Predicting Student Dropout Rates

ISDS 574 Final Project

Dave Widjaja, Joseph Mcclain, JP Millan, Keerthanaa Ellur,
Khue Nguyen, Temi Oyefeso, Vince Luong

Agenda

01

Introduction

- Background
- Problem
- Data Description

02

Data Preprocessing

- Data Cleaning
- Data Partitioning

03

Modelling & Analysis

- Models
- Analysis
- Performance comparisons

04

Conclusions

- Takeaways
- Citations

Introduction

Background

Goal

Reduce the student dropout rate at Belvedere University

Purpose

Identify the major factors causing high student dropout rates

Expectations

Create data-driven policies to increase the number of graduates

Top Causes of Dropout

1. Financial Challenges
2. Mental Health Issues
3. Lack of Academic Support
4. Disconnection from Campus Life
5. Competing Responsibilities

*According to the World Economic Forum

Data Description

Dataset Origin:

- UCI Machine Learning Repository
- https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

Scope of Data:

- Focuses on student demographics, academic performance, and socio-economic context.
- Includes factors such as grades, parental education, and national economic indicators.

Target Variable:

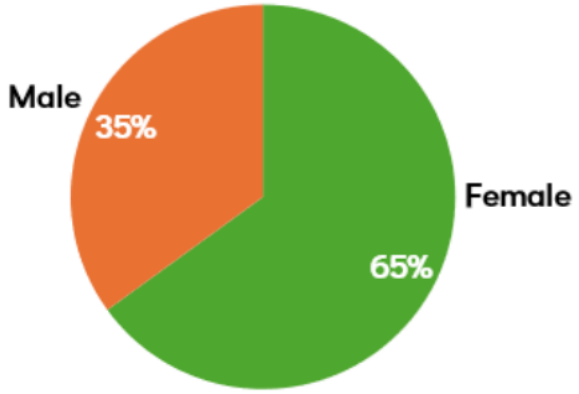
- Predicts outcomes: Dropout, Graduate, or Enrolled.

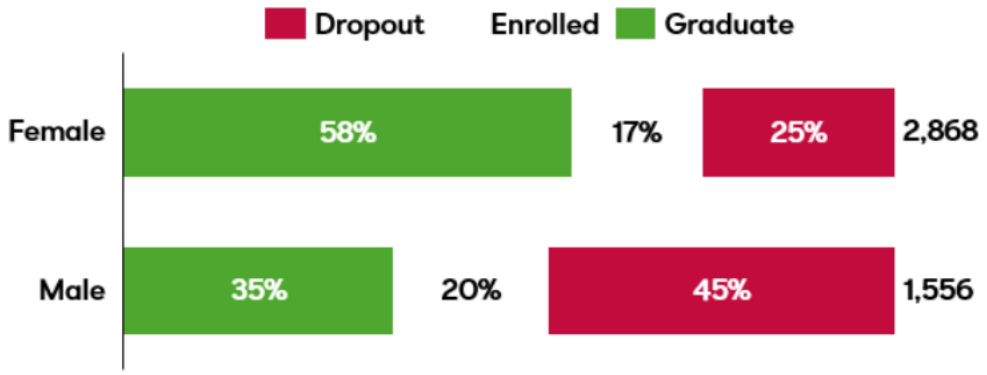
Sample Features:

- Categorical: Marital Status, Course, Nationality, Educational Needs.
- Continuous: Admission Grade, Previous Qualification Grade, GDP, Unemployment Rate.

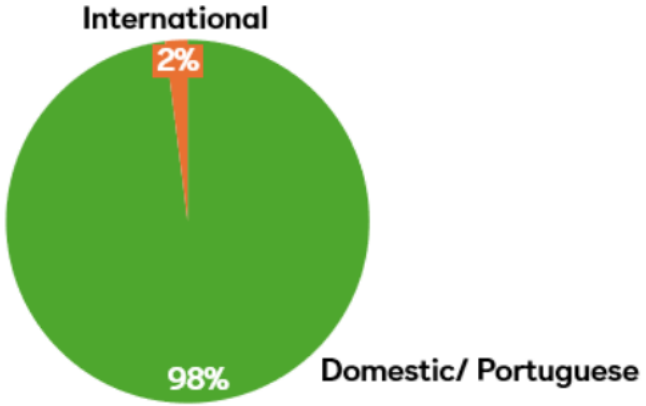
Descriptive Statistics

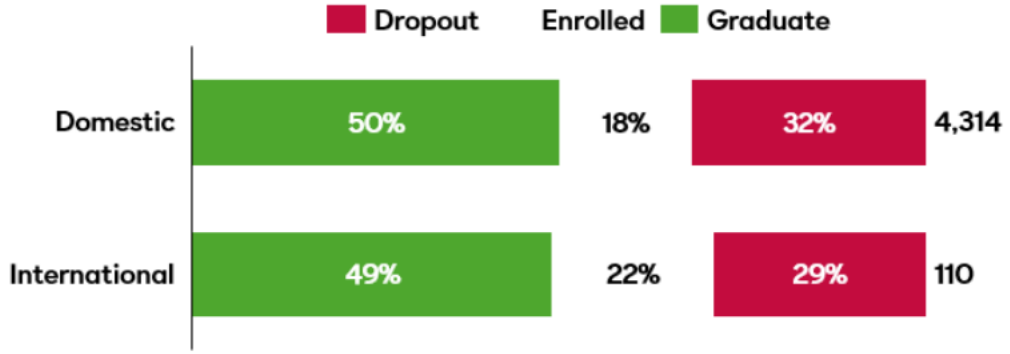
Gender



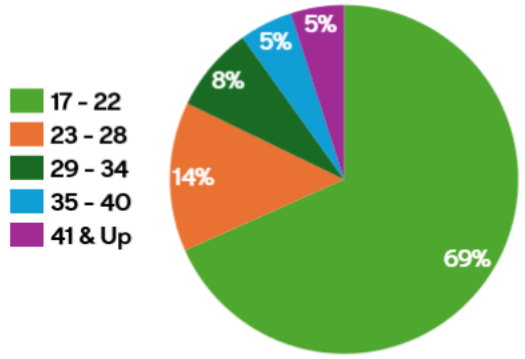


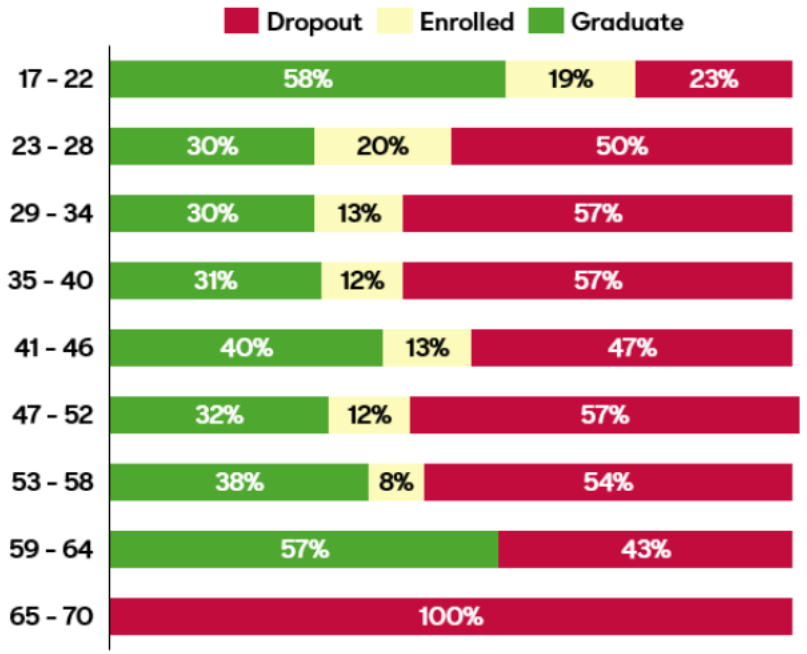
Nationality



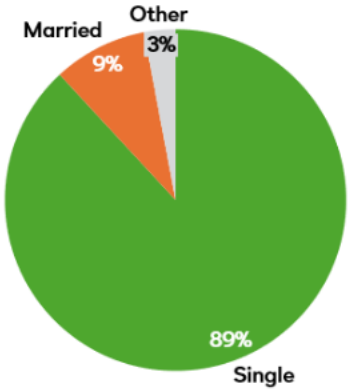


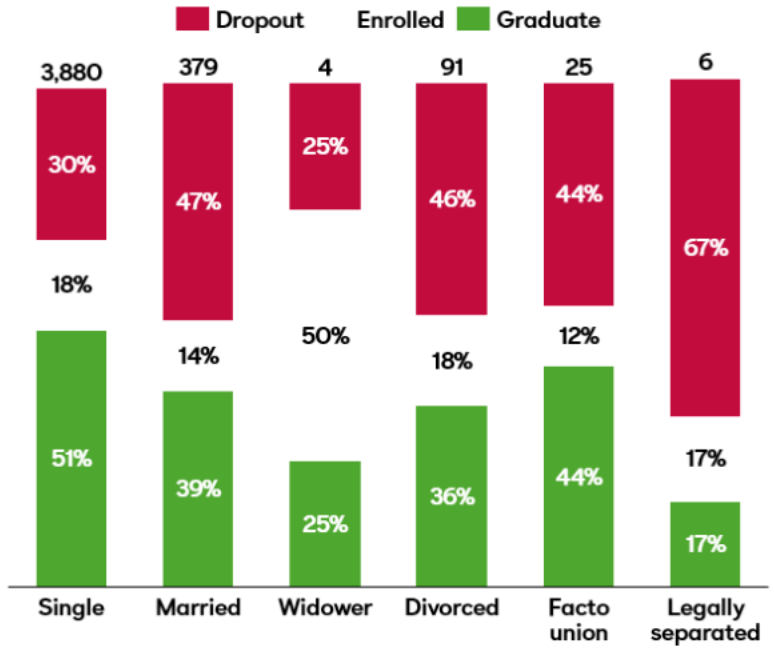
Age





Marital Status





Data Preprocessing

Data Exploration and Cleaning

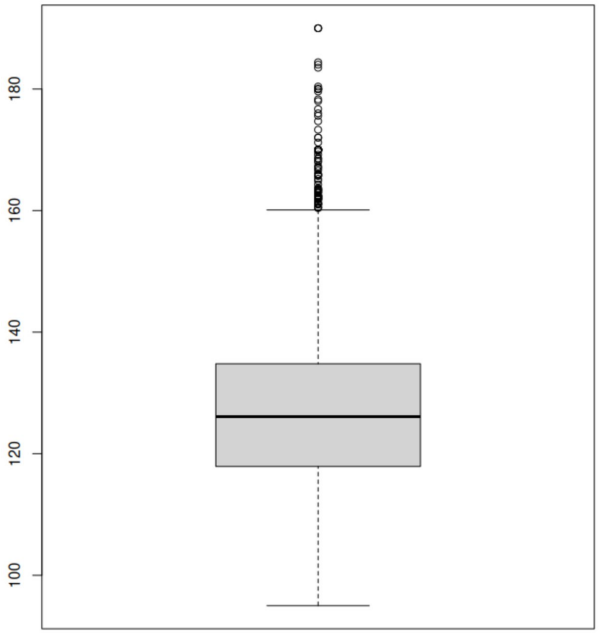
Dataset Summary:

- Rows: 4424, Columns: 37
- No missing values.

Steps Taken:

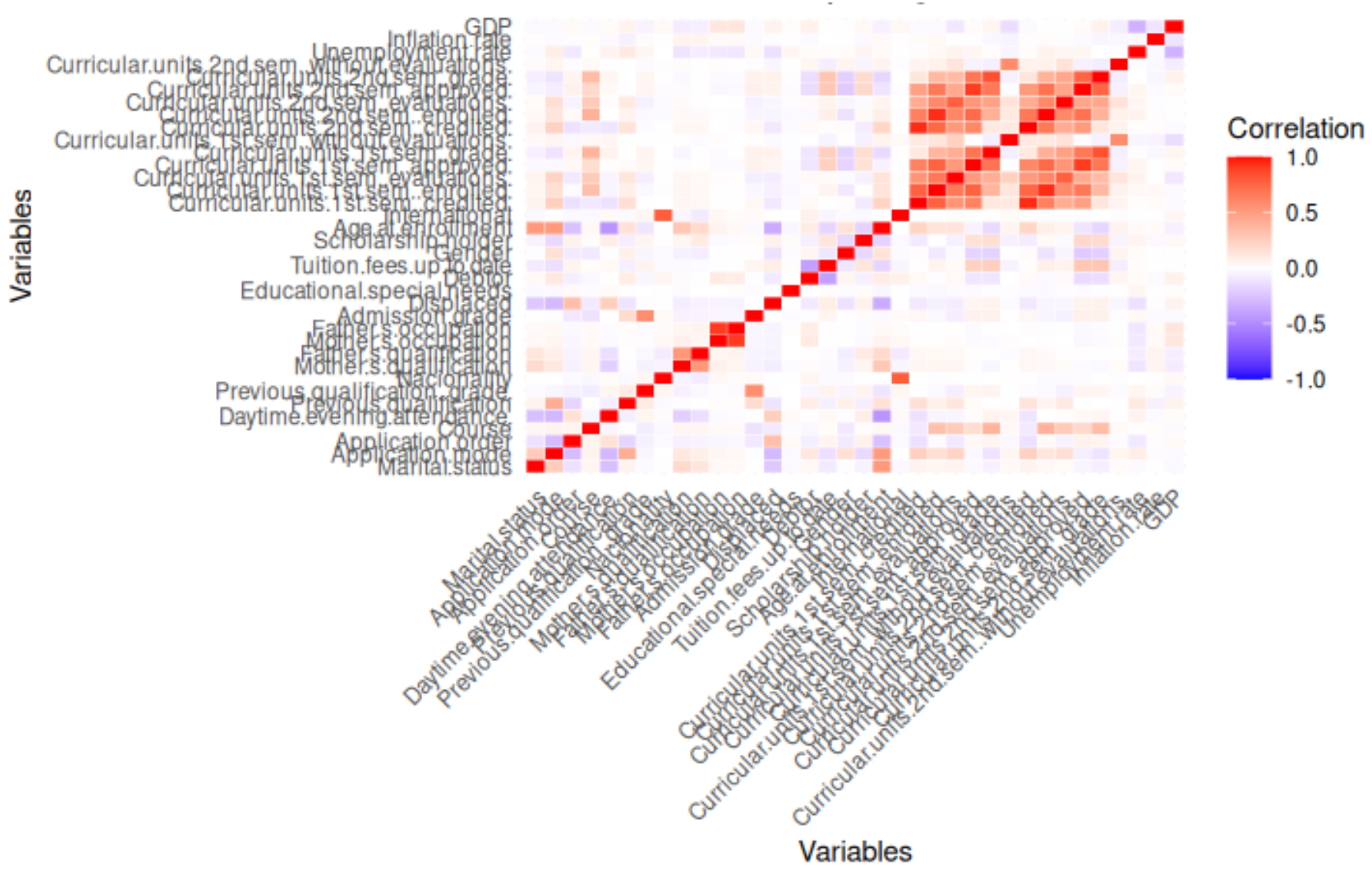
- Checked distributions of continuous variables.
- Removed outliers and highly correlated variables
- Recoded categorical variables (e.g., Marital Status, Nationality).

Admission Grade Box Plot



Heatmap of Missing Values

Correlation Heatmap: Original Data



Transforming Variables for Insights

Categorical Variables:

- Grouped and labeled variables (e.g., Courses, Application Mode).
- Converted categorical predictors into dummy variables.

Continuous Variables:

- Standardized and grouped into meaningful ranges (e.g., Grades, Age).

Optimization:

- Removed low-variance and highly correlated variables.
- Removed outliers

Cleaned Dataset

Columns: 85 after preprocessing.

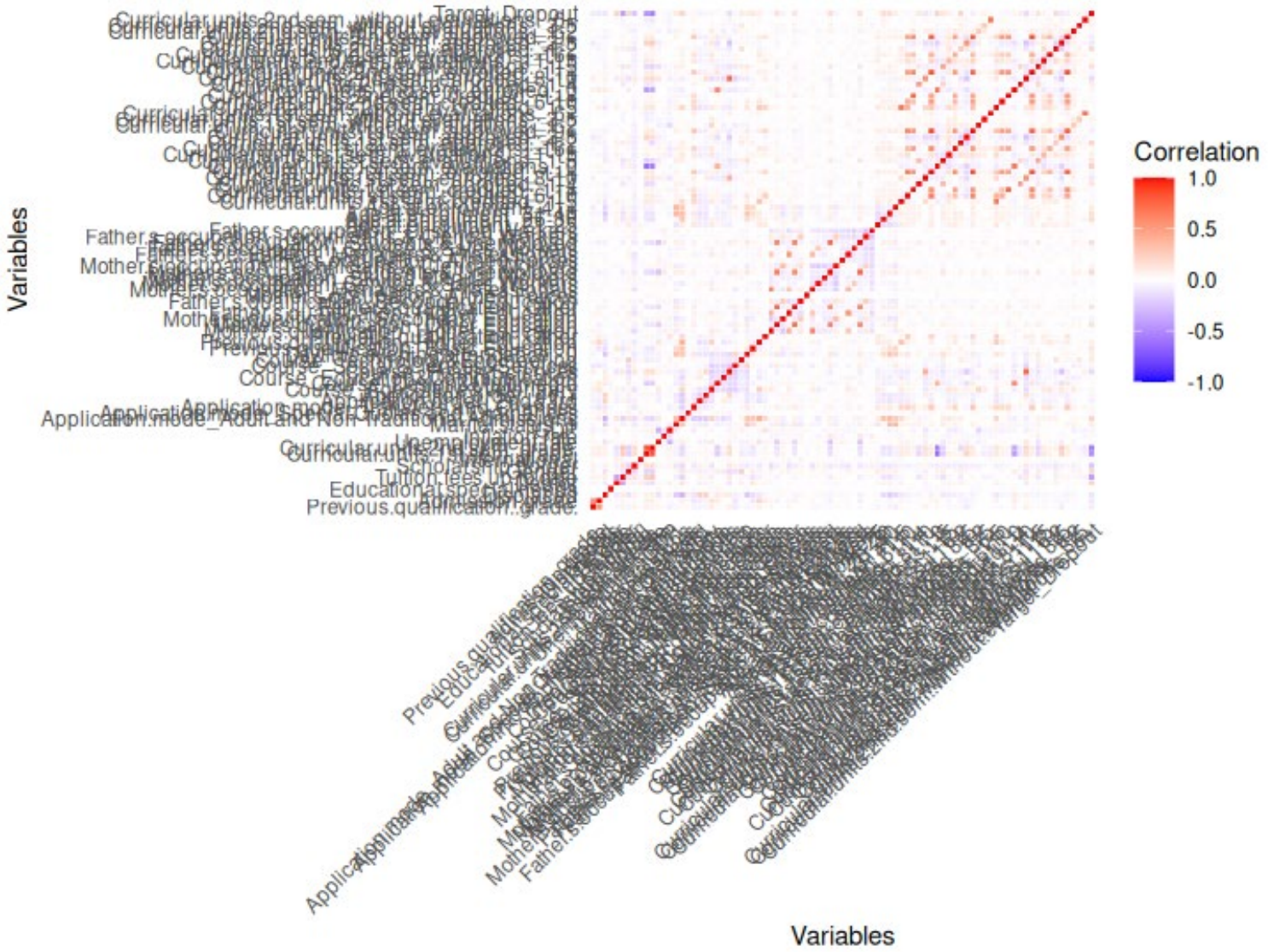
Rows: 4338 after removing outliers

- Retained only significant predictors by removing:
  - Low-variance variables.
  - Highly correlated variables (correlation > 0.9).
- Dummy variables created for categorical features to prepare for analysis.

Optimization Goals:

- Ensure data is free from redundancies.
- Improve model interpretability and efficiency.
- Further optimization (as needed) when exploring models

Correlation Heatmap



Modelling & Analysis

Model & Analysis

Logistic Regression

Total 48 variables, 35 significant and 13 insignificant

	OR	SE	95% CI, lowe	95% CI, uppe	p value
`Curricular.units.2nd.sem..approved. 1-2`	10.07	2.59	6.08	16.67	0.00
`Curricular.units.2nd.sem..approved._9+`	0.28	0.17	0.08	0.91	0.03
`Age.at.enrollment_26-30`	4.46	1.28	2.54	7.84	0.00
`Age.at.enrollment_41+`	3.40	1.21	1.69	6.83	0.00
`Mother.s.occupation_Students & Unemployed`	3.10	1.36	1.32	7.31	0.01
`Course_Education&Communication`	2.83	0.59	1.89	4.25	0.00

Model & Analysis

Logistic Regression

		Logistic Regression		
		Forward	Backward	Stepwise
Cutoff=0.5	**Val ER**	0.1348703	0.1308357	0.1348703
	Sensitivity	0.718638	0.7222222	0.718638
	Specificity	0.9345794	0.9388275	0.9345794
Cutoff=0.4	**Val ER**	0.1435159	0.1423631	0.1435159
	Sensitivity	0.7688172	0.7688172	0.7688172
	Specificity	0.8980459	0.8997451	0.8980459
Cutoff=0.3	**Val ER**	0.1665706	0.1613833	0.1665706
	Sensitivity	0.8225806	0.8243728	0.8225806
	Specificity	0.8385726	0.8453696	0.8385726
Cutoff=0.2	**Val ER**	0.2011527	0.2023055	0.2011527
	Sensitivity	0.8637993	0.8602151	0.8637993
	Specificity	0.7680544	0.7680544	0.7680544
Cutoff=0.1	**Val ER**	0.2939481	0.2945245	0.2939481
	Sensitivity	0.9336918	0.9301075	0.9336918
	Specificity	0.5981308	0.5989805	0.5981308

- At higher cutoff values (0.4, 0.5): correctly identify students who will not drop out (true negatives)
- **At lower cutoff values (0.1, 0.2, 0.3)**: correctly identify students who will drop out (true positives)
- Accuracy also decreases

Model & Analysis

KNN

Variable used: all variables from cleaned data

Best K chosen from k with lowest error rate on **validation** data

```
> bestK_dropOut
$k.optimal
[1] 7

$error.min
[1] 0.1576037

$error.all
      k=1       k=3       k=5       k=7       k=9      k=11      k=13      k=15      k=17
0.2018433 0.1741935 0.1668203 0.1576037 0.1640553 0.1640553 0.1622120 0.1640553 0.1640553
```

Model & Analysis

KNN

		kNN
Cutoff=0.5	Val ER	0.157604
	Sensitivity	0.587896
	Specificity	0.96206
Cutoff=0.4	Val ER	0.164977
	Sensitivity	0.682997
	Specificity	0.906504
Cutoff=0.3	Val ER	0.164055
	Sensitivity	0.685879
	Specificity	0.906504
Cutoff=0.2	Val ER	0.190783
	Sensitivity	0.835735
	Specificity	0.796748
Cutoff=0.1	Val ER	0.360369
	Sensitivity	0.945245
	Specificity	0.495935

Lower cutoffs – higher sensitivity:

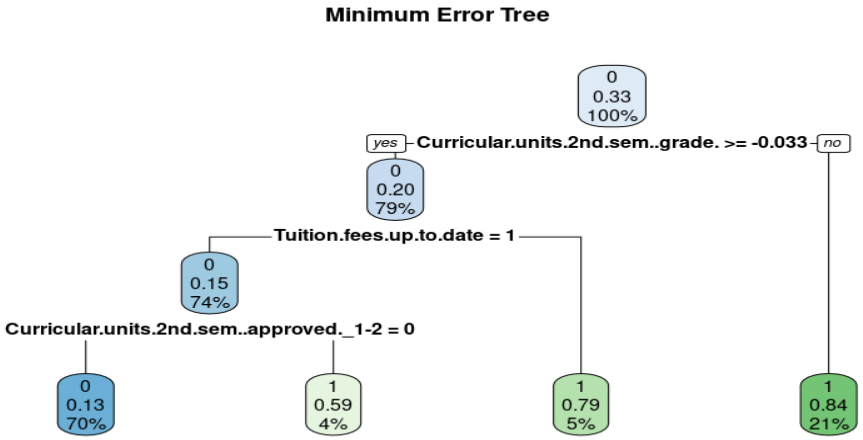
- correctly identify students who will drop out (true positives)

Higher cutoffs – higher specifity:

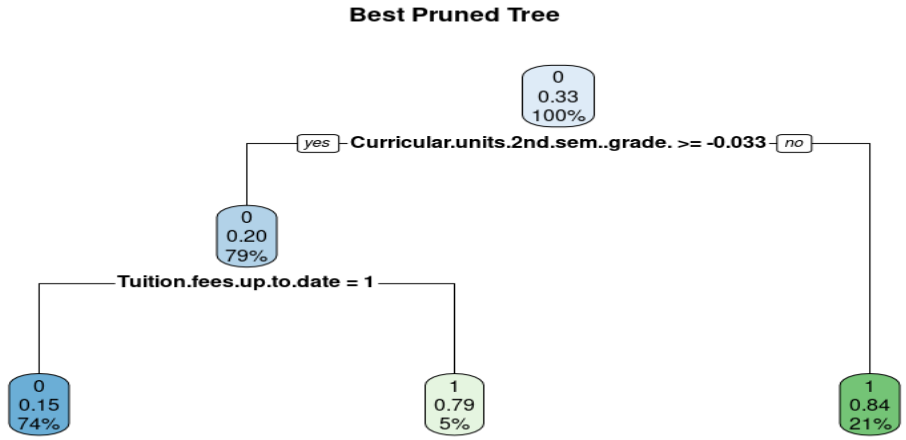
- correctly identify students who will not drop out (true negatives)

For student dropout prediction, it is generally more important to have higher sensitivity because preventing Dropouts is the priority => choose cutoff 0.1

Model Comparisons



Minimum Error Tree has more splitting because it gives precedence to all variable data but can lead to overfitting.



Best Pruned Tree sacrifices granularity to preserve flexibility of model by giving less precedence to insignificant variables resulting in less nodes.

Model & Analysis

CART

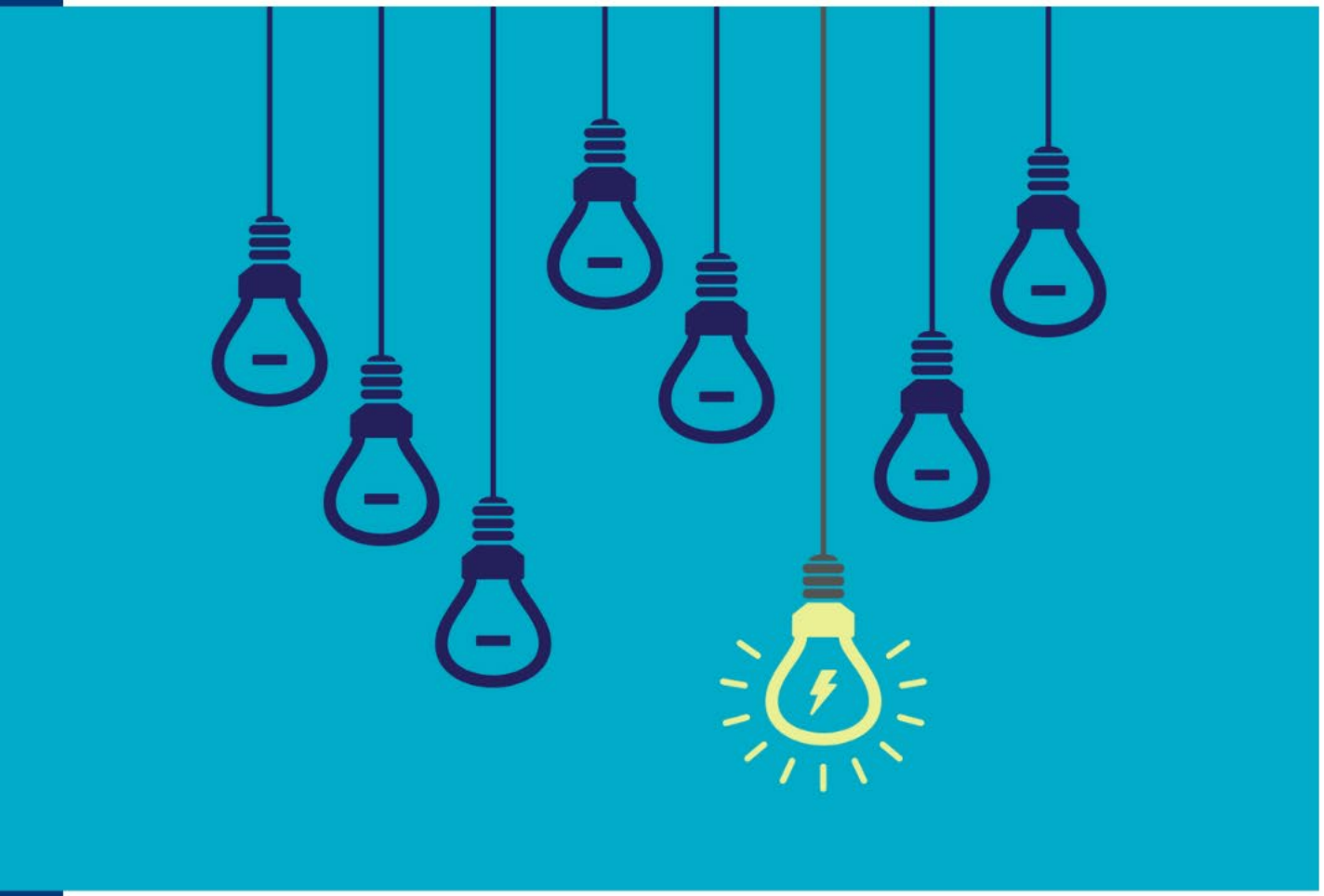
	Results	Classification Tree	
		Minimum Error	Best pruned
Cutoff=0.5	Val ER	0.1526	0.1526
	Sensitivity	0.7796	0.7019
	Specificity	0.8779	0.913
Cutoff=0.4	Val ER	0.1526	0.1526
	Sensitivity	0.7796	0.7019
	Specificity	0.8779	0.913
Cutoff=0.3	Val ER	0.1526	0.1526
	Sensitivity	0.7796	0.7019
	Specificity	0.8779	0.913
Cutoff=0.2	Val ER	0.1526	0.1526
	Sensitivity	0.7796	0.7019
	Specificity	0.8779	0.913
Cutoff=0.1	Val ER	0.6889	0.6889
	Sensitivity	1	1
	Specificity	0	0

Our CART models provided a lower sensitivity rate but a higher specificity rate indicating more success at identifying negative cases

Having different cut off levels did not have significant impact on metrics

Based on our desire to identify students more likely to drop out using BPT is better model because it responds better to new data

Key Insights



Key Learnings

Predicting student dropouts using classification methods (such as logistic regression, kNN classifications, and decision trees) can offer valuable insights into the factors contributing to student dropout:

- Identifying Key Predictors of Student Dropout
- Understanding Patterns and Relationships
- Improving Student Retention Strategies

Comparison with Prior Research

Findings Aligned with Industry Research

Key factors in our analysis that are consistent with other industry findings:

- Age
- Gender
- Parental Education and Occupation
- Work and Family Responsibilities
- Credit Accumulation
- Course Load

Findings Contradictory to Industry Research

Key factor in our findings that may contradict other industry's findings:

- Admission grade
- Course selection

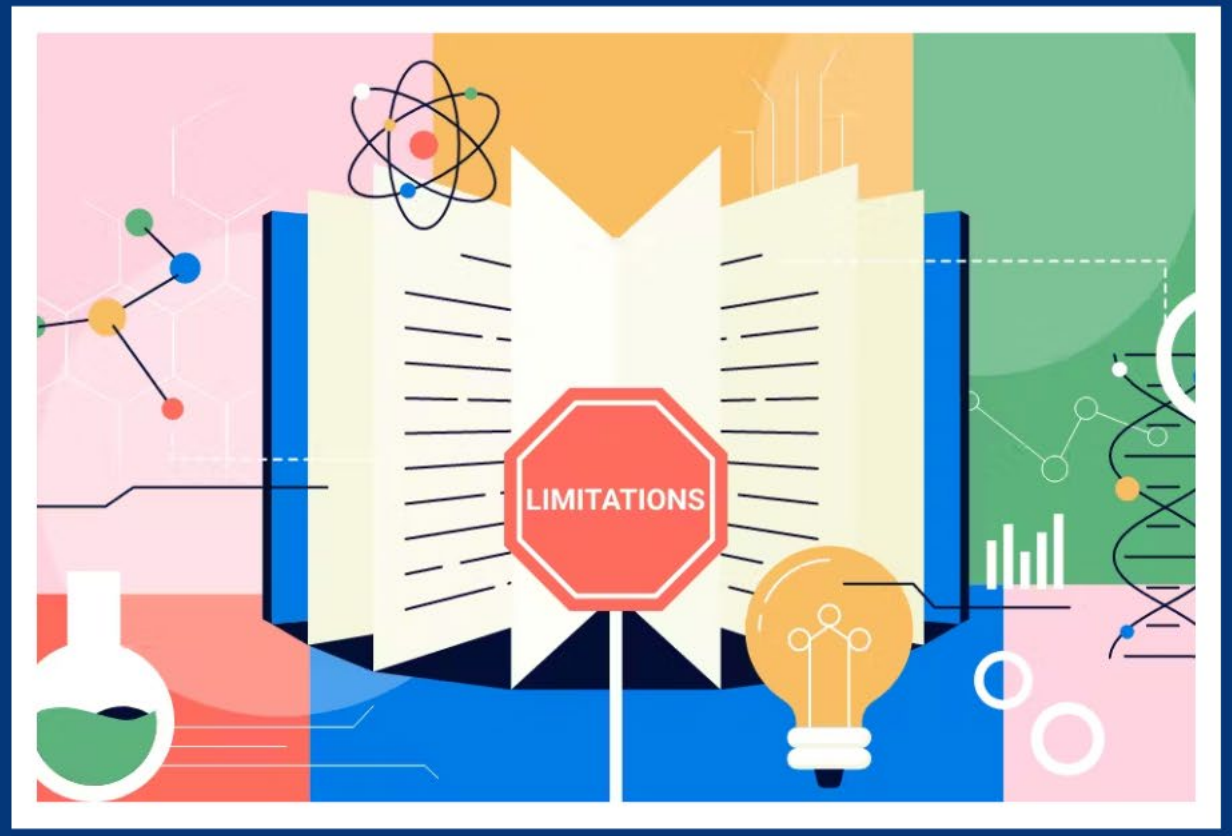
Reason for these discrepancies is that our data might involve a specific educational institution, region, or demographic group that behaves differently from broader industry trends. Dropout rates can vary significantly based on location or culture, or sample size.

Implications & Applications

- Educational Policy and Interventions
- Resource Allocation

Analysis Limitations

- Data availability and quality
- Model specific limitation
- Generalization – too demographic specific
- Unmeasured variables
- Static Data - Academic performance from first and second semester (lack of dynamic data)



Future Research

- Additional variables
- Incorporate dynamic variables (data overtime)
- Clustering models (in various level)

Conclusion & Takeaways

- Complex interplay factors
- Importance of targeted interventions
- Value of diverse modeling techniques

Thank You!