People and Results: Tying it All Up

Presented by Angel, Michelle, Joshua, Valentin & Tomy
In Partnership with the **Portugal university Board**

ANALYTICS FOR GOOD

Executive Summary

MANDATE	KEY ELEMENTS TO ASSESS	RECOMMENDATIONS	IMPACTS
Present recommendations to evaluate key factors affecting university graduation and dropout rates, aligning with Portugal University Board's (PUB) goal of boosting higher education graduates	What are the **main predictors** related to graduation? What is key to consider given the **public context** of this initiative? Ensure **ethical practices** while harnessing data what steps can be taken to initiate **actionable measures**?	**The DATA Methodology** (**D**issect-**A**ssemble-**T**ailor-**A**ctivate) A methodology **tying up** the findings of our work to **tangible actions** driving results for the **Portugal University Board**	Offer a **scalable** and reusable **tool** applicable **nationwide.** **An insightful framework** for guiding future initiatives **to enhance PUB's graduation rates.**

1. Dissect
2. Assemble
3. Tailor
4. Activate

1. Dissect

Gaining Insights into the Context and Defining the Role of Our Solution

Portugal is committed to improving its tertiary education rate

1999

The Ministry of Science, Technology, and Higher Education issued the Order no. 6659/99 **requesting HEIs undertake studies to identify root causes of failure / dropout**

2000-2021

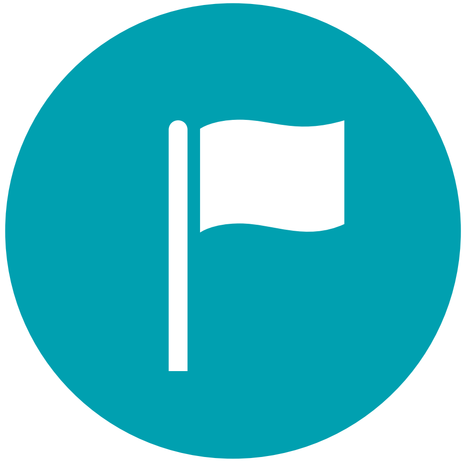
The share of 25-34 years old who completed tertiary education **increased from 13% to 47%**

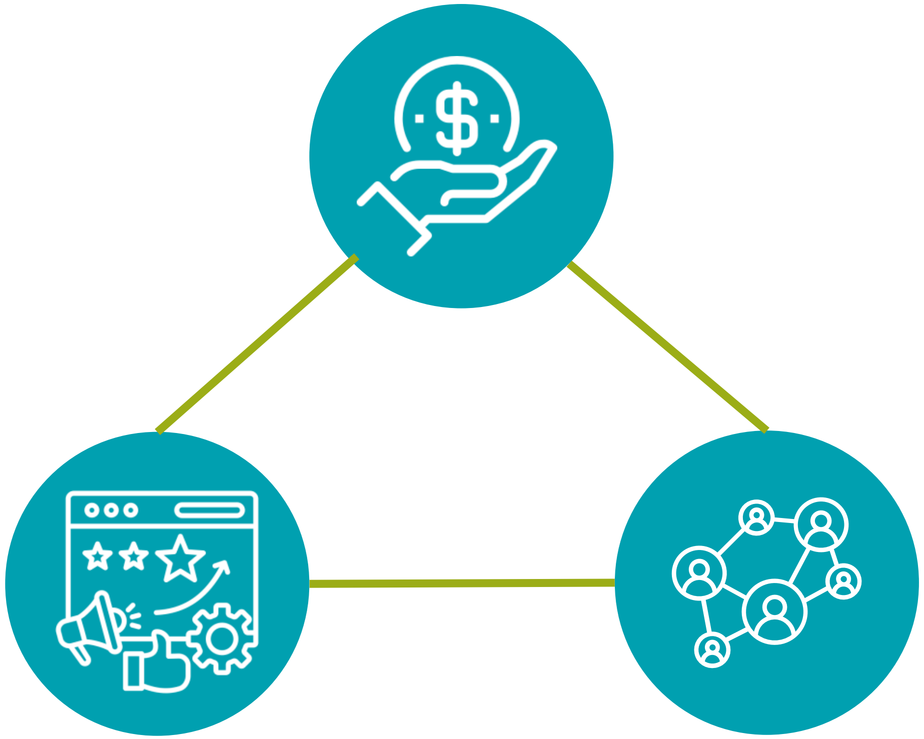
On track, but still behind the OECD average by 3 percentage points...

2013

The Portuguese government issued Resolution no. 60/2013 **requesting an annual report on dropout rates** in higher education

This matters for Portugal's economy and for individual HEIs





A Solution That Benefits All Stakeholders

Identifying at-risk students

- Use a classification model to determine which students have a higher risk of dropping out of university
- Analyze key variables to distinguish graduates from dropouts

Tailored Intervention Strategies

- Create a dynamic system to identify admitted students as “more likely to drop out”
- Targeted intervention to increase their chances of graduating

Benefits:

School Impact: Lowering dropout rates enhances the overall success and reputation of the institution

Student Impact: Improved academic outcomes and increased chances of graduation for at-risk students

A Solution That Benefits All Stakeholders

Generalization Across Schools:
Can extend successful interventions to a wider range of schools and students as most schools collect similar information from students during the admission process.

Common Intervention Strategies:

- **Financial support:** Can offer financial support in the form of scholarships, grants, work-study opportunities, and flexible payment plans.
- **Mentorship programs:** Can provide students with guidance and support from more experienced peers or faculty members.
- **Academic support services:** Support services, such as tutoring centers, writing labs, and academic skills workshops, can help students improve their study skills, understand course material, and succeed in their classes.

Finding Data to Tackle a Nationwide Issue

Researchers studying high university dropout rates in Portugal face the challenge of obtaining comprehensive data

What do we want to do?

1. Solution that will charm **PUB**

2. Build a **scalable** solution that have **strong explicative power**

What dataset was used?

1. A **dataset supported by program SATDAP** – Public Administration
2. Assembly of **many datasets, from various majors, all across the country**

https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

Step 1: Exploratory Data Analysis

Steps

1. Column Inspection

2. Frequency Analysis for Categorical Variables

3. Distribution & Summary Statistic of Numerical Variables

4. Outliers test

5. Correlation Matrix

Manipulations & Examples

```
Column Name: Marital status, Data Type: int64
Column Name: Application mode, Data Type: int64
Column Name: Application order, Data Type: int64
Column Name: Course, Data Type: int64
Column Name: Daytime/evening attendance , Data Type: int64
Column Name: Previous qualification, Data Type: int64
Column Name: Previous qualification (grade), Data Type: float64
Column Name: Nacionality, Data Type: int64
Column Name: Mother's qualification, Data Type: int64
Column Name: Father's qualification, Data Type: int64
Column Name: Mother's occupation, Data Type: int64
Column Name: Father's occupation, Data Type: int64
Column Name: Admission grade, Data Type: float64
Column Name: Displaced, Data Type: int64
Column Name: Educational special needs, Data Type: int64
Column Name: Debtor, Data Type: int64
Column Name: Tuition fees up to date, Data Type: int64
Column Name: Gender, Data Type: int64
Column Name: Scholarship holder, Data Type: int64
Column Name: Age at enrollment, Data Type: int64
Column Name: International, Data Type: int64
Column Name: Curricular units 1st sem (credited), Data Type: int64
Column Name: Curricular units 1st sem (enrolled), Data Type: int64
Column Name: Curricular units 1st sem (evaluations), Data Type: int6
Column Name: Curricular units 1st sem (approved), Data Type: int64
```

Understanding what type of data is held in each column

Step 1: Exploratory Data Analysis

Steps

1. Column Inspection

2. Frequency Analysis for Categorical Variables

3. Distribution & Summary Statistic of Numerical Variables

4. Outliers test

5. Correlation Matrix

Manipulations & Examples

```
Frequency Analysis for: Marital status
Marital status
1    3919
2     379
4      91
5      25
6       6
3       4
Name: count, dtype: int64
```

Identifying key trends and patterns within the dataset by reviewing the occurrences of specific categorical variables

Step 1: Exploratory Data Analysis

Steps

1. Column Inspection

2. Frequency Analysis for Categorical Variables

3. Distribution & Summary Statistic of Numerical Variables

4. Outlier test

5. Correlation Matrix

Manipulations & Examples

Detect potential anomalies through visual and computational analysis
[Q1 -1.5 *IQR ; Q3 + 1.5* IQR]

Step 1: Exploratory Data Analysis

Steps

1. Column Inspection

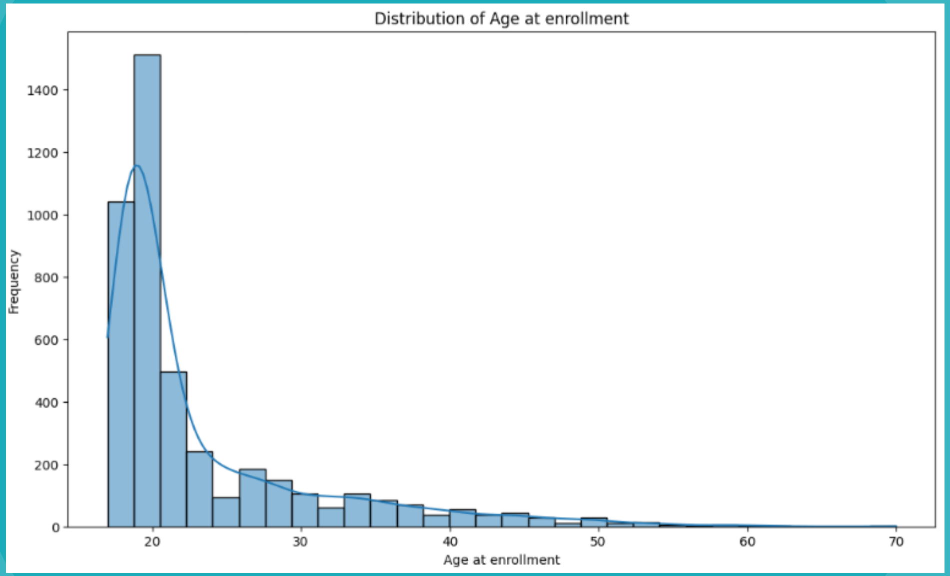
2. Frequency Analysis for Categorical Variables

3. Distribution & Summary Statistic of Numerical Variables

4. Outliers test

5. Correlation Matrix

Manipulations & Examples



Identify central tendencies and variability, detecting skewnesss and identifying which variables follow a normal distribution

Step 1: Exploratory Data Analysis

Steps

1. Column Inspection

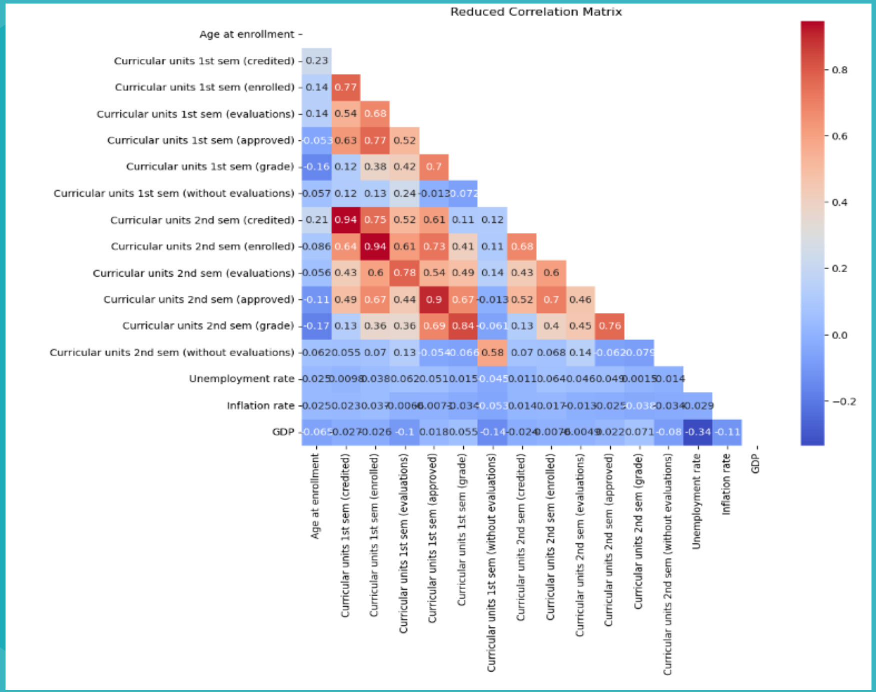
2. Frequency Analysis for Categorical Variables

3. Distribution & Summary Statistic of Numerical Variables

4. Outliers test

5. Correlation Matrix

Manipulations & Examples



Identify positive and negative linear relationships between pairs of variables

Step 1: Exploratory Data Analysis

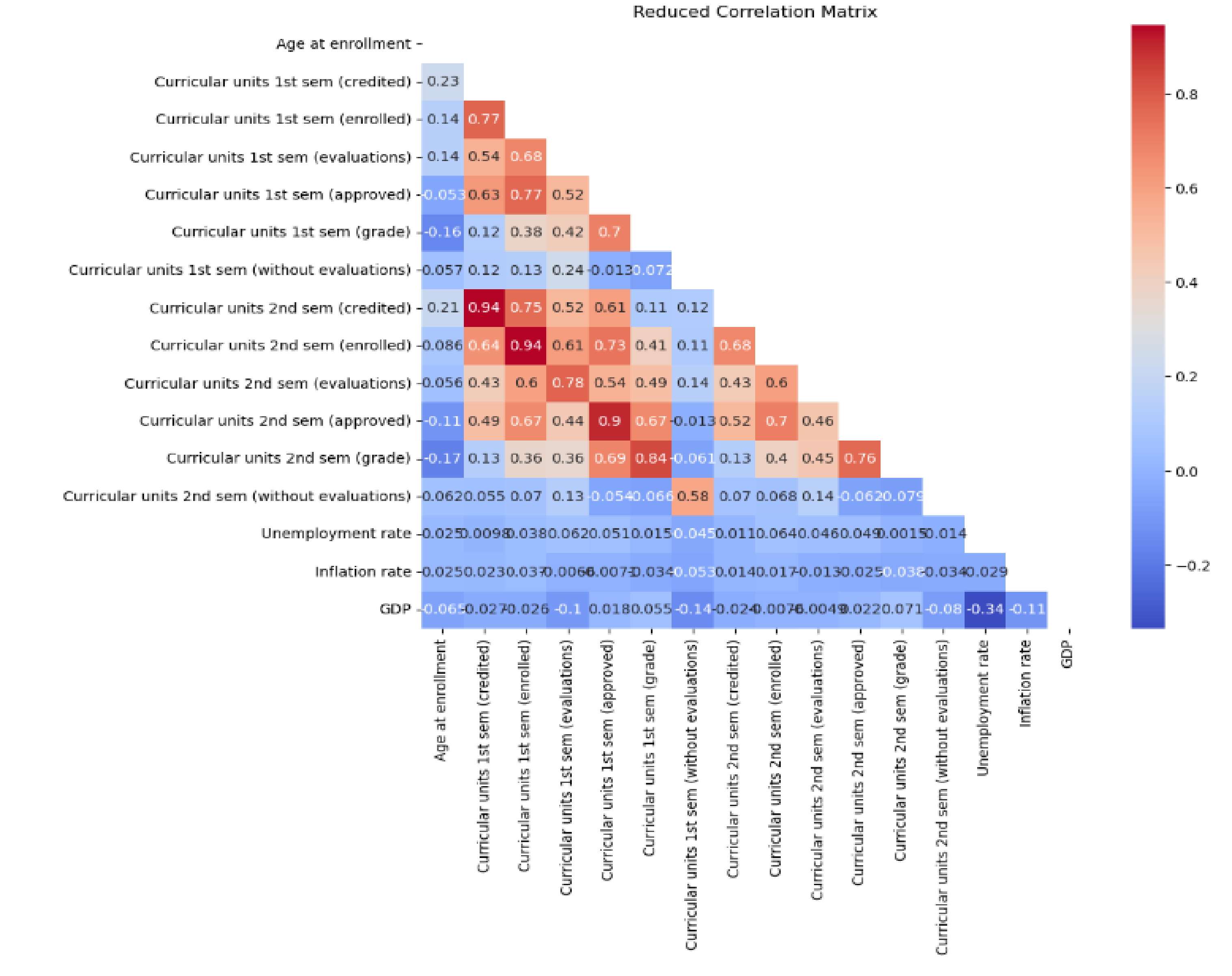
1. Colu

2. Freq
for Cat
Variab

3. Distr
Summ
Nume

4. Outl

5. Corr

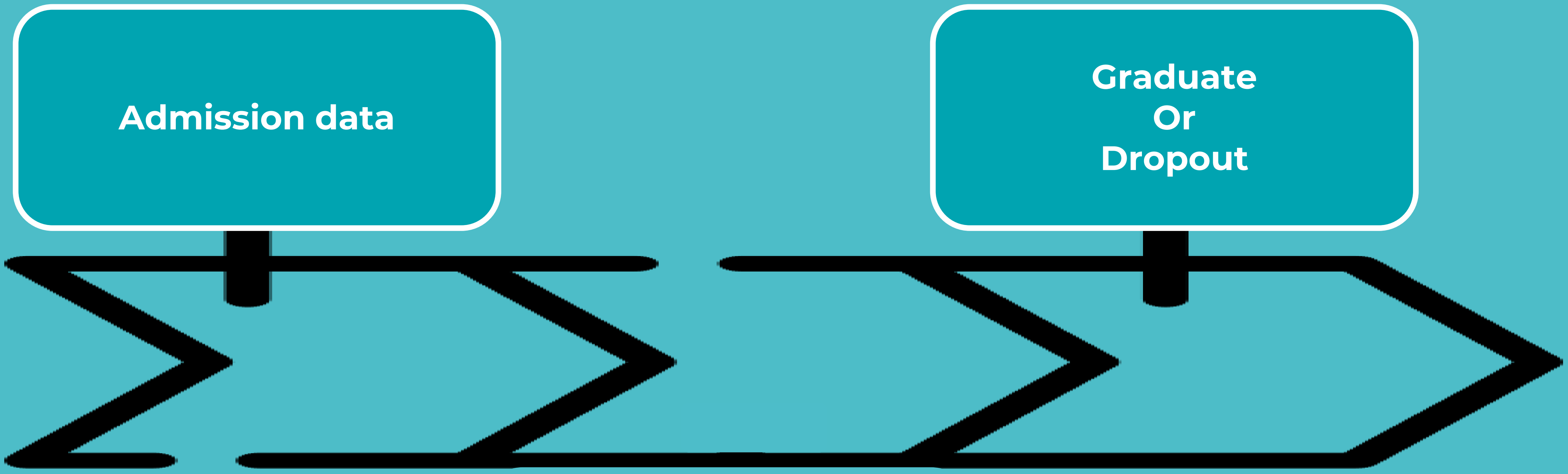


tive linear
of variables

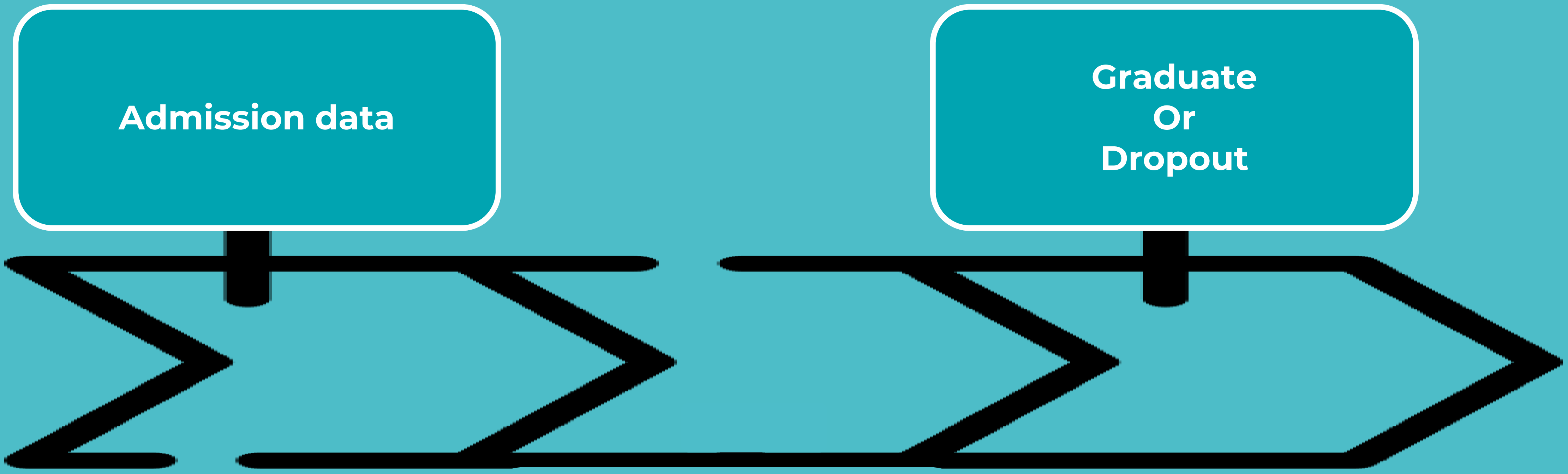
2. Assemble

Identifying Predictors to Build a Comprehensive Model

So… Our solution targets students at risk at the admission to enhance their chance of success…

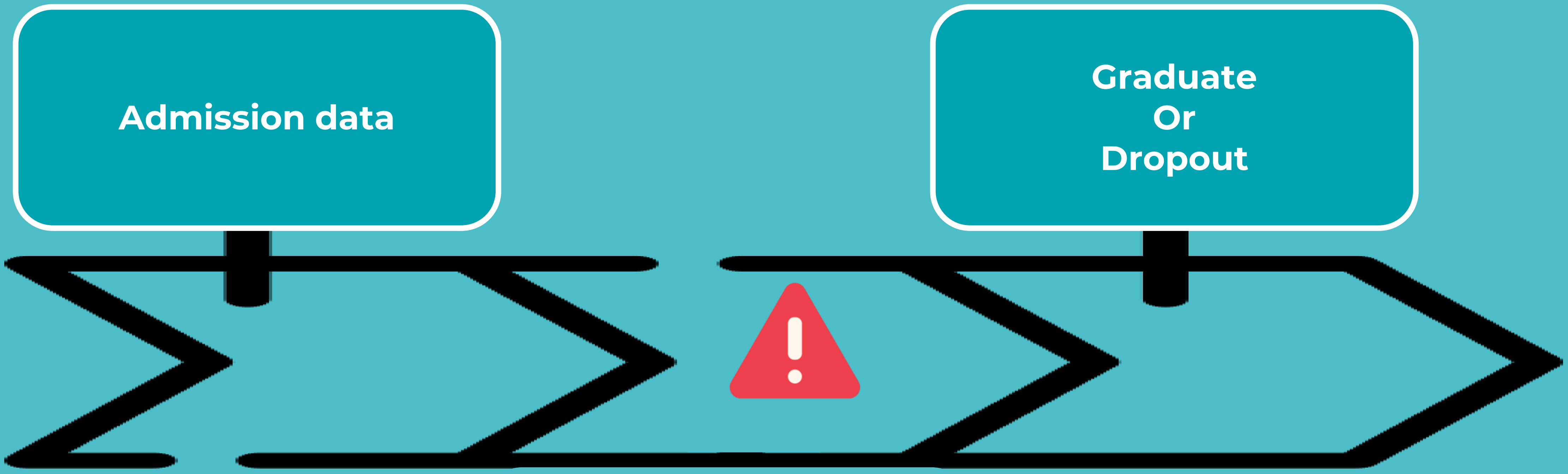


So... Our solution targets students at risk at the admission to enhance their chance of success...



And that would’ve been great if getting a degree was only based on the information we provided in our admission forms...

But the degree itself presents a challenge that must be taken into account to accurately understand the reasons for dropout...



But the degree itself presents a challenge that must be taken into account to accurately understand the reasons for dropout...

Therefore, we decided to equip the Portugal School Board with another model that would identify who became at risk due to the challenges of their studies !

But the degree itself presents a challenge that must be taken into account to accurately understand the reasons for dropout...

Model 1

Admission data

Graduate
Or
Dropout

Model 2

Data pertaining to courses and academic progress

Therefore, we decided to equip the Portugal School Board with another model that would identify who became at risk due to the challenges of their studies !

The extra step that goes a long way

Data Preprocessing- Variable Types & Initial Steps

Numerical Variables

Include:

- GDP
- Inflation Rate
- Unemployment Rate
- Age at enrollment

Categorical Variables

Include:

- Mother's occupation
- Father's occupation
- Gender
- Scholarship holder

Data Preprocessing- Formatting and Translating

Mapping and Dropping

- Numerical values in the dataset were mapped to descriptive strings for better interpretability.
- This mapping facilitated easier analysis and understanding of the data.
- “Enrolled” students were dropped from the dataset (to be able to accurately determine factors that influence whether a student will drop out or graduate).

Data Preprocessing- Formatting and Translating

Binning

- Binning was used to reduce noise and data dimensionality.
- Sparse data observations were combined for practicality.

```
mothers_qualification_dict = {
    1: "Secondary Education - 12th Year of Schooling or Eq.",
    2: "Higher Education - Bachelor's Degree",
    3: "Higher Education - Degree",
    4: "Higher Education - Master's",
    5: "Higher Education - Doctorate",
    6: "Frequency of Higher Education",
    9: "12th Year of Schooling - Not Completed",
    10: "11th Year of Schooling - Not Completed",
    11: "7th Year (Old)",
    12: "Other - 11th Year of Schooling",
    14: "10th Year of Schooling",
    18: "General commerce course",
    19: "Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv.",
    22: "Technical-professional course",
    26: "7th year of schooling",
    27: "2nd cycle of the general high school course",
    29: "9th Year of Schooling - Not Completed",
    30: "8th year of schooling",
    34: "Unknown",
    35: "Can't read or write",
    36: "Can read without having a 4th year of schooling",
    37: "Basic education 1st cycle (4th/5th year) or equiv.",
    38: "Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv.",
    39: "Technological specialization course",
    40: "Higher education - degree (1st cycle)",
    41: "Specialized higher studies course",
    42: "Professional higher technical course",
    43: "Higher Education - Master (2nd cycle)",
    44: "Higher Education - Doctorate (3rd cycle)"
}
```

```
mothers_qualification_dict = {
    1: "Secondary Education",
    2: "Higher Education",
    3: "Higher Education",
    4: "Higher Education",
    5: "Higher Education",
    6: "Unknown",
    9: "Did Not Finish High School",
    10: "Did Not Finish High School",
    11: "Did Not Finish High School",
    12: "Did Not Finish High School",
    14: "Did Not Finish High School",
    18: "General Commerce Course",
    19: "Did Not Finish High School",
    22: "Technical-Professional Course",
    26: "Did Not Finish High School",
    27: "Secondary School",
    29: "Did Not Finish High School",
    30: "Did Not Finish High School",
    34: "Unknown",
    35: "Illiterate",
    36: "Did Not Finish High School",
    37: "Did Not Finish High School",
    38: "Did Not Finish High School",
    39: "Technological Specialization Course",
    40: "Higher Education",
    41: "Specialized Higher Studies Course",
    42: "Professional Higher Technical Course",
    43: "Higher Education",
    44: "Higher Education"
}
```

Data Preprocessing- Preparing Data for the Model

Standardization of Numerical Variables

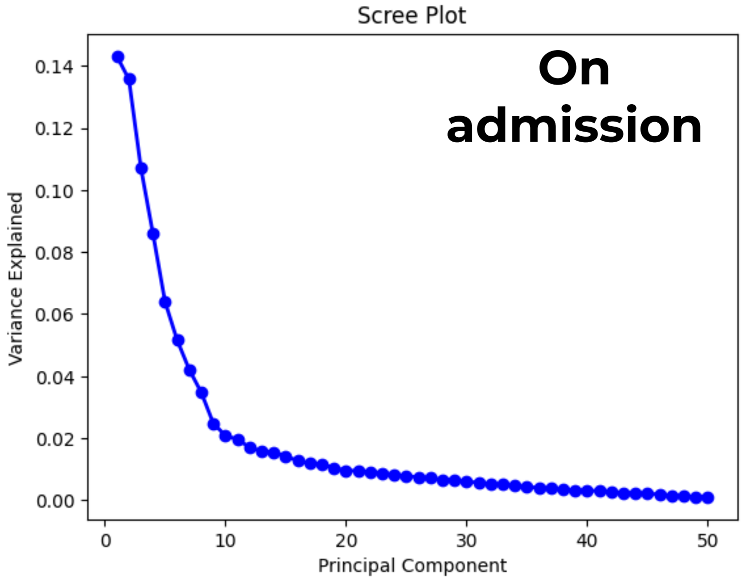
- Numerical values in the dataset were mapped to descriptive strings for better interpretability.
- This mapping facilitated easier analysis and understanding of the data.

Dummifying Categorical Variables

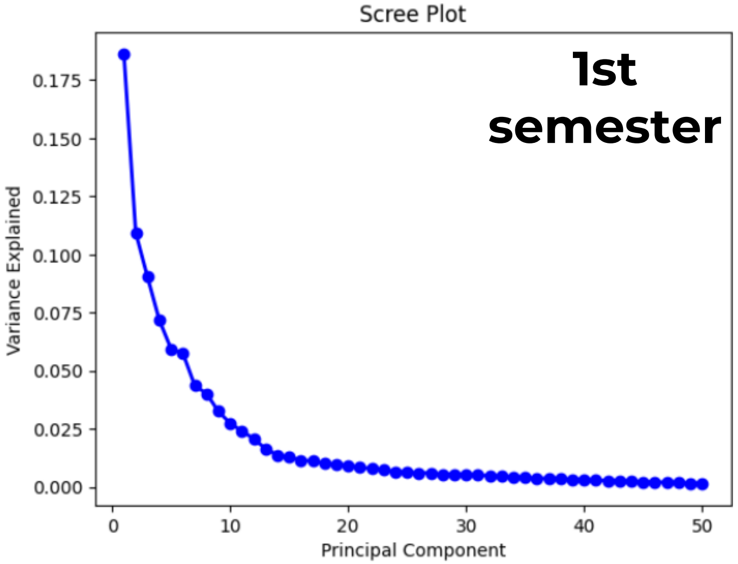
- Dummification was done using `pd.get_dummies`
- This process converted labels to numerical formats for model compatibility
- Making the dataset more suitable for efficient analysis.

Variable selection: Number of Variables

- Used scaled data (scaled numerical variables and 0-1 dummies)
- Created a Scree plot for each model:



We chose n=9, on admission (to account for the extra variables)



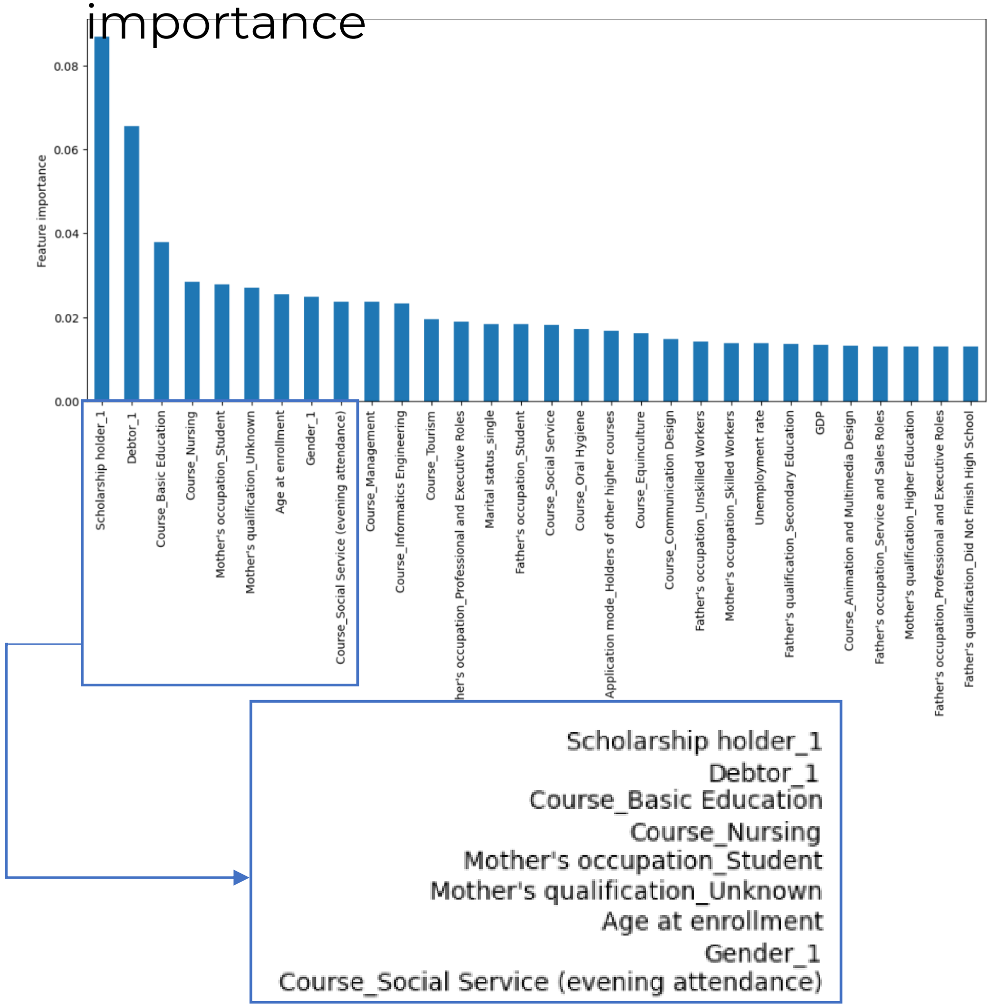
We chose n=13 with 1st semester grades and tuition payment information (to account for the extra variables)

- We use n as the number of variables to be used in the rest of the model

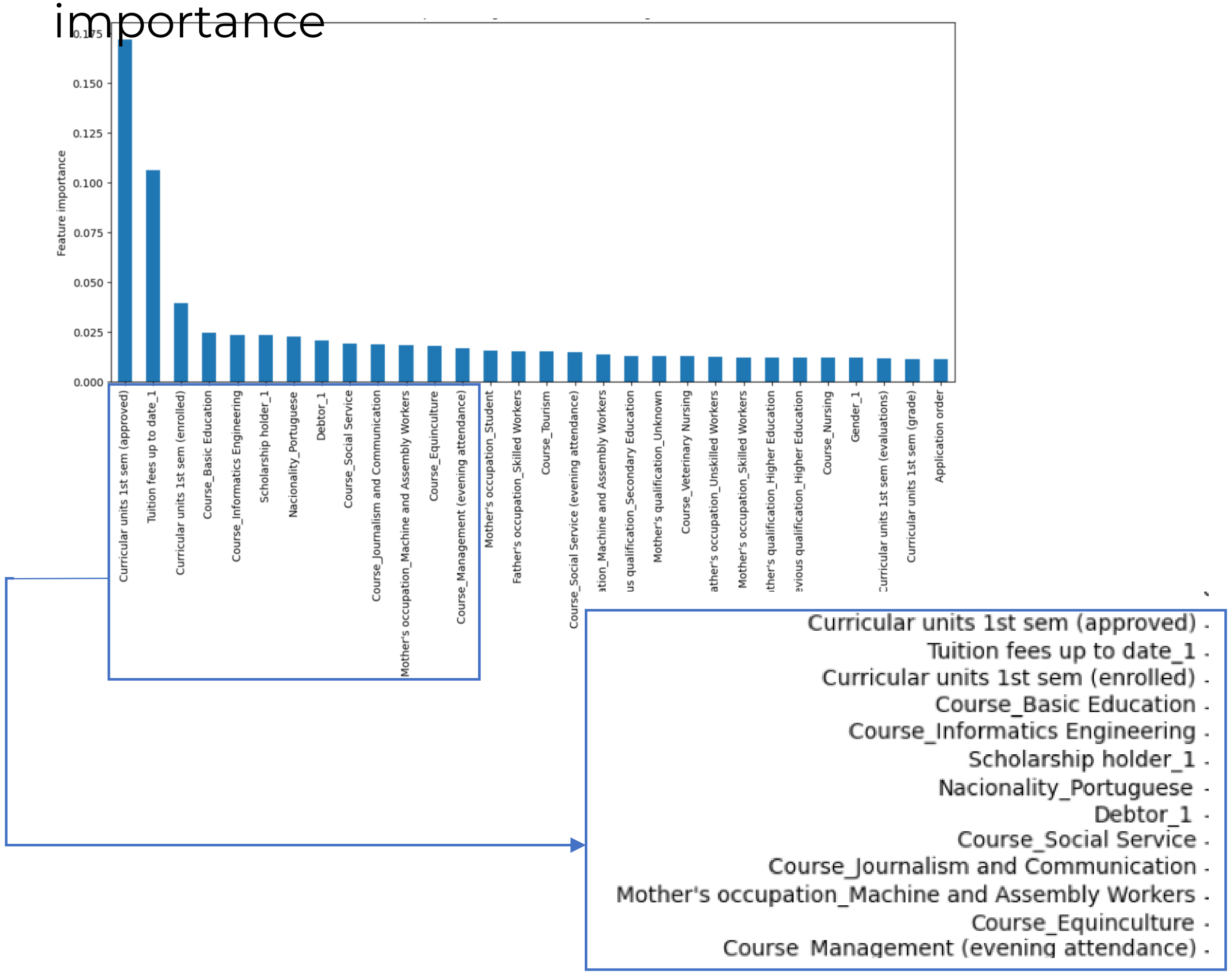
Variable selection: Top n Variables

- Using two tree-based models (random forest and boosting) to get the top n variables:

On entrance n=9: **boosting** feature importance



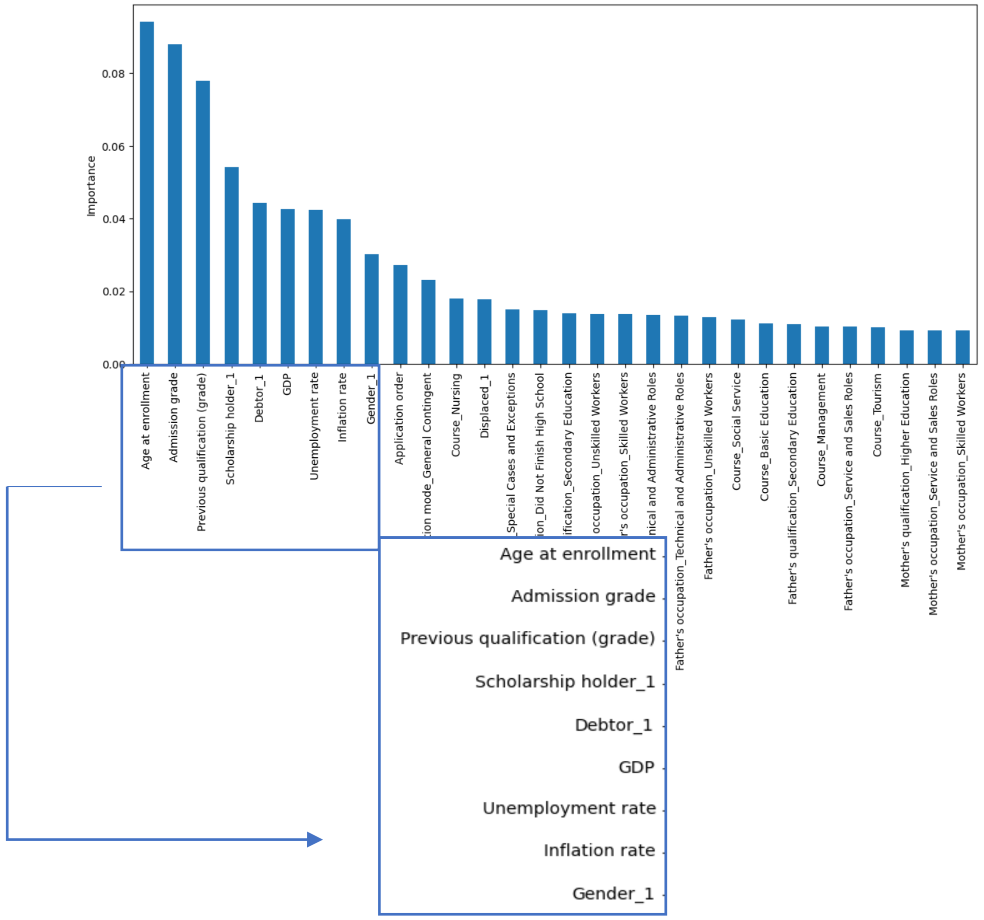
First semester n=13: **boosting** feature importance



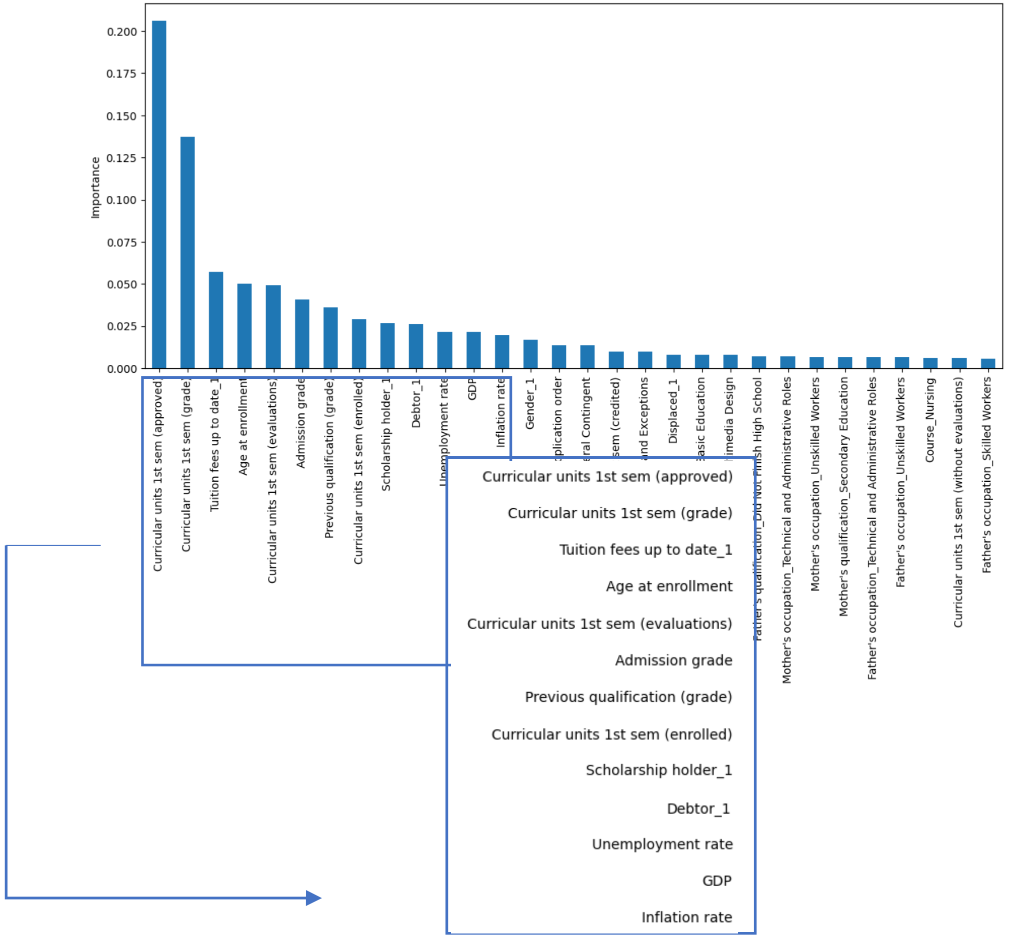
Variable selection: Top n Variables

- Using two tree-based models (random forest and boosting) to get the top n variables:

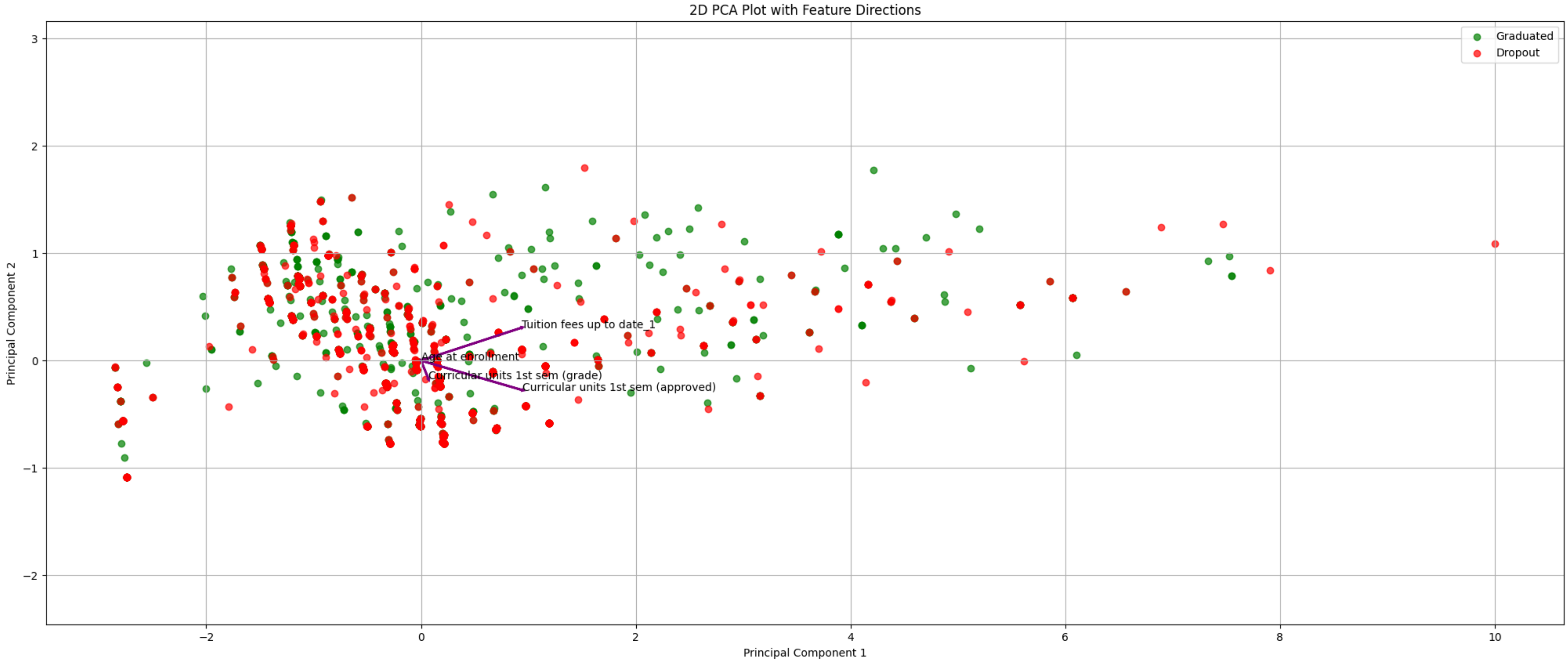
On entrance n=9: **random forest** feature importance



First semester n=13: **random forest** feature importance

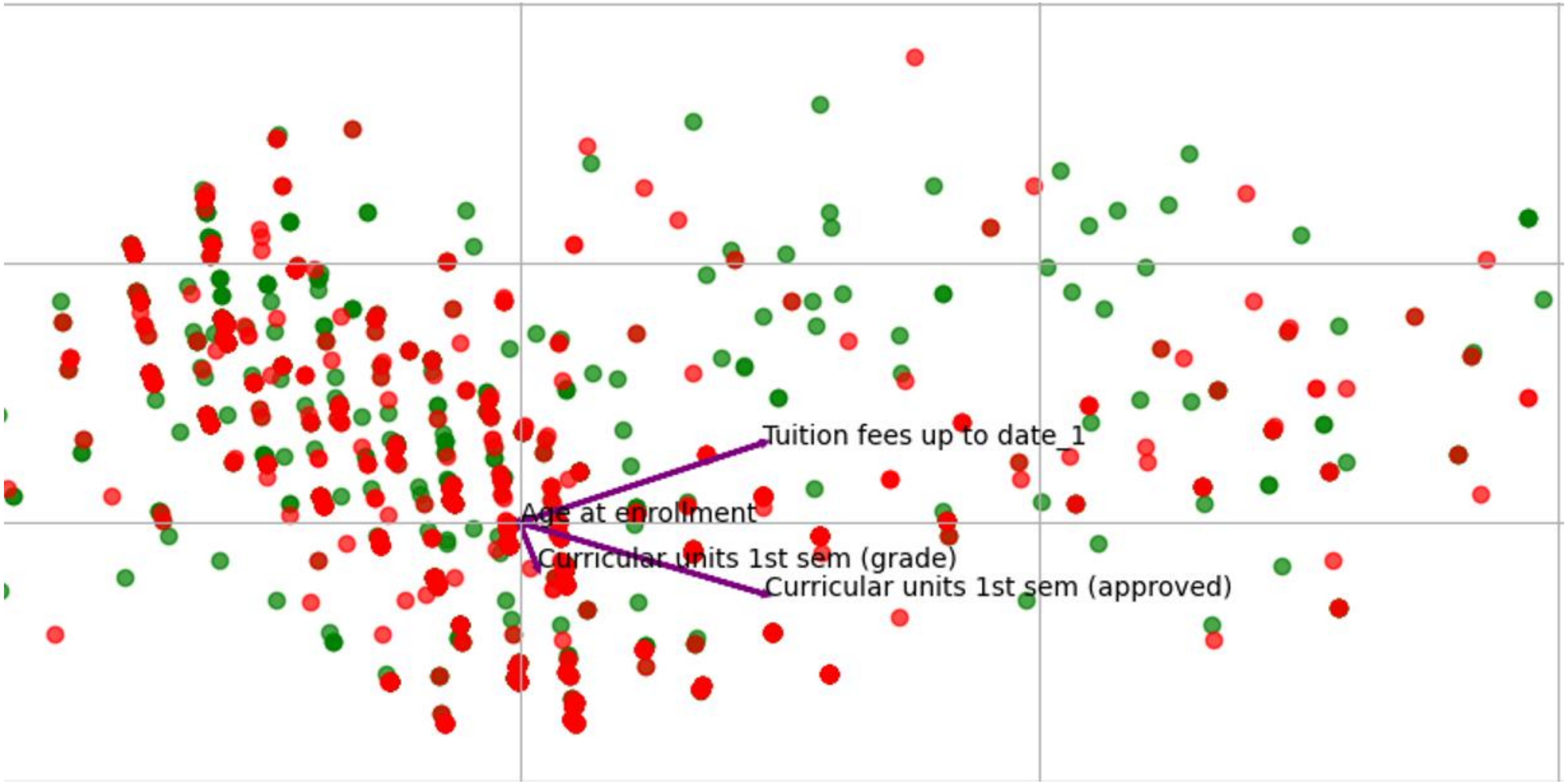


Variable interpretation PCA



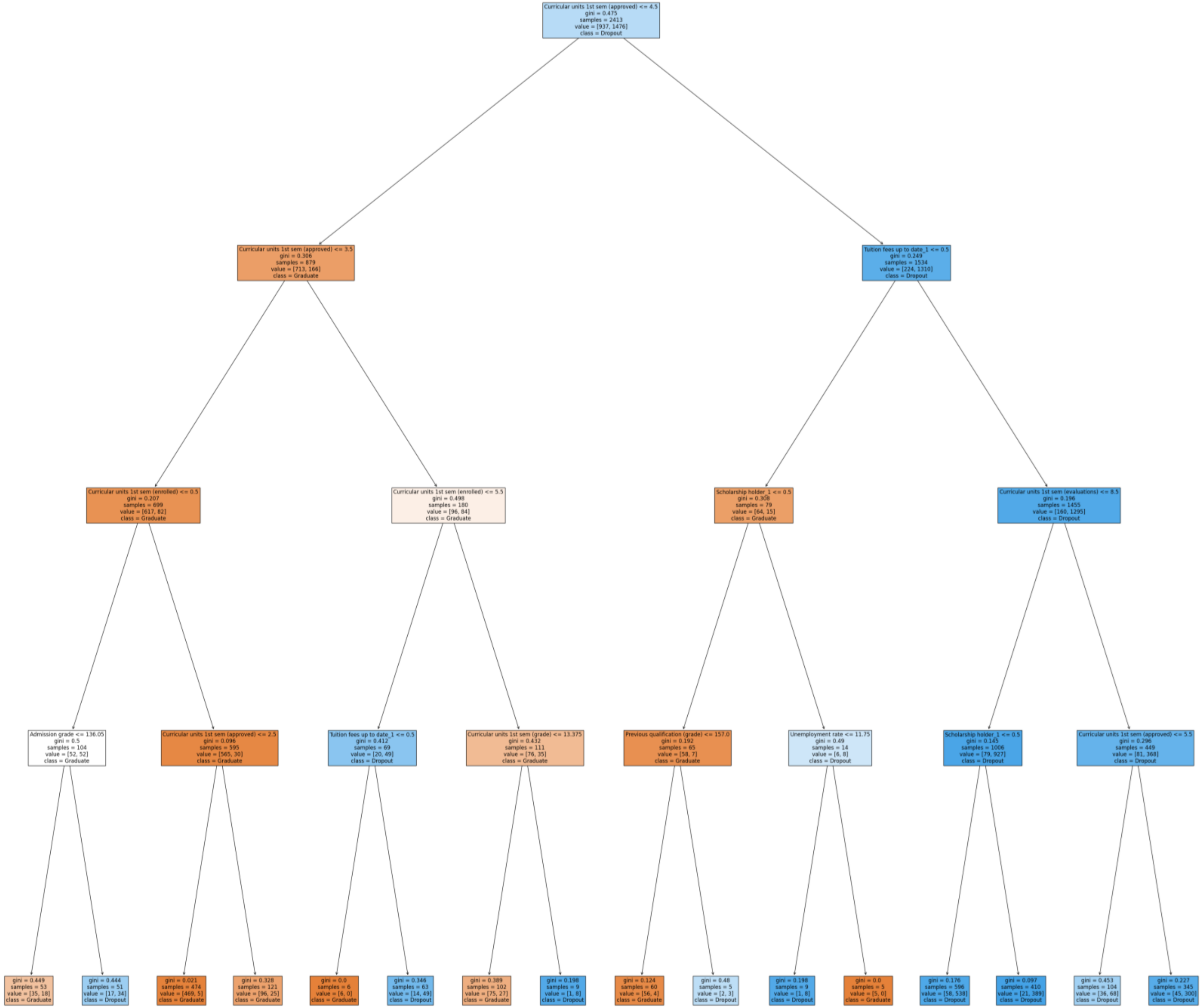
- PCA with top 4 variables from boosting (first semester)

Variable interpretation PCA



- No pattern can be observed in all PCA plots

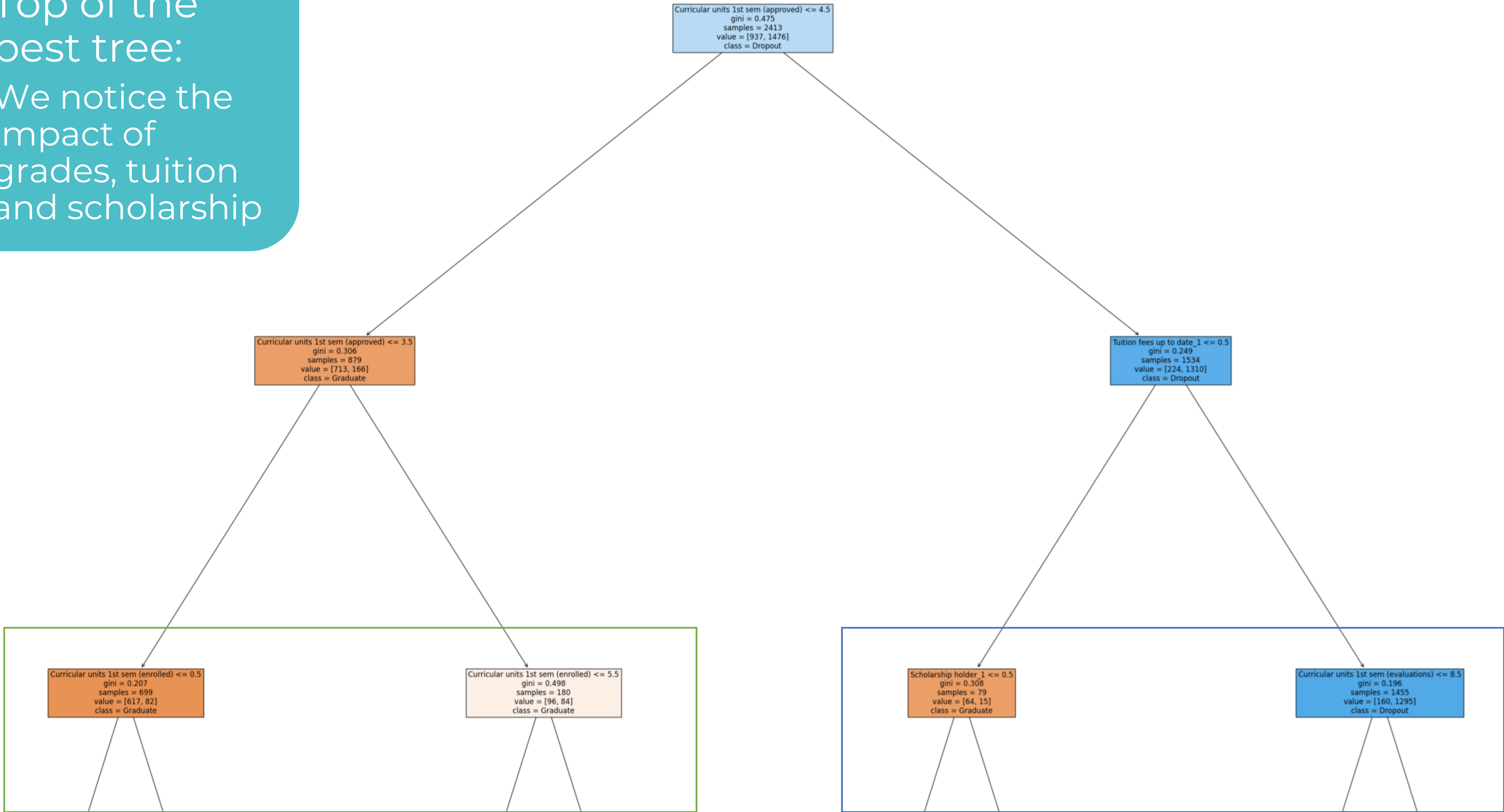
Variable interpretation trees

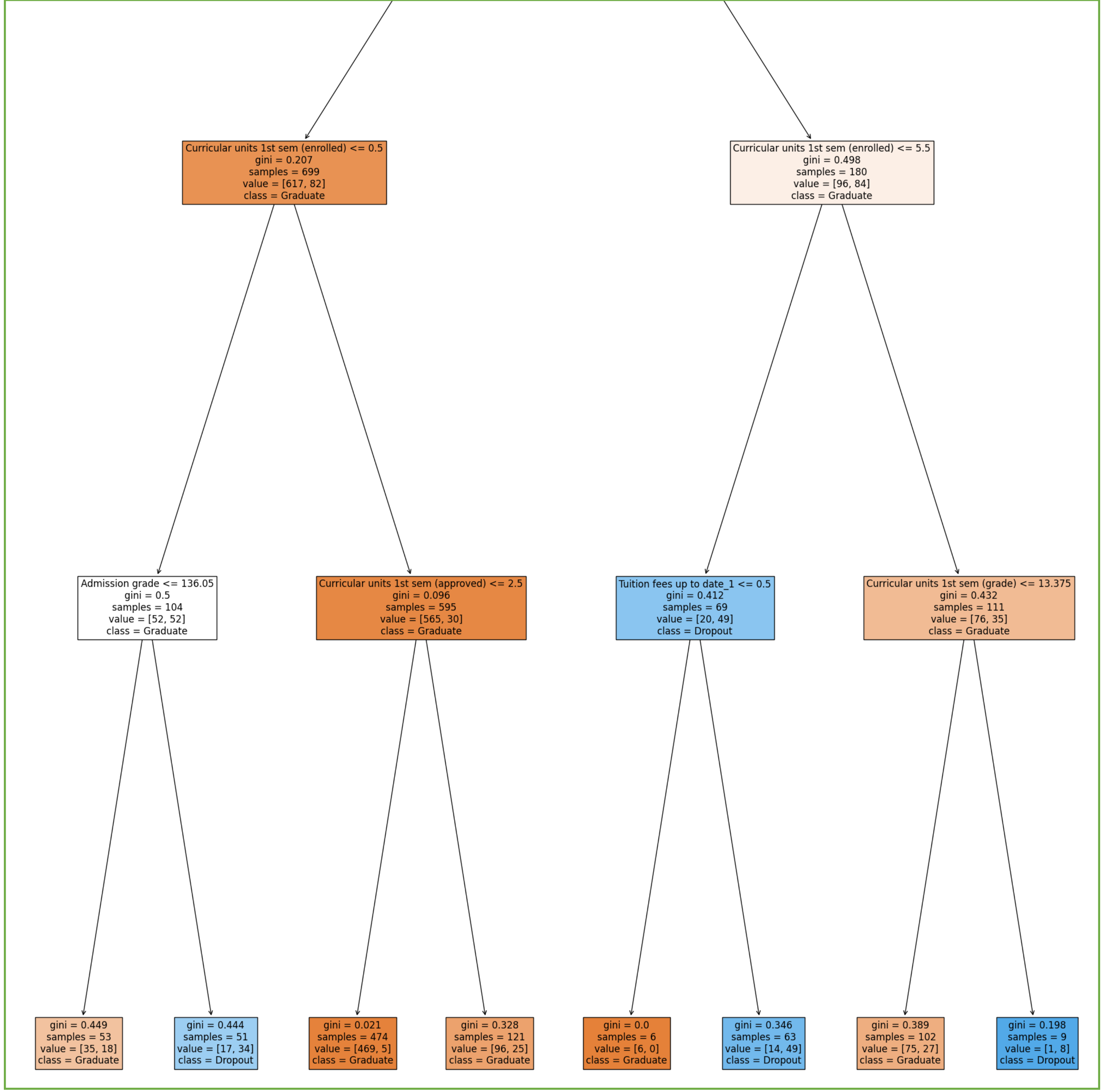


Example here:
1st semester, top 13 variables from random forest feature selection
(One of 4 trees)

Top of the best tree:

We notice the impact of grades, tuition and scholarship



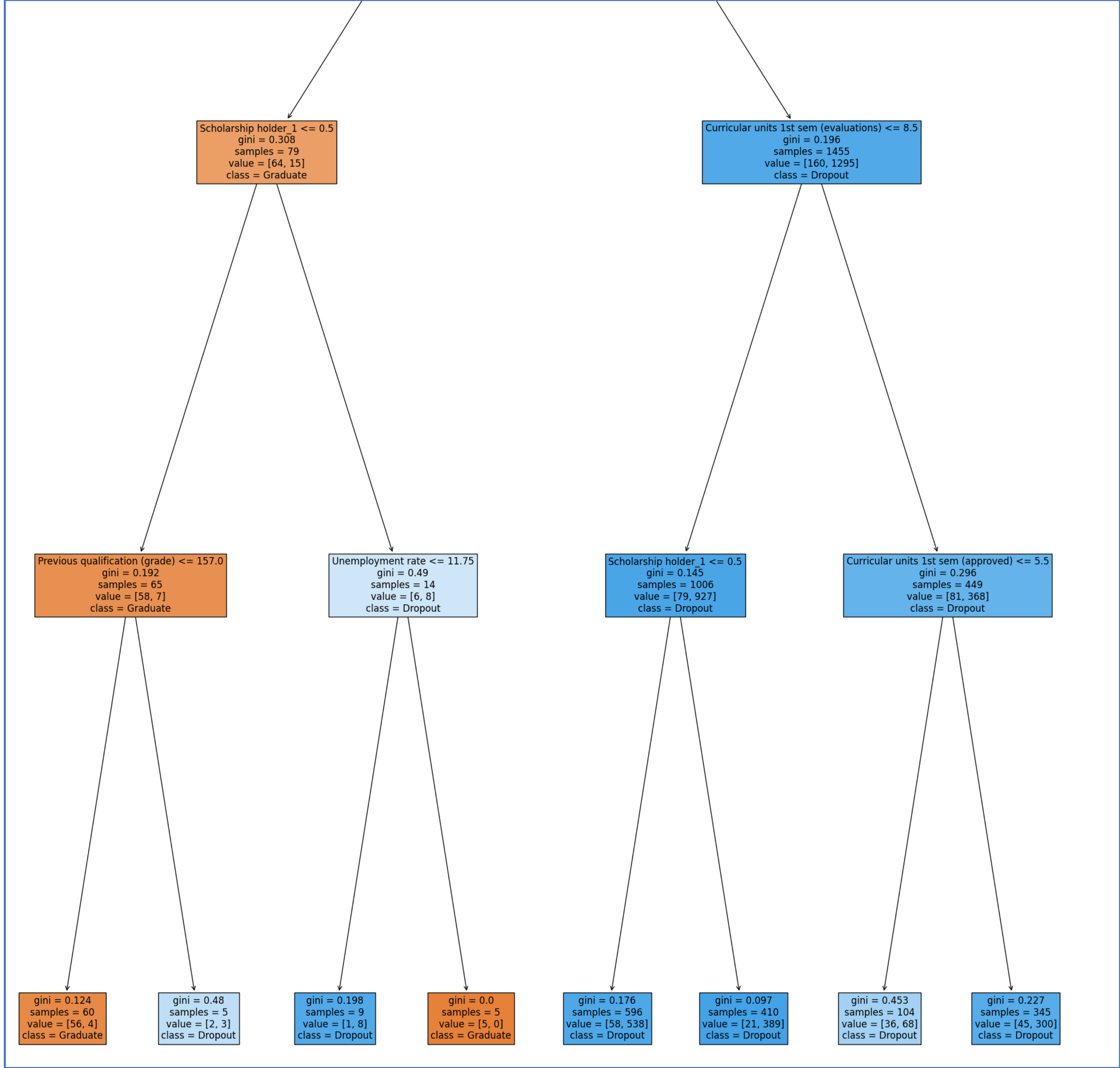


Bottom left of the best tree:

We notice the impact of grades, tuition and scholarship

Bottom right of the best tree:

We notice the impact of grades, unemployment and scholarship



Final Modeling Decision

Variable selection recap:

- We used PCA to find the number of variable (n) to include in each model
- We used random forest and boosting to get the best n features
- We now need to choose the most appropriate model

Model choice:

- In this context considering interpretation is vital we have use supervised learning.
- Since our goal is classification, we opted for logistic regression.
- Logistic regression allows us to use coefficients and understand the implications of each variable.

Logistic Regression Model Selection at T-0

	Precision	Recall	F1-score	Support
0 (dropout)	0.75	0.86	0.80	712
1 (graduate)	0.72	0.57	0.64	477
Accuracy			0.74	1189
Macro average	0.74	0.71	0.72	1189
Weighted average	0.74	0.74	0.73	1189

T0 – Logistic Regression Model Based on Gradient Boosting Features

	Precision	Recall	F1-score	Support
0 (dropout)	0.68	0.53	.60	477
1 (graduate)	0.73	0.84	.78	712
Accuracy			0.71	1189
Macro average	0.71	0.68	0.69	1189
Weighted average	0.71	0.71	0.71	1189

T0 – Logistic Regression Model Based on Random Forest Selected Features

Model selection was determined based off recall scores for our T0 model the Logistic Regression model based on the features selected from Gradient Boosting

Logistic Regression Model Selection at T-0

	Precision	Recall	F1-score	Support
0 (dropout)	0.75	0.86	0.80	712
1 (graduate)	0.72	0.57	0.64	477
Accuracy			0.74	1189
Macro average	0.74	0.71	0.72	1189
Weighted average	0.74	0.74	0.73	1189

T0 – Logistic Regression Model Based on Gradient Boosting Features

	Precision	Recall	F1-score	Support
0 (dropout)	0.68	0.53	.60	477
1 (graduate)	0.73	0.84	.78	712
Accuracy			0.71	1189
Macro average	0.71	0.68	0.69	1189
Weighted average	0.71	0.71	0.71	1189

We carried on with the model based on random forest feature selection because:

In our case **recall** (ability to identify at-risk student) is the most important measure.

Precision has also a decent level which makes sure we are spending public funds diligently

T0 – Logistic Regression Model Based on Random Forest Selected Features

Model selection was determined based off recall scores for our T0 model the Logistic Regression model based on the features selected from Gradient Boosting

Logistic Regression Model Selection at T-1

	Precision	Recall	F1-score	Support
0 (dropout)	0.86	0.95	0.90	712
1 (graduate)	0.90	0.77	0.83	477
Accuracy			0.88	1189
Macro average	0.88	0.86	0.87	1189
Weighted average	0.88	0.88	0.87	1189

T1 – Logistic Regression Model Based on **Random Forest** Feature Selection

	Precision	Recall	F1-score	Support
0 (dropout)	0.88	0.79	0.83	477
1 (graduate)	0.87	0.93	0.90	712
Accuracy			0.87	1189
Macro average	0.87	0.86	0.86	1189
Weighted average	0.88	0.87	0.87	1189

T1 – Logistic Regression Model Based on **Gradient Boosting** Feature Selection

- Model selection was determined based off recall scores
- For our T1 model the Logistic Regression based on the features identified from the Random Forest were used

Logistic Regression Model Selection at T-1

	Precision	Recall	F1-score	Support
0 (dropout)	0.86	0.95	0.90	712
1 (graduate)	0.90	0.77	0.83	477
Accuracy			0.88	1189
Macro average	0.88	0.86	0.87	1189
Weighted average	0.88	0.88	0.87	1189

T1 – Logistic Regression Model Based on **Random Forest** Feature Selection

We carried on with the model based on random forest feature selection because:

In our case **recall** (ability to identify at-risk student) is the most important measure.

Precision has also a decent level which makes sure we are spending public funds diligently

	Precision	Recall	F1-score	Support
0 (dropout)	0.88	0.79	0.83	477
1 (graduate)	0.87	0.93	0.90	712
Accuracy			0.87	1189
Macro average	0.87	0.86	0.86	1189
Weighted average	0.88	0.87	0.87	1189

T1 – Logistic Regression Model Based on **Gradient Boosting** Feature Selection

- Model selection was determined based off recall scores
- For our T1 model the Logistic Regression based on the features identified from the Random Forest were used

3. Tailor

Adjusting Models to Better Reflect the Portuguese Student's Reality

Model 1 - Admission

Findings from Analysis – At Admission

6x

more likely
if in debt

3x

more likely if
mother's education
unknown

Financial status, family education history, family occupations significantly affect the odds of dropping out

Socioeconomic factors are the strongest predictors for likelihood of dropping out

Findings from Analysis – At Admission

80%

less likely as scholarship holder

Financial status, family education history, family occupations significantly affect the odds of dropping out

Socioeconomic factors are the strongest predictors for likelihood of dropping out

Model 1 - Admission

Findings from Analysis – At Admission

70%
less likely in Social Services

47%
less likely in Nursing

Students in these programs are much less likely to drop out than those in other programs

Nursing & Social Services are the courses with the highest likelihood of producing graduates

Findings from Analysis – At Admission

~2x

more likely
if male

Males have a much higher likelihood of dropping out, while the impact of age at enrollment is modest

Varying impact from demographic predictors

4. Activate

Providing Actionable Takeaways for the Portugual university Board

Model 1 - Admission

Implications & Interventions – At Admission

Enhance Community Support Structures

Counteract insufficient support networks for navigating higher education

Learning from successful programs

Undertake case study of pedagogical techniques that could be piloted in other programs

Subsidize services for at-risk students

Use needs-based approach to provide subsidized support services to most vulnerable students

Investigate gender disparities

Determine whether gender variable proxying other factors, else target interventions for male students

Targeted support to promote the success of incoming students most at risk of not graduating

Model 1 - Admission

Implications & Interventions – At Admission

Leverage the model for strategic planning

Maintain ethical and fairness considerations

Targeted support to promote the success of incoming students most at risk of not graduating

Model 2 – After 1st Semester

The Extended Model- After 1st Semester

First semester grades introduced

New features about first-semester grades and tuition payment were added

- Number of credited* courses
- Number of enrolled courses
- Number of courses with evaluations
- Number of approved* (passed) courses
- Average grade for the semester
- Number of courses without evaluations

Overlap with T0 features

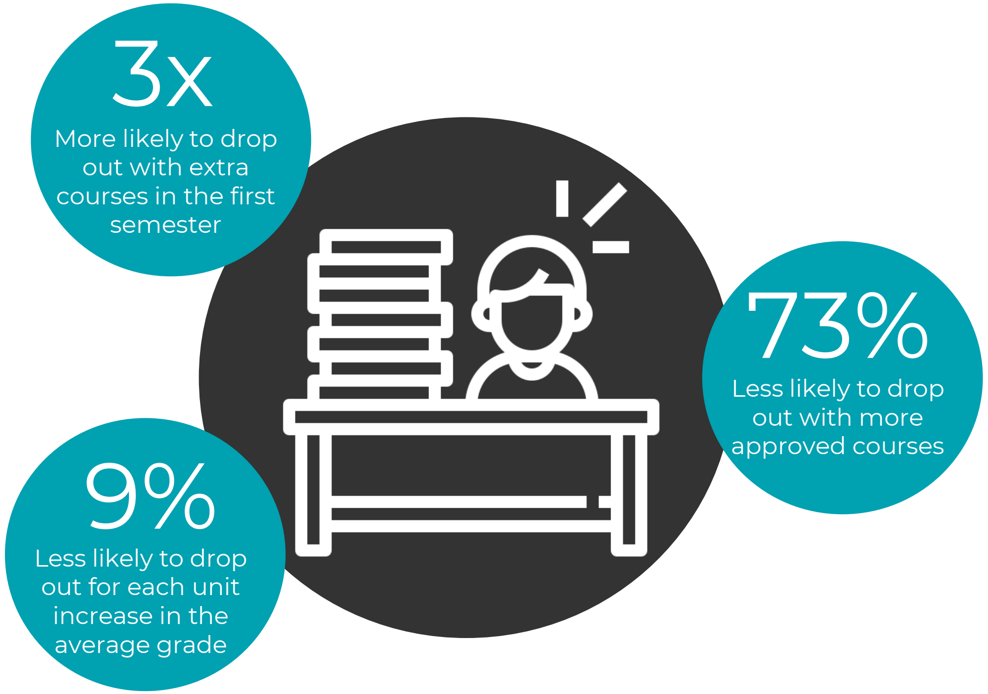
Features including **scholarship holder, debtor, age at enrollment** remained relevant predictors in the extended model

Even with new first semester grade features, some key features like scholarships and debtor remained important predictors.

*We are assuming "approved" refers to a course that has been passed, while "credited" refers to a course that the student has not only passed but has earned the corresponding academic credits towards their degree.

Model 2 – After 1st Semester

Findings from Analysis – After 1st Semester



Overloading on courses associated with greater dropout risk, while successfully completing more courses is associated with a greater probability of graduating. Having a greater average from first semester slightly reduces risk of dropping out.

Academic Performance in the First Semester is highly influential in predicting dropout and graduation rates.

Model 2 – After 1st Semester

Findings from Analysis – After 1st Semester

49%

Less likely to drop out if a scholarship holder

91%

More likely to drop out if in debt

Financial status in terms of debt and scholarship are still impactful on likelihood of drop out, just as they were in the admission model.

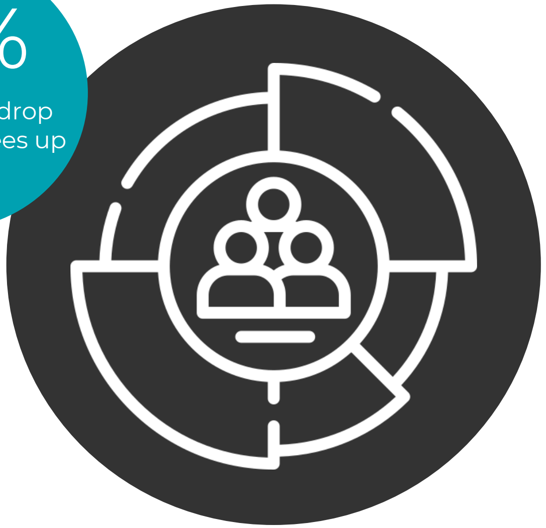
Certain socioeconomic indicators are still critical

Model 2 – After 1st Semester

Findings from Analysis – After 1st Semester

61%

Less likely to drop out if tuition fees up to date



Financial status in terms of adherence to tuition payment schedules is impactful on likelihood of dropping out.

Introducing a new socioeconomic factor proves to be relevant

Model 2 – After 1st Semester

Findings from Analysis – After 1st Semester

11%
Less likely to drop out with higher inflation rates

7%
More likely to drop out with higher unemployment rates

4%
More likely to drop out for older students at time of enrollment

Macroeconomic indicators rise in importance and age at enrollment yields interesting results

Macroeconomic and demographic factors show slight relevance in the model

Model 2 – After 1st Semester

Implications & Interventions – After 1st Semester

Manage course load

Provide guidance on managing their course load effectively via academic advisors

Academic success programs

Implement programs to improve the number of approved units in the first semester

Academic performance monitoring

Regular and balanced approach to providing feedback through evaluations

Targeted support to promote the success of current students most at risk of not graduating

Executive Summary

MANDATE	KEY ELEMENTS TO ASSESS	RECOMMENDATIONS	IMPACTS
Present recommendations to evaluate key factors affecting university graduation and dropout rates, aligning with Portugal University Board's (PUB) goal of boosting higher education graduates	What are the **main predictors** related to graduation? What is key to consider given the **public context** of this initiative? Ensure **ethical practices** while harnessing data what steps can be taken to initiate **actionable measures**?	**The DATA Methodology** (**D**issect-**A**ssemble-**T**ailor-**A**ctivate) A methodology **tying up** the findings of our work to **tangible actions** driving results for the **Portugal University Board**	Offer a **scalable** and reusable **tool** applicable **nationwide.** **An insightful framework** for guiding future initiatives **to enhance PUB's graduation rates.**

Appendix

Why Does Prioritizing Student Success Matter?

Tackling a Domestic Issue

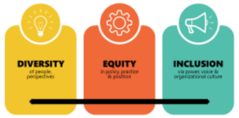
Improving domestic human capital is important for Portugal's sustained economic development – providing a key ingredient for improvements in productivity, innovation, and competitiveness of the economy

Control Financial Impacts

Sustaining their financial viability as enrolled students affect direct revenues from tuition, and indirect revenues from public funding allocations and/or external donors which may depend on metrics such as enrollment and graduation rates.

Maintaining Reputation

Improving external reputation, as dropout rates can be seen as reflective satisfaction in addition to performance of students and teachers.



Providing equal chances

Enhancing their social responsibility and accountability, as dropout can affect the equity and diversity of the student body, as well as the contribution of institutions to the development of their surrounding communities

Adaptation and validation to EuropeanPortuguese of the Reasons for Higher Education Dropout Scale

Portugal | Education at a Glance 2022 : OECD Indicators

Why supporting student success matters

Why it matters for Portugal

- Improving domestic human capital is important for Portugal’s sustained economic development – providing a key ingredient for improvements in productivity, innovation, and competitiveness of the economy

Why it matters for specific institutions

- Improving external reputation, as dropout rates can be seen as reflective satisfaction in addition to performance of students and teachers.
- Sustaining their financial viability as enrolled students affect direct revenues from tuition, and indirect revenues from public funding allocations and/or external donors which may depend on metrics such as enrollment and graduation rates.
- Enhancing their social responsibility and accountability, as dropout can affect the equity and diversity of the student body, as well as the contribution of institutions to the development of their surrounding communities

At Admission Model Results (1: Dropout, 0: Graduate)

Feature	Variable Type	Feature Meaning	Feature Category	Coefficient	Odds Ratio
Debtor_1	Categorical	Is a debtor	Socioeconomic data	1.79183834	6.000473241
Scholarship holder_1	Categorical	Is a scholarship holder	Socioeconomic data	-1.517730421	0.219208835
Course_Social Service (evening attendance)	Categorical	Enrolled to study Social Services - evening classes	Course	-1.217860027	0.295862628
Mother's qualification_Unknown	Categorical	Mother's highest educational attainment is unknown	Socioeconomic data	1.185700254	3.272977937
Course_Basic Education	Categorical	Enrolled to study Basic Education	Course	1.060293342	2.887217807
Gender_1	Categorical	Male	Demographic data	0.754421559	2.126381181
Course_Nursing	Categorical	Enrolled to study Nursing	Course	-0.617621776	0.539225312
Mother's occupation_Student	Categorical	Mother is currently a student	Socioeconomic data	0.489300065	1.631174105
Age at enrollment	Numerical	Age of student at enrollment	Demographic data	0.057293848	1.05896694

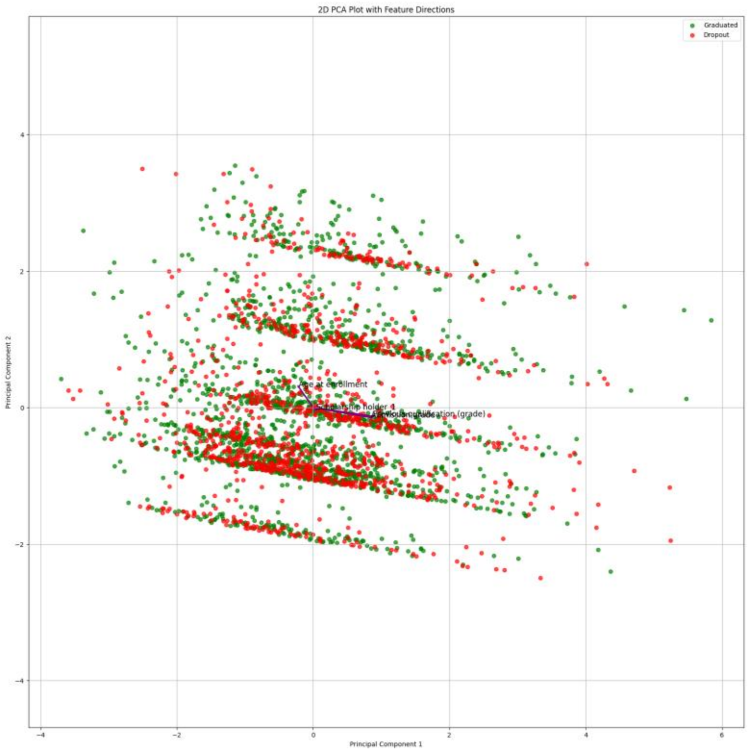
After 1st Semester Model Results (1: Dropout, 0: Graduate)

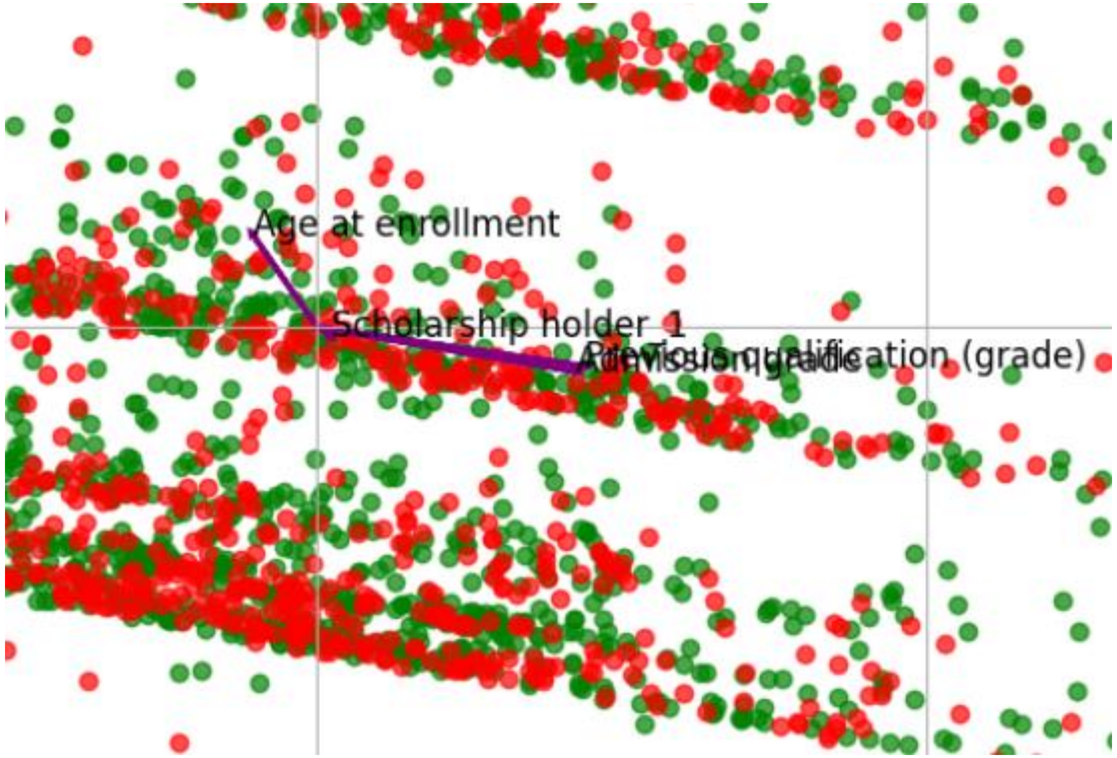
Feature	Variable Type	Feature Meaning	Feature Category	Coefficient	Odds Ratio
Curricular units 1st sem (approved)	Numerical	The number of courses a student has successfully completed in the first semester	Academic data at the end of 1st semester	-1.307420	0.270517
Curricular units 1st sem (enrolled)	Numerical	The number of courses a student has enrolled in during the first semester	Academic data at the end of 1st semester	1.095973	2.992092
Tuition fees up to date_1	Categorical	Has tuition fees up to date	Socioeconomic data	-0.946533	0.388084
Scholarship holder_1	Categorical	Is a scholarship holder	Socioeconomic data	-0.676500	0.508393
Debtor_1	Categorical	Is a debtor	Socioeconomic data	0.648269	1.912227
Inflation rate	Numerical	Inflation rate of the economy of the country where the student is from	Macroeconomic data	-0.106620	0.898867
Curricular units 1st sem (grade)	Numerical	Average grade of a student's first semester courses	Academic data at the end of 1st semester	-0.096624	0.907897
Unemployment rate	Numerical	Unemployment rate of the economy of the country where the student is from	Macroeconomic data	0.072293	1.074970
Age at enrollment	Numerical	Age of student at enrollment	Demographic data	0.039217	1.039996
GDP	Numerical	Gross Domestic Product of the country where the student is from	Macroeconomic data	-0.014571	0.985535
Admission grade	Numerical	The grade achieved by the student during the admission process	Academic data at enrollment	-0.007318	0.992709
Curricular units 1st sem (evaluations)	Numerical	Number of evaluations a student has undergone in the first semester	Academic data at the end of 1st semester	0.005328	1.005342
Previous qualification (grade)	Numerical	Grade of student's previous qualification before enrolling in the institution	Academic data at enrollment	0.003227	1.003232

Variable interpretation PCA

- PCA with top 4 variables from random forest (first semester)

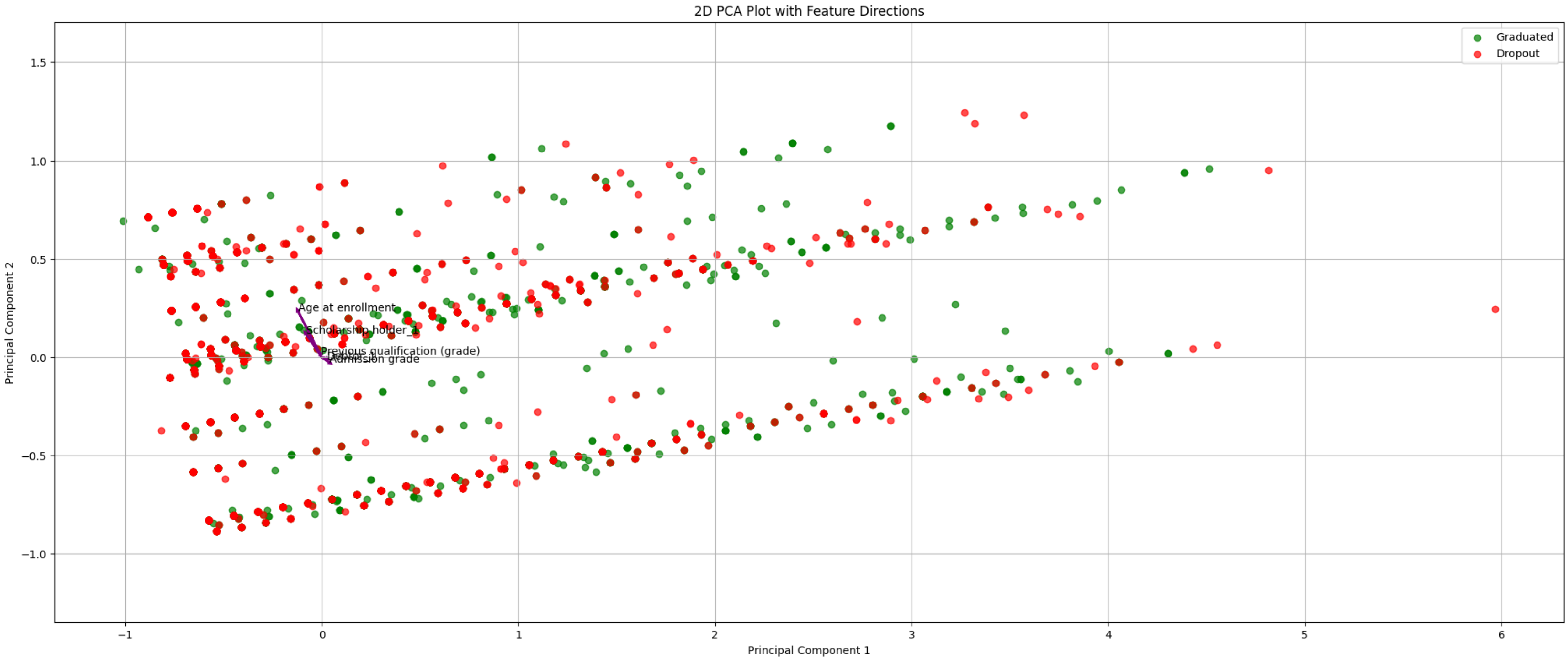
Variable interpretation PCA





- PCA with top 4 variables from random forest (entrance)

Variable interpretation PCA



- PCA with top 4 variables from boosting (entrance)

Findings from Analysis – At Admission

Socioeconomic indicators are the strongest predictors for likelihood of dropping out

- Being a debtor significantly increases the odds of a student dropping out. The odds ratio indicates that debtors are about 6 times more likely to drop out compared to non-debtors.
- Looking at the predictor with the next highest odds ratio, we see that Students whose mother's highest educational attainment is unknown (which collection methods indicate likely means a low level of educational attainment) are over 3 times more likely to drop out compared to student's whose mothers have higher levels of educational attainment
- In a similar vein, students whose mothers are also currently students are 1.6 times more likely to drop out than students whose mothers have other occupations.
- The predictor with the strongest upward effect on the likelihood of a student graduating is also a socioeconomic one – whether they are a scholarship holder. This is consistent with the expectation that scholarship holders would have higher academic performance and commitment to their studies than non-scholarship holders.

Nursing & Social Services are the courses with the highest likelihood of producing graduates

- Students enrolled in the evening Social Service program and the Nursing program are 70% and 47% less likely to drop out, respectively, than those enrolled in other programs

There is varying impact from demographic predictors

- Being male is associated with a higher likelihood of dropping out. The odds ratio indicates that male students have about twice the odds of dropping out compared to female students.
- An increase in the age of a student at the time of enrollment slightly increases the likelihood of dropping out. The effect is modest but indicates that older students face slightly higher dropout risks.

Implications & Interventions – At Admission (1/2)

Enhancing community support structures

- The significant influence of a mother's education and occupation on graduation likelihood underscores the potential lack of community support for some students, particularly in areas such as developing study habits, accessing tutoring, and generally navigating higher education
- To address this, expanding and promoting support programs that impart these skills is crucial. Implementing a peer support system, which emulates a community support network, may prove more effective than individualized approaches like webinars. This could include incentivizing academically successful scholarship students with stipends to participate in such peer support programs, fostering a more engaging and supportive learning environment.

Subsidized support for at-risk students

- Given that students in debt are among the most vulnerable, it is important that support services provided to them are subsidized wherever possible. In scenarios where budget constraints limit widespread subsidization, adopting a needs-based approach to allocate these services could ensure that resources are directed to those who need them most, maximizing the impact of available funding.

Learning from successful programs

- The higher graduation rates in nursing and social services programs in the country merit a detailed qualitative case study to uncover pedagogical differences compared to other programs. Insights gained from such a study about effective teaching techniques and student engagement strategies could then be piloted in other programs, aiming to replicate these successes

Exploring gender disparities

- The gender-based findings warrant a deeper investigation. In past studies where this phenomenon was observed – it was found that have males were more often enrolled in more challenging programs, potentially making gender a proxy indicator for enrollment in high-intensity courses. Should a genuine gender disparity be identified, targeted interventions for male students should be developed to address this imbalance.

Implications & Interventions – At Admission (1/2)

Strategic planning and budget allocation

- This model offers a valuable tool for estimating the dropout risk profile of incoming student cohorts, enabling more informed planning and budget allocation for support programs and services. The ability to transform generally-held principles about student success into customized predictions for the risk-profile of specific institutions will be instrumental in tailoring institutional resources to meet anticipated needs effectively.

Ethical Considerations and Fairness Constraints

- It is vital to ensure that the findings of this model are not used to discriminate against students from disadvantaged socioeconomic backgrounds, as this would perpetuate existing inequalities and contravene the broader societal role of higher education institutions in fostering economic development. To mitigate this risk, the model could be further refined by incorporating group-level fairness constraints.

Findings from Analysis – First Semester (T1)

Strongest Predictors for Dropping Out:

- Academic Performance in the First Semester: A student taking additional curricular units in the first semester faces approximately three times higher odds of dropping out compared to their counterparts.
- Socioeconomic Indicators: Individuals with outstanding debts have about 91% higher odds of dropping out compared to those without debts.

Strongest Predictors for Graduating:

- Academic Performance in the First Semester: Each additional approved (passed) unit in the first semester reduces the odds of dropping out by approximately 73%, suggesting a positive impact on graduation.
- Socioeconomic Indicators: Being a scholarship holder decreases the odds of dropping out by about 49%, highlighting the positive impact of financial support on student retention. Similarly, students who have their tuition fees up to date are 61% less likely to drop out. This could imply that financial stability plays a role in student retention.

Other Predictors:

- Macroeconomic: Higher inflation rates decreases the likelihood of dropping out by 11%. This could suggest that in times of higher inflation, students are slightly less likely to drop out. Higher unemployment rates increase the likelihood of dropping out by 7%. This could suggest that in times of higher unemployment, students are slightly more likely to drop out. Higher GDP slightly decreases the likelihood of dropping out by 2%. This could suggest that in countries with higher GDP, students are slightly less likely to drop out (these are macroeconomic indicators are of Portugal’s economy at the given time).
- Demographic: Older students at the time of enrollment are about 4% more likely to drop out, indicating unique challenges and considerations for this demographic.
- Academic performance in the first semester: For each unit increase in the average grade, the odds of dropping out decrease by approximately 9%, suggesting that higher grades reduce the risk of dropout. For each additional evaluation, the odds of dropping out increase by approximately 1%, indicating a subtle impact of evaluations on dropout likelihood.
- Academic performance before enrolling: Higher admission grades reduce the odds of dropping out by approximately 1%, underscoring the importance of prior academic achievement. For each unit increase in the grade of the previous qualification, the odds of dropping out increase by approximately 1%, indicating a subtle influence of past academic performance on current outcomes.

Interpreting the Coefficients and Odds Ratios

The logistic regression output presents a set of coefficients along with their respective odds ratios, offering insights into the factors influencing student dropout rates. The coefficients indicate the change in the log odds of the outcome for a one-unit change in the predictor variable, holding other variables constant.

For instance, a coefficient of -1.307420 for 'Curricular units 1st sem (approved)' suggests a negative relationship with the likelihood of dropping out; for each additional approved unit, the log odds of dropping out decrease by this value. The corresponding odds ratio of 0.270517 further translates this log odds change into multiplicative odds of the outcome, indicating that with each additional approved unit, the odds of dropping out are about 73% lower (1 - 0.270517).

Conversely, a positive coefficient, such as 1.095973 for 'Curricular units 1st sem (enrolled)', implies an increased likelihood of dropping out with additional enrolled units. The odds ratio of nearly 3 (2.992092) suggests that students enrolling in more units have approximately three times the odds of dropping out compared to those enrolling in fewer units.

Implications & Interventions – First Semester

Many of the interventions explained in the admission model are still relevant here as we see many socioeconomic factors are quite relevant in this extended model. However, as it is relevant that how a student performs in their first semester is critical in determining whether or not they will drop out, we can introduce some new interventions that the school can implement during the first semester.

Given that students who enroll in more curricular units in the first semester are more likely to drop out, institutions could provide guidance to students on managing their course load effectively. This could include academic advising sessions where advisors help students select an appropriate number of courses. Additionally, the school can implement programs that aim to improve the number of approved (passed) units in the first semester such as tutoring programs, study groups, or workshops for study skills and time management. Regular evaluations and feedback can help students understand their academic standing and areas of improvement. It is important to note that just implementing these programs is not enough, the school needs to advertise these and encourage students to use them so that the students are able to experience their benefits.