Predicting Successful Graduation from Performance in Mathematics Courses

Gonzalo Cazes
Caleb Hughes
David Larue
Kristopher Lee
Arvind Suresh

Erdős Institute
Data Science Boot Camp
Fall, 2024

Student Success Prediction from Performance in Mathematics Classes

Our Data and the Question: Mathematics courses taken and outcome predicts successful graduation.

Our reduced data set: No individual grades, just

- Above C-
- Satisfactory/Pass
- C- or Below

What can be predicted now?

Data Cleaning & Pre-Processing

- Iowa Stats: List of all math courses taken with a grade of C- or above. We know if a student got below a C or not, if/when they graduated, when they took the course and when they enrolled in the university

Data Cleaning & Pre-Processing

- Iowa Stats: List of all math courses taken with a grade of C- or above. We know if a student got below a C or not, if/when they graduated, when they took the course and when they enrolled in the university
- Focused on undergraduates only

Data Cleaning & Pre-Processing

- Iowa Stats: List of all math courses taken with a grade of C- or above. We know if a student got below a C or not, if/when they graduated, when they took the course and when they enrolled in the university
- Focused on undergraduates only
- Removed courses that wouldn't contribute (no longer offered, unique experiences, low enrollment)

Data Cleaning & Pre-Processing

- Iowa Stats: List of all math courses taken with a grade of C- or above. We know if a student got below a C or not, if/when they graduated, when they took the course and when they enrolled in the university
- Focused on undergraduates only
- Removed courses that wouldn't contribute (no longer offered, unique experiences, low enrollment)
- Spring/Fall were one semester, Summer was 0.5. We looked at students enrolled from 2008 until before 2021.

Data Cleaning & Pre-Processing

- Iowa Stats: List of all math courses taken with a grade of C- or above. We know if a student got below a C or not, if/when they graduated, when they took the course and when they enrolled in the university
- Focused on undergraduates only
- Removed courses that wouldn't contribute (no longer offered, unique experiences, low enrollment)
- Spring/Fall were one semester, Summer was 0.5. We looked at students enrolled from 2008 until before 2021.
- Added a column for the target variable Y, indicating if the student graduated within 9.5 semesters

Data Cleaning & Pre-Processing

- Iowa Stats: List of all math courses taken with a grade of C- or above. We know if a student got below a C or not, if/when they graduated, when they took the course and when they enrolled in the university
- Focused on undergraduates only
- Removed courses that wouldn't contribute (no longer offered, unique experiences, low enrollment)
- Spring/Fall were one semester, Summer was 0.5. We looked at students enrolled from 2008 until before 2021.
- Added a column for the target variable Y, indicating if the student graduated within 9.5 semesters
- Created an analog of GPA

EDA

	Cleaned Dataset	Original Dataset
Total Students	9,181	13,065
Total Courses	28	110
Graduation Rate	49%	58%

EDA

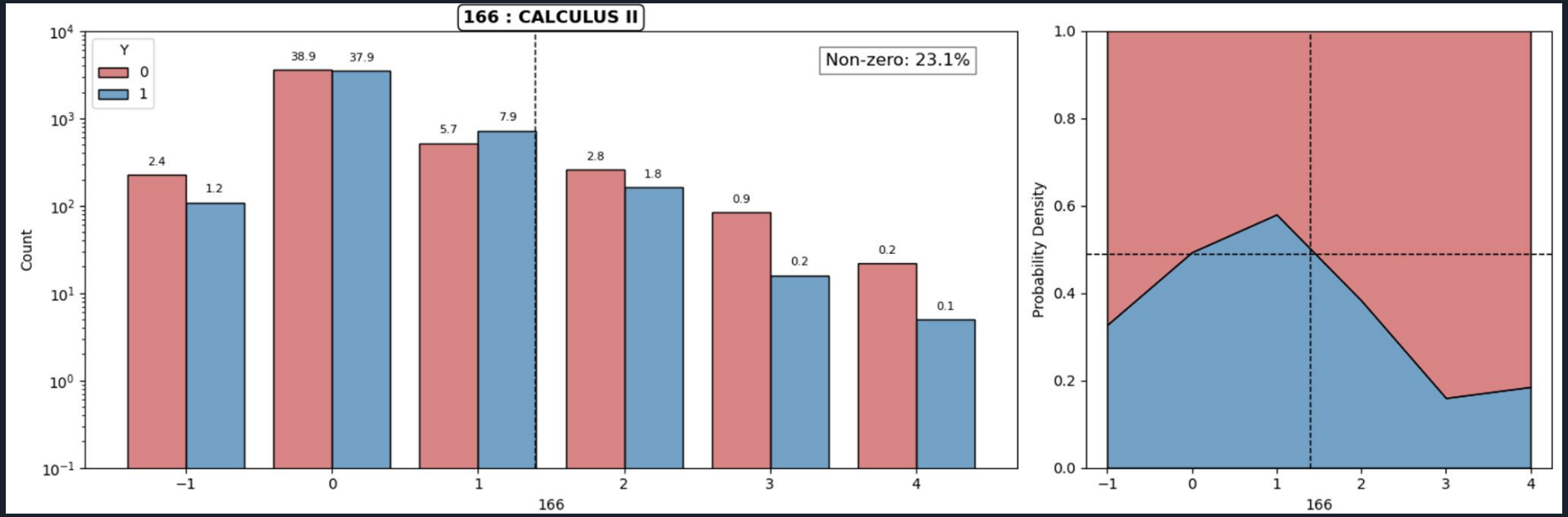
	Math Major Courses	General Courses Only	Full Dataset
Total Students	1,245	7,936	9,181
Total Courses	16	12	28
Graduation Rate	61%	47%	49%

EDA

	Math Major Courses	General Courses Only	Full Dataset
Total Students	1,245	7,936	9,181
Total Courses	16	12	28
Graduation Rate	61%	47%	49%

CONCLUSION: Taking at least 1 math major course may improve odds of graduating!

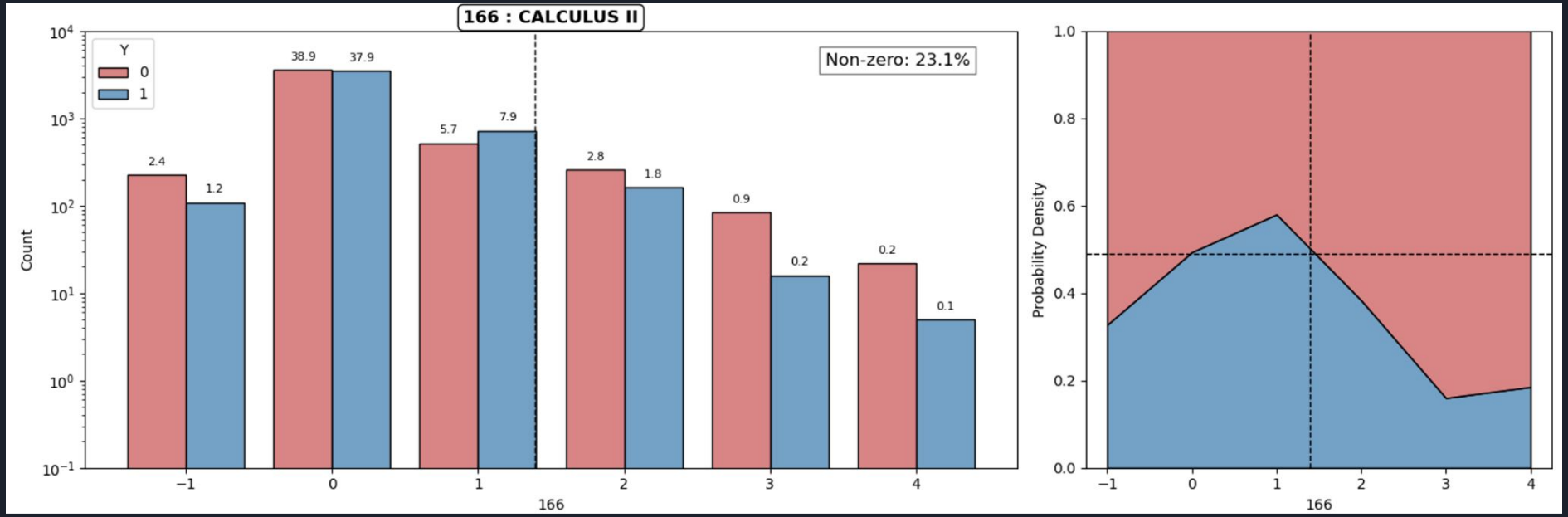
EDA



Weighted performance: 0 if student did not take the course -1 if student took the course and received less than a C

1-4 if student received a C or above in year 1-4, respectively

EDA



CONCLUSION: the earlier a math major course can be taken, the more likely the student is to succeed

Feature engineering

- Well-established: Completing more math in high school is correlated with academic success.
- Question: Is taking more math courses in college correlated with graduating on time?

Feature engineering

- Well-established: Completing more math in high school is correlated with academic success.
- Question: Is taking more math courses in college correlated with graduating on time?
- Measuring “math performance” in college to account for passing time:
  - Performance in semester n:

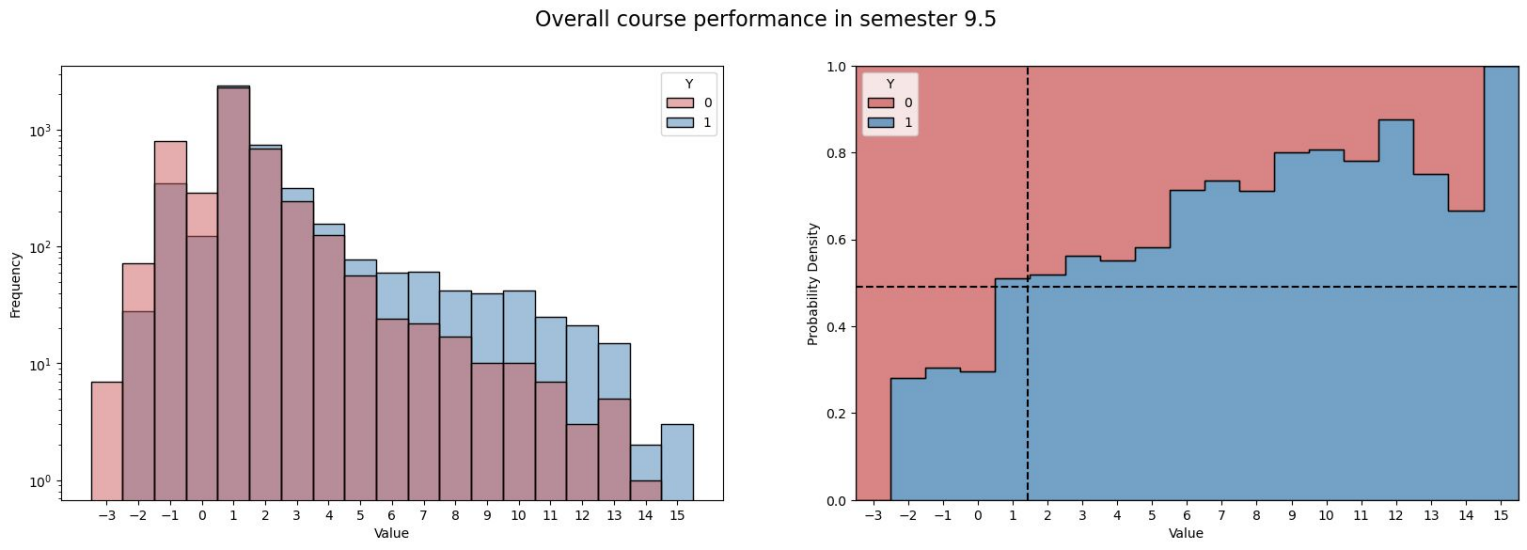
    SEM_n_cdf = #(math courses passed in sem <=n) - #(math courses failed in sem <=n)
  - E.g. Alice passes 2 courses in 1st sem, fails one in 2nd sem, and takes no more math. Then

    SEM_1_cdf = 2,

    SEM_n_cdf = 1, for all n>1

- Expectation: Greater math performance implies greater odds of graduating in 4 years.

Feature engineering



Uncanny connection:

- Performance < Mean performance
- Performance = Mean performance
- Performance > Mean performance

Probability of grad < Mean prob of grad
Probability of grad = Mean prob of grad
Probability of grad > Mean prob of grad

Modeling Approach

- Graduation is a binary target

Modeling Approach

- Graduation is a binary target
- Multiple binary selection models; focused efforts on four:

Modeling Approach

- Graduation is a binary target

- Multiple binary selection models; focused efforts on four:
  - Logistic Regression (baseline)

Modeling Approach

- Graduation is a binary target

- Multiple binary selection models; focused efforts on four:
  - Logistic Regression (baseline)
  - Support Vector Classifier with Radial Basis Function

Modeling Approach

- Graduation is a binary target

- Multiple binary selection models; focused efforts on four:
  - Logistic Regression (baseline)
  - Support Vector Classifier with Radial Basis Function
  - XGBoost Classifier

Modeling Approach

- Graduation is a binary target

- Multiple binary selection models; focused efforts on four:
  - Logistic Regression (baseline)
  - Support Vector Classifier with Radial Basis Function
  - XGBoost Classifier
  - Custom Stacked Classifier

Modeling Results

Model	CV Accuracy	Test Accuracy	Accuracy Change
Logistic	63.08	62.44	-1.01
SVC	64.86	64.53	-0.5
XGBoost	66.5	65.7	-1.19
Stacked	69.47s	65.78s	-5.27

Feature Importance

- Logistic models preferred cumulative performance
  - Top 16 features by absolute value

Feature Importance

- Logistic models preferred cumulative performance
  - Top 16 features by absolute value

- XGBoost preferred individual courses
  - Top 3 courses: Introduction to Probability, Preparation for Calculus, and Calculus II

So... how did we do?

- Best test accuracy we got was around 65%.
- Claim: Subject to the inherent limitations of out dataset, this is not bad!
- Reason: Our dataset suffers from a high ***Bayes Error Rate***:

So... how did we do?

- Best test accuracy we got was around 65%.
- Claim: Subject to the inherent limitations of out dataset, this is not bad!
- Reason: Our dataset suffers from a high ***Bayes Error Rate***:
  - Theoretical minimum error for any binary classifier for the dataset.
  - Unavoidable prediction error comes from groups of students
    - with *completely identical features* (so models can't distinguish them),
    - but evenly distributed into both grads as well as non-grads!
  - 9181 students —-> Only 2580 distinct rows of features!

So... how did we do?

- Best test accuracy we got was around 65%.
- Claim: Subject to the inherent limitations of out dataset, this is not bad!
- Reason: Our dataset suffers from a high ***Bayes Error Rate***:
  - Theoretical minimum error for any binary classifier for the dataset.
  - Unavoidable prediction error comes from groups of students
    - with *completely identical features* (so models can't distinguish them),
    - but evenly distributed into both grads as well as non-grads!
  - 9181 students —-> Only 2580 distinct rows of features!
  - Computed a (highly non-sharp) lower bound for Bayes Error Rate = **25.3**%!
  - Equivalently, upper bound on accuracy = **74.7**%!

- Conclusion: *The predictive power of our dataset was limited from the start!*

Future directions

- Despite the limited nature and sparsity of our dataset, we found some interesting correlations between math courses and successful graduation.

- Expect: with sufficiently large and robust data, can predict graduation with much higher probability.
  (*With great data comes great predictability!*)

- In particular, we believe that the math performance indicators can be further refined to increase predictive power.

- Long-term application: help departments develop tools to identify and support students at risk of dropping out.

Acknowledgements

- We thank the Iowa State University's math department for generously sharing their data with us.
- We thank our project mentor Alec Traaseth for his invaluable advice that helped us stay on track throughout this project.
- We thank the Erdos Institute for organizing this great bootcamp, and in particular, Steven Gubkin for providing great, informative lectures.