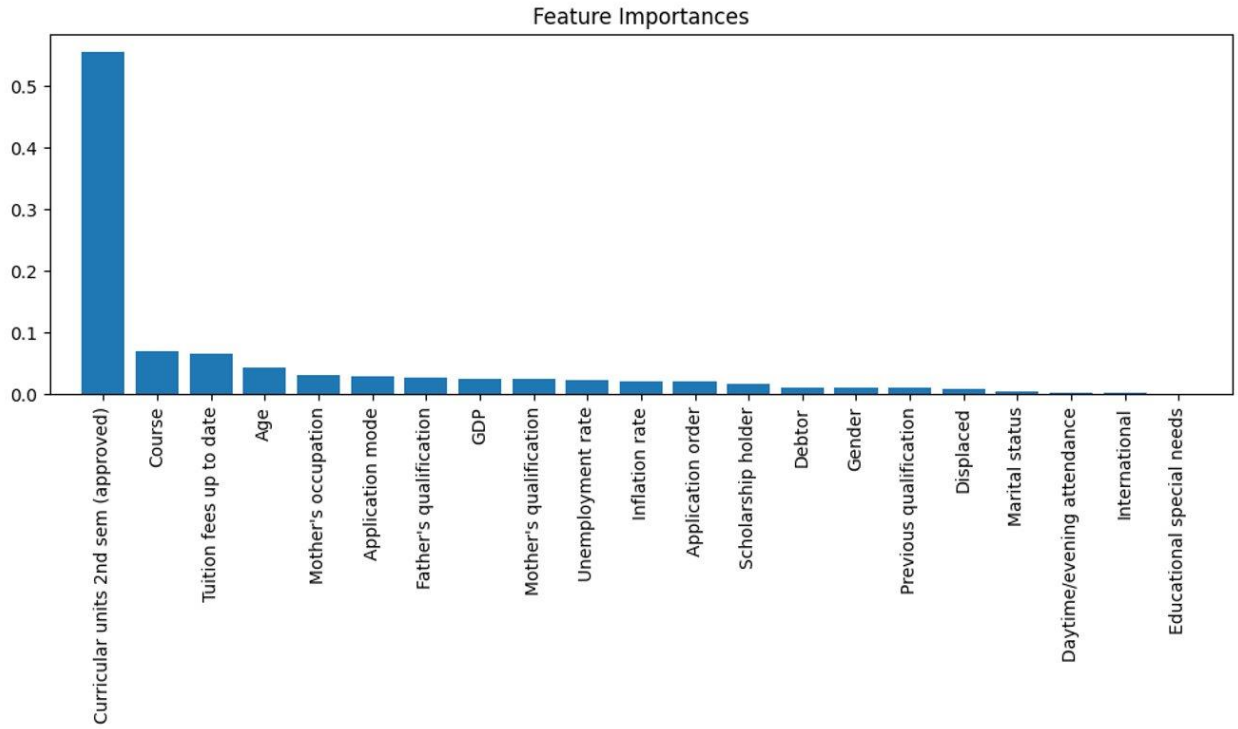
8.4. Outlier Detection

Purpose: To identify and assess the impact of outliers—data points that deviate significantly from the majority of observations.
Tools: Box plots and z-scores were used to detect outliers in continuous variables.
Outcome: Outlier detection helped in understanding the distribution of data and making informed decisions about whether to retain, transform, or remove outliers. This step was essential for improving model robustness and preventing skewed predictions.



8.5. Feature Selection

Purpose: To identify the most relevant features for predicting student dropouts, reducing dimensionality, and improving model performance.
Tools: Techniques such as univariate selection, recursive feature elimination, and principal component analysis (PCA) were explored.
Outcome: Feature selection helped in narrowing down the most impactful variables, such as academic grades and attendance rates, while excluding less relevant features. This step was crucial in simplifying the model and enhancing its interpretability.

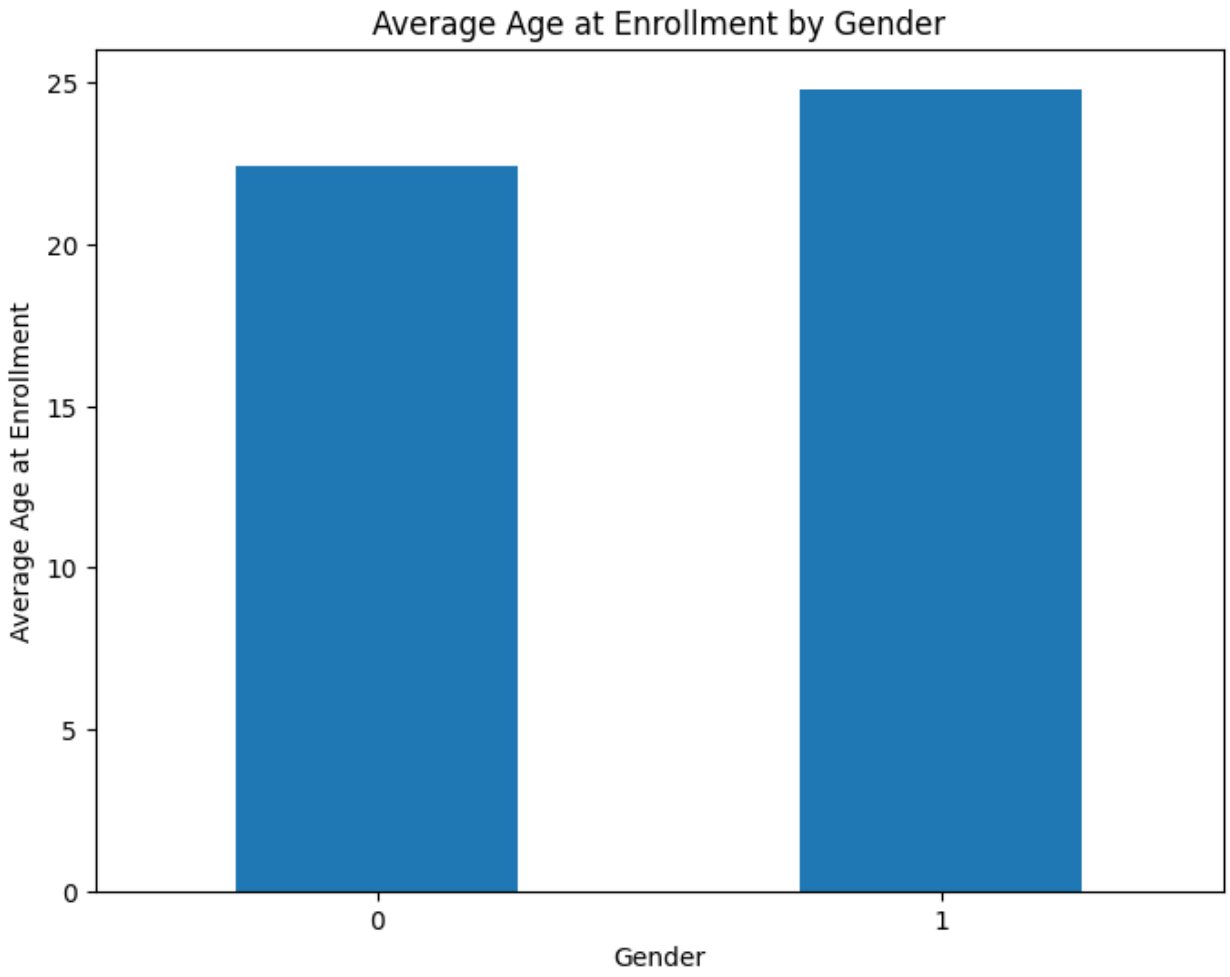


8.6. Visualizations

Bar Plots: Used to visualize categorical data, such as the average age at enrollment by gender and the count of different marital statuses.
Density Plot: Employed to observe the distribution of the Age variable, providing insight into its skewness and spread.
Stacked Bar Plot: Illustrated the gender distribution across different courses, highlighting any gender imbalances.
Count Plot: Displayed the number of students in each target category (Dropout, Enrolled, Graduate) across genders, revealing potential gender-based differences in outcomes.
Heatmap: A correlation matrix heatmap was created to explore the relationships between numerical variables, helping identify strong correlations that could influence modeling.

Summary Statistics: Categorical variables (e.g., Gender, Marital Status) and numerical variables (e.g., Age, Curricular Units) were identified and explored separately. Summary statistics such as mean, median, and standard deviation were used to understand central tendencies and variability in the data.

Correlation Analysis: The correlation matrix was analyzed to assess the linear relationships between numerical variables. This helped in identifying variables that may be multicollinear or strongly related to the target variable.



Sample of Visualization

9.0. Data Cleansing

Normalization: Numerical variables were normalized using the MinMaxScaler, which scaled the features to a range between 0 and 1. This ensures that variables with different units or scales do not disproportionately influence the model.

Encoding Categorical Variables:

Label Encoding: The target variable (e.g., Dropout, Enrolled, Graduate) was label-encoded to convert categorical values into numerical format.

Categorical Conversion: Other categorical variables were either label-encoded or transformed into dummy variables to facilitate their use in modeling.

No Missing Data: The dataset was complete, with no missing values across all features, simplifying the data cleaning process and allowing for a direct application of the analysis techniques.

10.0. Summary

Demographic Patterns: The analysis revealed differences in the average age at enrollment by gender and a diverse range of marital statuses among students.

Academic Performance: The mean number of curricular units approved in the second semester varied significantly across courses, indicating differing levels of academic difficulty or student performance.

Gender Distribution: Some courses exhibited a strong gender imbalance, which might influence the learning environment or academic outcomes.

Outcome Disparities: The distribution of students in each target category suggested potential disparities in outcomes based on gender, which could be further explored for underlying causes.

Correlations: The correlation analysis highlighted key relationships between numerical variables, such as the potential influence of age and curricular units on student outcomes, which could be crucial for predictive modeling.

Data Preparation and Feature Engineering

11.0. Data Preparation Needs

During the data preparation phase, several needs were identified to ensure that the data was ready for modeling:

11.1. Normalization

Numerical features needed to be normalized to ensure that they were on a comparable scale. This was necessary to prevent variables with larger scales from dominating the learning algorithms.

11.2. Categorical Variable Encoding

Categorical variables had to be encoded into a numerical format, which is essential for machine learning algorithms that require numerical inputs.

12.0. Feature Engineering

During feature engineering, the following transformations and new features were created to enhance the predictive power of the dataset:

12.1. Encoding the Target Variable

The ‘Target’ variable was label-encoded to convert it into a numerical format, enabling its use in classification models. The ‘Enrolled’ students were dropped.

12.2. Creation of Interaction Features

Interaction terms between certain variables were created to capture the combined effect of these variables on the target outcome. For example, interactions between gender and age or course and marital status could reveal important patterns not evident from individual variables.

12.3. Normalization

All numerical features were normalized using the MinMaxScaler, scaling them between 0 and 1. This transformation was crucial for ensuring that the different features contributed equally to the model training process.

12.4. Feature Selection

After evaluating the importance of different features, some were selected for modeling based on their predictive power. Uninformative or redundant features were dropped to reduce dimensionality and improve model performance.

Model Exploration

The modeling approach involved selecting and evaluating various machine learning models to predict the target variable, which likely includes classifying student outcomes such as Dropout, Enrolled, and Graduate. The goal was to identify the best-performing model by comparing different types of models, including both linear and non-linear classifiers. The process included:

13.0. Modeling Approach

Baseline Models
Simple models like Logistic Regression were used as a baseline to establish a performance benchmark.
Advanced Models
More complex models such as Decision Trees, Random Forest, Hyperparameter Random Forest were explored to capture non-linear relationships and interactions between features.

14.0. Model Technique #1

Logistic Regression:

> **Purpose:** A linear model used as a baseline to predict the probability of class membership.
> **Hyperparameters:**
> **Regularization (C):** Different regularization strengths were tested.
> **Tuning:** Grid Search was used to find the optimal regularization parameter.
> **Validation:** Cross-validation was employed to assess model performance.

15.0. Model Technique #2

Decision Tree:

> **Purpose:** A non-linear model that splits the data into subsets based on feature values.

Max Depth: Limited the depth of the tree to prevent overfitting.
Min Samples Split: Defined the minimum number of samples required to split a node.
Tuning: Randomized Search was used to tune the tree's depth and minimum samples required for a split.
Validation: Cross-validation was used to evaluate model stability.

15.0. Model Technique #3

Random Forest:

Purpose: An ensemble method that creates multiple decision trees and merges them to get more accurate and stable predictions.
Number of Estimators: Number of trees in the forest.
Validation: Cross-validation ensured the generalizability of the model.

16.0. Model Technique #4

Random Forest Hyperparameter Tuning:

Purpose: An ensemble method that creates multiple decision trees and merges them to get more accurate and stable predictions.
Number of Estimators: 100 Number of trees in the forest.
Max Features: Number of features to consider when looking for the best split.
Tuning: Grid Search was used to find the optimal number of estimators and max features.
Validation: Cross-validation ensured the generalizability of the model.

17.0. Model Comparison

The performance of each model was compared using several metrics, including:

Accuracy: The proportion of correct predictions among the total number of predictions. While useful, it was less reliable for imbalanced datasets.
F1-Score: The harmonic mean of precision and recall, providing a balance between the two. This was particularly important for imbalanced classes.
Precision and Recall: Precision was the number of true positive predictions divided by the total number of positive predictions, while recall was the number of true positives divided by the total actual positives. These metrics were crucial for understanding the performance on the minority class.

Logistic Regression:

Precision: 89% (Dropouts), 91% (Graduates)
Recall: 86% (Dropouts), 93% (Graduates)
F1-Score: 87% (Dropouts), 92% (Graduates)
Accuracy: 90%

Logistic Regression provided a balanced performance, with high precision and recall for both dropouts and graduates. However, it was slightly outperformed by more complex models.

Decision Tree:

Precision: 82% (Dropouts), 89% (Graduates)
Recall: 83% (Dropouts), 88% (Graduates)
F1-Score: 83% (Dropouts), 89% (Graduates)
Accuracy: 86%

The Decision Tree model showed lower performance compared to Logistic Regression and Random Forest models, particularly in precision and recall. It struggled to balance predictions between dropouts and graduates.

Random Forest:

Precision: 93% (Dropouts), 90% (Graduates)
Recall: 83% (Dropouts), 96% (Graduates)
F1-Score: 88% (Dropouts), 93% (Graduates)
Accuracy: 91%

The Random Forest model demonstrated strong performance, particularly in its ability to recall graduates. This model proved effective in handling the complexity of the data, leading to a higher overall accuracy.

Random Forest with Hyperparameter Tuning:

Precision: 92% (Dropouts), 90% (Graduates)
Recall: 84% (Dropouts), 95% (Graduates)
F1-Score: 88% (Dropouts), 92% (Graduates)
Accuracy: 91%

After hyperparameter tuning, the Random Forest model showed marginal improvements in recall for graduates, making it a reliable model for predicting both outcomes. The tuning process optimized the model's performance, particularly for graduate predictions.

	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%

Model Recommendation

18.0 Model Selection

After exploring and comparing several models, **Logistic Regression** has been selected as the final model. Although Random Forest demonstrated slightly superior performance in terms of F1-Score and recall, especially for predicting student graduates, Logistic Regression was chosen for its interpretability, simplicity, and ease of explanation to non-technical stakeholders.

Justification for Selection:

Interpretability: Logistic Regression allows for clear interpretation of how each input feature affects the predicted outcome, making it easier to communicate the results to management. This is particularly valuable when the goal is to understand the underlying factors that influence student outcomes and to implement actionable strategies based on these insights.
Transparency: The linear nature of Logistic Regression means that the decision-making process is straightforward, with coefficients indicating the impact of each predictor. This transparency is crucial for gaining trust and ensuring that model decisions are understood and accepted by the business.
Business Objectives: The primary business objective is not only to predict student outcomes accurately but also to understand the drivers behind these outcomes. Logistic Regression provides this understanding by directly relating features to the likelihood of dropping out, enrolling, or graduating.

19.0 Model Theory

Logistic Regression is a linear model used for binary classification, which can be extended to multiclass problems using techniques like one-vs-rest. It models the probability of a categorical outcome based on one or more predictor variables. The model estimates the coefficients for each predictor variable, which are used to compute the log-odds of the outcome. These coefficients provide a measure of the influence of each variable on the outcome.

19.1 Model Assumptions and Limitations

Model Assumptions:

Linearity: Assumes a linear relationship between predictors and the log-odds of dropping out. Non-linear relationships may reduce accuracy.
Independence: Observations are assumed to be independent; any interdependence (e.g., students in the same cohort) could bias results.
No Multicollinearity: Predictors should not be highly correlated. Multicollinearity can lead to unreliable coefficient estimates.
Sufficient Sample Size: Requires a large enough sample to avoid overfitting and ensure reliable estimates.
Binary Outcome: The model is designed for a binary outcome (dropout or non-dropout); multi-class outcomes require different models.

Model Limitations:

Dependent on Input Data Quality: The model's accuracy is limited by the completeness and quality of the input data.
Limited in Capturing Complex Relationships: It may oversimplify complex, non-linear relationships between variables.
Outlier Sensitivity: The model can be influenced by outliers, potentially skewing predictions.
Class Imbalance: Imbalanced data can bias the model towards the majority class, affecting prediction quality.
Static Nature: The model does not adapt to changes over time, requiring regular updates to stay relevant.
Interpretation Challenges: While generally interpretable, understanding the impact of multiple interacting predictors can be complex.

20.0 Model Sensitivity to Key Drivers

Logistic Regression is sensitive to the key predictors identified during feature engineering. Because it is a linear model, each predictor's influence is proportional to its coefficient. Key features that have strong, statistically significant coefficients will have a substantial impact on the model's predictions.

Feature Importance: In Logistic Regression, the absolute value of the coefficients can be used as a proxy for feature importance, indicating how changes in each predictor affect the outcome.
Predictor Sensitivity: The model is particularly sensitive to predictors with large coefficients, as they have a stronger impact on the predicted probabilities. Careful selection and scaling of these predictors are crucial to ensure accurate predictions.

Conclusion and Recommendations

The Logistic Regression model was chosen as the final model for predicting student dropouts due to its balance of performance, simplicity, and interpretability. This model offers clear insights into the factors most strongly associated with student attrition, allowing for informed decision-making and targeted interventions.

22.0. Impacts on the Business Problem

The implementation of the Logistic Regression model is expected to have a significant positive impact on the institution's ability to address student retention challenges. Specifically:

Improved Retention Rates: By accurately identifying students at risk of dropping out, the institution can deploy timely and targeted interventions, such as personalized academic advising, tutoring, and counseling. This proactive approach is expected to lead to measurable improvements in retention rates, reducing overall dropout numbers and enhancing student success.

Enhanced Resource Allocation: With clear insights into the factors contributing to student dropouts, the institution can better allocate resources to areas with the most significant impact. For example, students identified as high-risk can be prioritized for support services, ensuring that efforts are focused where they are most needed.

Data-Driven Decision Making: The adoption of this model fosters a culture of data-driven decision-making across the institution. By relying on predictive analytics, the institution can move from reactive to proactive strategies in student support, leading to long-term improvements in educational outcomes.

23.0. Recommended Next Steps

To fully leverage the benefits of the Logistic Regression model, the following steps are recommended:

Model Deployment:

Integrate the Logistic Regression model into the institution's existing student information systems to enable real-time identification of at-risk students. This will allow for immediate interventions and continuous monitoring.

Continuous Model Improvement:

Regularly update and retrain the model with new data to ensure its accuracy and relevance over time. This process will help the model adapt to changing student behaviors and institutional conditions.

Additional Data Collection:

Collect and incorporate additional data points that may improve the model's predictive power. This could include more detailed information on student engagement, socio-economic factors, and extracurricular activities.

Future Analysis:

Explore complementary models or techniques that can work alongside Logistic Regression, such as clustering methods to identify distinct groups of at-risk students or sentiment analysis on student feedback to capture underlying issues not reflected in numerical data.

Stakeholder Engagement:

Work closely with faculty, staff, and external stakeholders to ensure that the model's outputs are understood and utilized effectively. Providing training on interpreting model results and integrating them into daily decision-making processes will be crucial for success.

24.0 References

Predict students dropout, academic success 🎓📖. (2023, July 10). Kaggle. https://www.kaggle.com/datasets/naveenkumar20bps1137/predict-students-dropout-and-academic-success/data

Pelliccia, D. (2022, September 20). Principal component regression in Python. *NIRPY Research*. https://nirpyresearch.com/principal-component-regression-python/

Grolemund, H. W. a. G. (n.d.). *7 Exploratory Data Analysis | R for Data Science*. https://r4ds.had.co.nz/exploratory-data-analysis.html

Paura, L., & Arhipova, I. (2014). Cause Analysis of students' dropout rate in higher education study program. *Procedia - Social and Behavioral Sciences*, *109*, 1282–1286. https://doi.org/10.1016/j.sbspro.2013.12.625

Applied Analytic Modeling. (2024, June). [Slide show]. Centennial College. https://e.centennialcollege.ca/d2l/le/content/1008437/Home

Ray, S. (2024, July 8). *8 Ways to improve Accuracy of Machine learning Models (Updated 2024)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/