12.04.25

Discovering Enrollment Patterns for Course Planning

Tasmia Alamgir Zubaer Chowdhury Mateo Goldman

1. Introduction & Motivation

2. Cleaning & Preprocessing

3. Feature Engineering

4. Modeling

5. Conclusion

Motivation

College students notoriously face significant stress when planning their academic trajectory, often leading to inefficiencies in course selection.

> 30% of U.S. college students change their major within three years.

> Only 46% of students graduate with a bachelor's degree within four years.

> High Stress Registration uncertainty creates significant anxiety and burdens academic advisors.

Project Goal

Offer better support to student, faculty and staff by making enrollment and academic planning for efficient for schools, colleges and universities.

Advisors

Automating the discovery of course popularity and seat availability

Reducing repetitive query burden on academic advisors, allowing them to focus on guidance pivotal to student success.

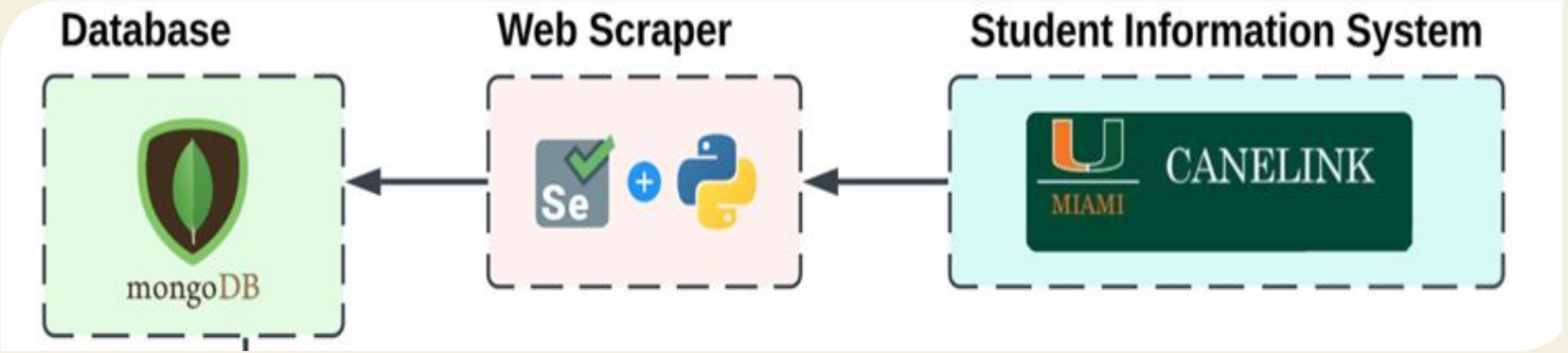
Students

Provide students with transparent metrics:

- Course Ratings (Fill Speed)
- Dropout Rates (Professor Quality)
- Enrollment Probability

Initial Database and Scrapping

Course information data from University of Miami's Student Information System was scrapped, capturing enrollment changes in real time for multiple semesters.
We only analysed the exported dataset in CSV format from MongoDB time-series data cluster.



1. Introduction & Motivation

2. Cleaning & Preprocessing

3. Feature Engineering

4. Modeling

5. Conclusion

Dataset Overview

We imported the static course sections file, which contains one row per class with metadata such as subject, catalog number, capacity, and delivery mode and the time-series file, which tracks how the enrollment for each course changes over time as open seats are taken.

Initial State:

sections.csv

_id	name	subjectName	subjectCode	catalogNumber	academicCareer	semester	year	sectionType	sectionCode	classNumber	session	days[0]	days[1]	days[2]
67c0	Principles	Accounting Bus	ACC	211	Undergraduate	Spring	2025	Lecture	1U	8429	Regular Acad	Tuesday		
67c0	Principles	Accounting Bus	ACC	211	Undergraduate	Spring	2025	Lecture	A	8431	Regular Acad	Monday	Wednesday	Friday
67c0	Principles	Accounting Bus	ACC	211	Undergraduate	Spring	2025	Lecture	B	8432	Regular Acad	Monday	Wednesday	Friday
67c0	Principles	Accounting Bus	ACC	211	Undergraduate	Spring	2025	Lecture	C	8433	Regular Acad	Monday	Wednesday	Friday
67c0	Principles	Accounting Bus	ACC	211	Undergraduate	Spring	2025	Lecture	E	8434	Regular Acad	Monday	Wednesday	Friday
67c0	Principles	Accounting Bus	ACC	211	Undergraduate	Spring	2025	Lecture	F	8435	Regular Acad	Monday	Wednesday	Friday

courses.sectionsTS.csv:

dateTimeRetrieved	courseInfo.classNumber	courseInfo.semester	courseInfo.year	_id	seatsAvailable	status	waitlistAvailable	reservedSeat	wailistAvailable
2024-11-04T00:14:5	8429	Spring	2025	67c083	45	Open	300		
2024-11-04T19:30:5	8429	Spring	2025	67c083	45	Open	300		
2024-11-05T12:39:4	8429	Spring	2025	67c083	45	Open	300		
2024-11-06T10:04:3	8429	Spring	2025	67c083	45	Open	300		
2024-11-06T17:14:3	8429	Spring	2025	67c083	45	Open	300		
2024-11-07T14:56:0	8429	Spring	2025	67c083	44	Open	300		

Final State:

We removed empty, irrelevant, or low-quality columns/rows from both datasets and save the cleaned versions for downstream analysis.

Section CSV size reduced from 13.5MB to 3.5MB and time series CSV by 3MB

sections-trimmed.csv:

_id	name	subjectName	subjectCode	catalogNumber	academicCareer	semester	year	sectionType	sectionCode	classNumber	session	days[0]	days[1]	days[2]
67c08	Principles	Accounting Bus	ACC	211	Undergraduate	Spring	2025	Lecture	1U	8429	Regular	Tuesday		
67c08	Principles	Accounting Bus	ACC	211	Undergraduate	Spring	2025	Lecture	A	8431	Regular	Monday	Wednesday	Friday
67c08	Principles	Accounting Bus	ACC	211	Undergraduate	Spring	2025	Lecture	B	8432	Regular	Monday	Wednesday	Friday
67c08	Principles	Accounting Bus	ACC	211	Undergraduate	Spring	2025	Lecture	C	8433	Regular	Monday	Wednesday	Friday
67c08	Principles	Accounting Bus	ACC	211	Undergraduate	Spring	2025	Lecture	E	8434	Regular	Monday	Wednesday	Friday
67c08	Principles	Accounting Bus	ACC	211	Undergraduate	Spring	2025	Lecture	F	8435	Regular	Monday	Wednesday	Friday

courses.sections-trimmedTS.csv

dateTimeRetrieved	courseInfo.classNumber	courseInfo.semester	courseInfo.year	_id	seatsAvailable	status	waitlistAvailable	reservedSeatsAvailable	
2024-11-04 00:14:53	8429	Spring	2025	67c0	45	Open	300		
2024-11-04 19:30:56	8429	Spring	2025	67c0	45	Open	300		
2024-11-05 12:39:48	8429	Spring	2025	67c0	45	Open	300		
2024-11-06 10:04:30	8429	Spring	2025	67c0	45	Open	300		
2024-11-06 17:14:37	8429	Spring	2025	67c0	45	Open	300		
2024-11-07 14:56:00	8429	Spring	2025	67c0	44	Open	300		

Data Validation Report

A data quality report summarizing row/column counts, duplicates, and identify 31 fully empty columns that can be safely ignored.

Data Validation Checks
1. All id values are unique (7513 unique values)
2. academicCareer values are all either 'Graduate', or ‘Undergraduate'
3. year value are all '2025'
4. semester value are all 'Spring'

Data Type Summary	
dtype	n_columns
bool	1
float64	33
int64	4
object	124

n_rows	n_columns	n_duplicates
7513	162	0

Numeric Summary										
	count	mean	std	min	25%	50%	75%	max	missing count	missing % capacity
capacity	7513.0	21.348	37.881	1	5	15	25	990	0	0.00

Cleaning & Cohort Selection

Cohort Filtering

Restricted analysis to Spring 2025 semester

Data Quality Repair

Merging incorrect column names into the correct field
(i.e. **wailistAvailable** -> **waitlistAvailable**)

Temporal Structuring

Ordered data for time-series analysis
- Grouped by **classNumber**
- Sorted by **dateTimeRetrieved**

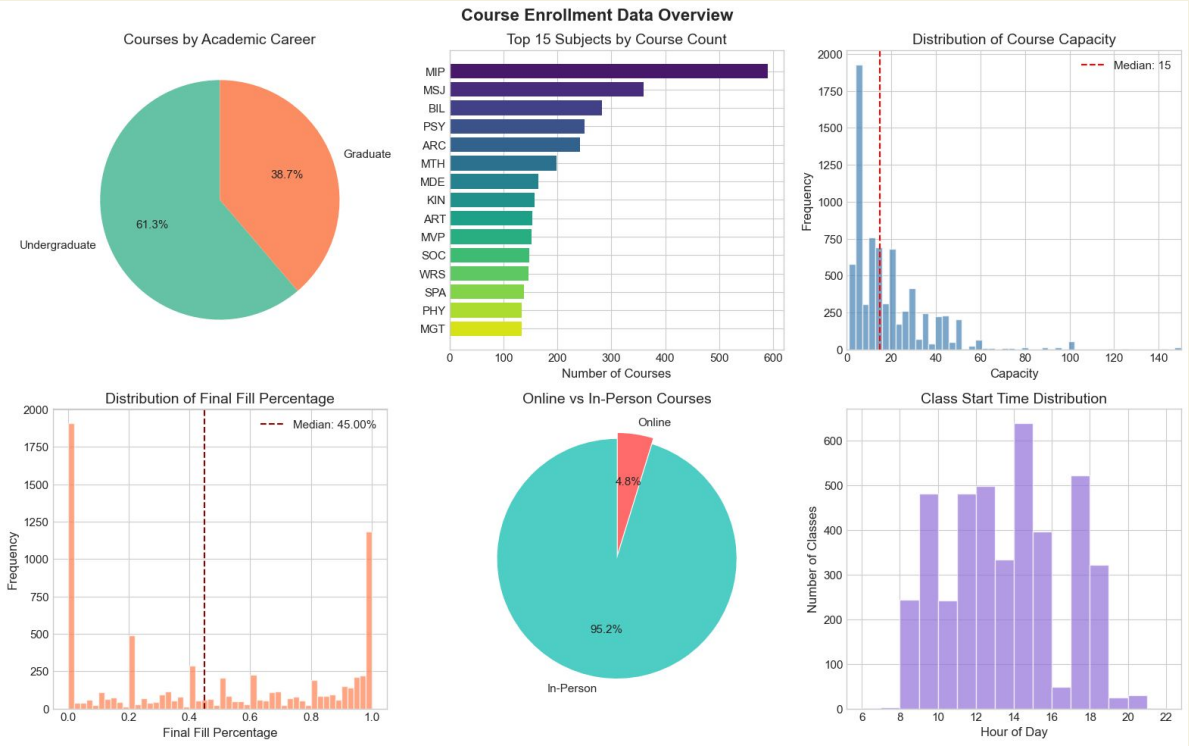
Dimensions

Unique Course Sections: 7,513
Time-Series Observations: 1,060,659

Data Quality and Aggregation

- Reporting Engine to profile time-series observations and static records
- Ensured **dateTimeRetrieved** is *increasing* for every course
- 100% Uniqueness of Course and Section IDs
- Removal of 31 empty columns, reducing dimensionality
  - Unused instructor slots
  - Incorrect calendar fields
- Intersection Merge strategy
  - Keys: **classNumber** + **semester** + **year**
  - Aligning static details with the dynamic logs of enrollment
- Overlapping columns are renamed with a suffix (-TS), preserving the original and new dynamic polling history
- Nested Data Structure construction
  - **seatsAvailable** becomes [45, 45, 44, ...]

Cleaned Dataset Overview



- Noise Reduction
  - Ghost/Cancelled courses removed to prevent model bias
  - 5,612 active courses with 798,136 time-series observations

1. Introduction & Motivation

2. Cleaning & Preprocessing

3. **Feature Engineering**

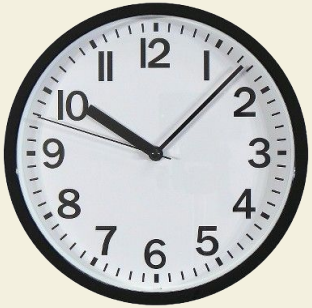
4. Modeling

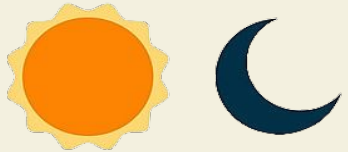
5. Conclusion

Capturing The Temporal Dynamics

Mapped linear time to Sine/Cosine coordinate pairs to ensure model understands that Sun night is adjacent to Monday morning

Velocity, **fill_velocity_per_day**, and Acceleration, **fill_acceleration**, quantifying the surging of registrations



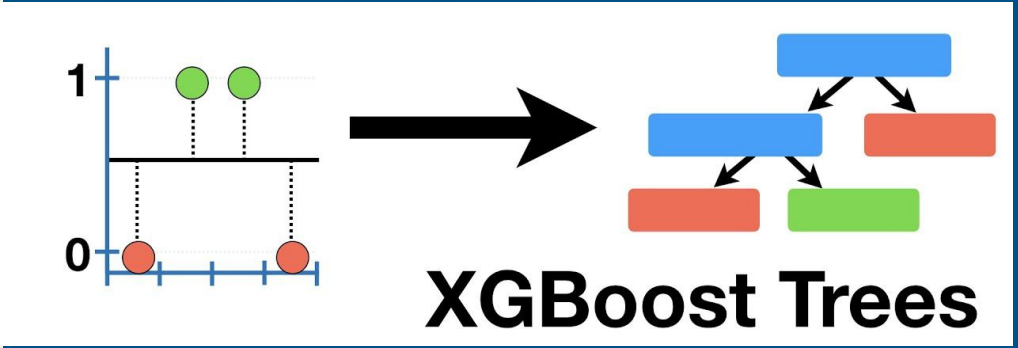


Dataset Construction Modeling

Constructed final training matrix, defining target variable as the **fill percentage**

One-Hot Encoding converts text to **binary vectors**

Final scrubbing to ensure ready for **XGBoost algorithm**



XGBoost

Selected for its ability to handle non-linear fill rates and robustness to missing data.

Leakage Prevention Strategy:

- Excluded the "current state" features (like **Current Seats Taken**). Representing the "answer" at time *t*.
- Included only Lagged Features and Rolling Statistics.

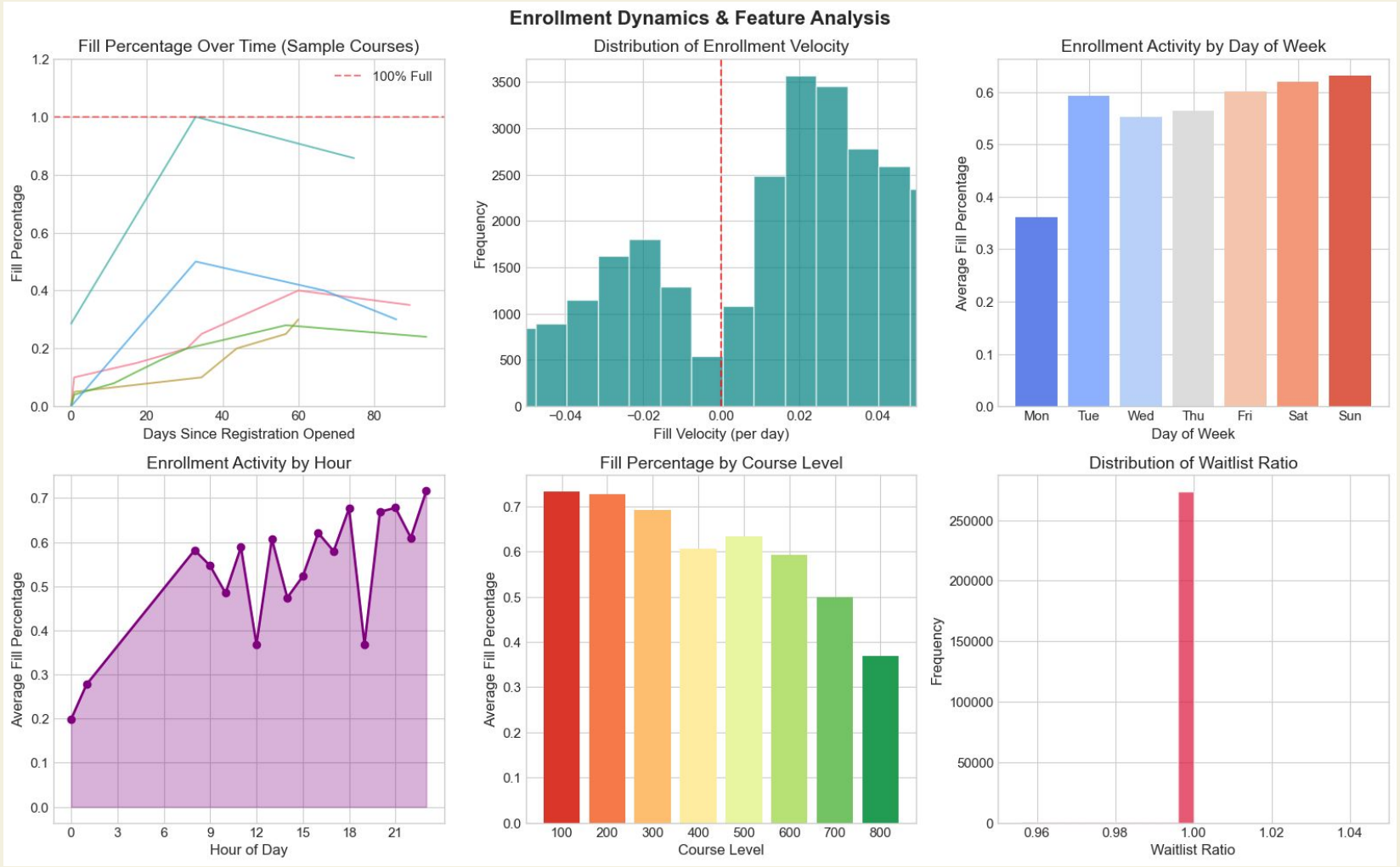
Training Split:

- Training on the first 80% of the timeline, testing on the final 20%.

Mean Absolute Error (MAE) target tolerance of +- 5%.

Feature Engineering

Data Mining



1. Defining the Problem

2. Making Observations

3. Feature Engineering

4. Modeling

5. Conclusion

Forecasting Model

Predicting enrollment 3, 7, and 14 days out

Model only sees past history. NOT current state

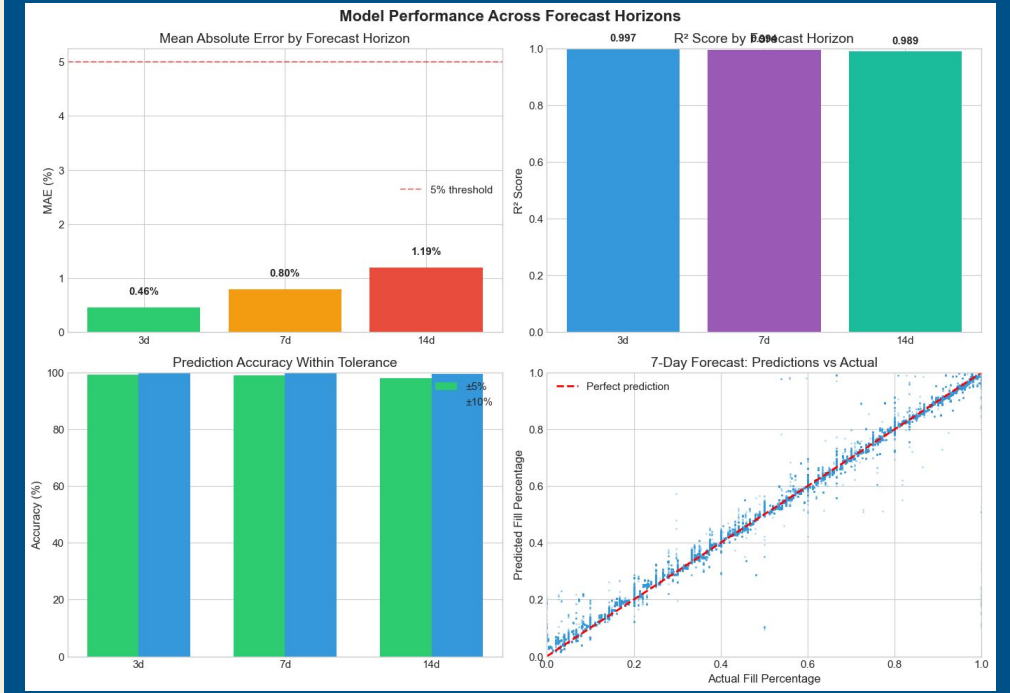
Very high accuracy for short and long term forecasts

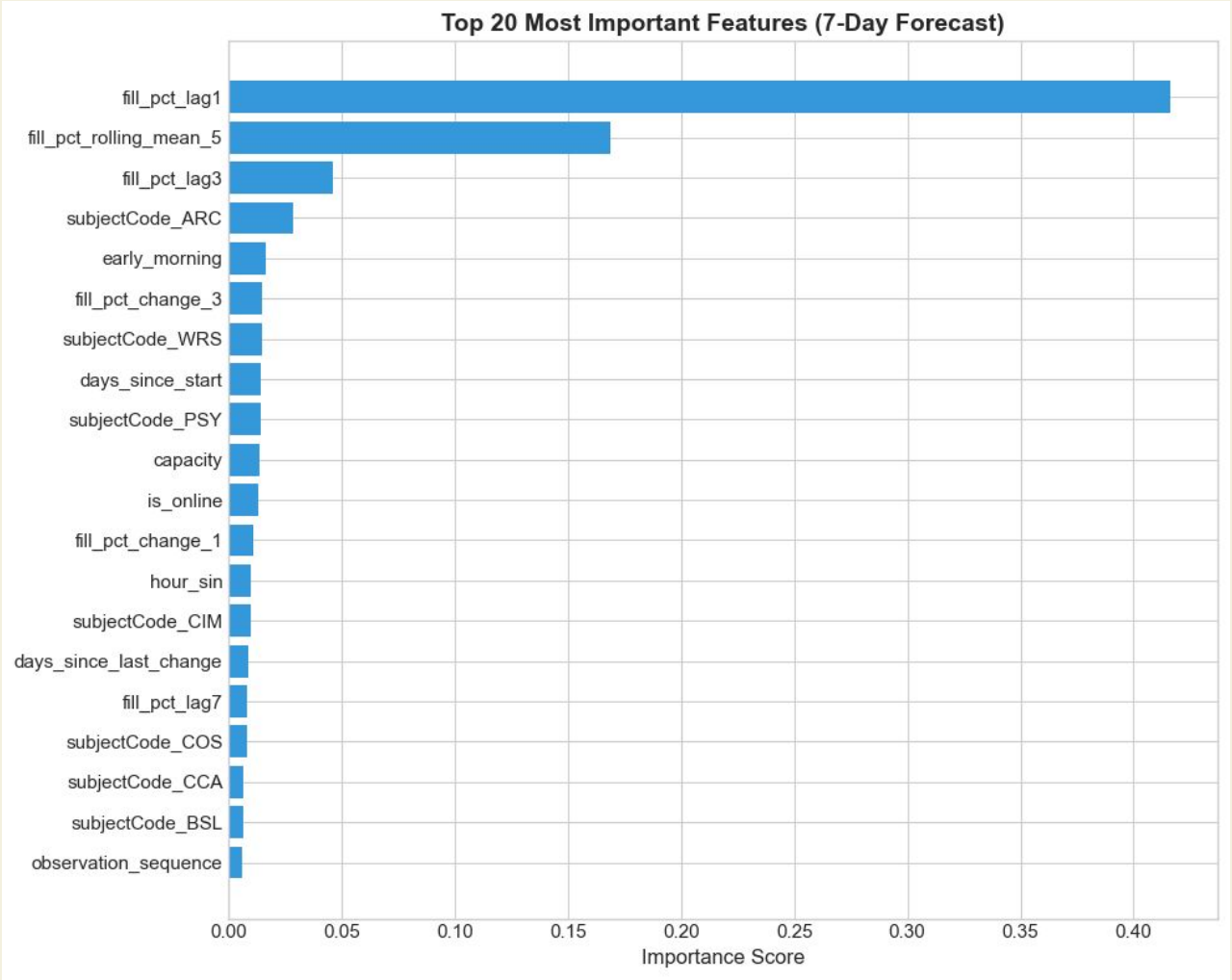
The momentum is the single best predictor!

Time Series Forecasting

XGBoost Classifier + Regressor

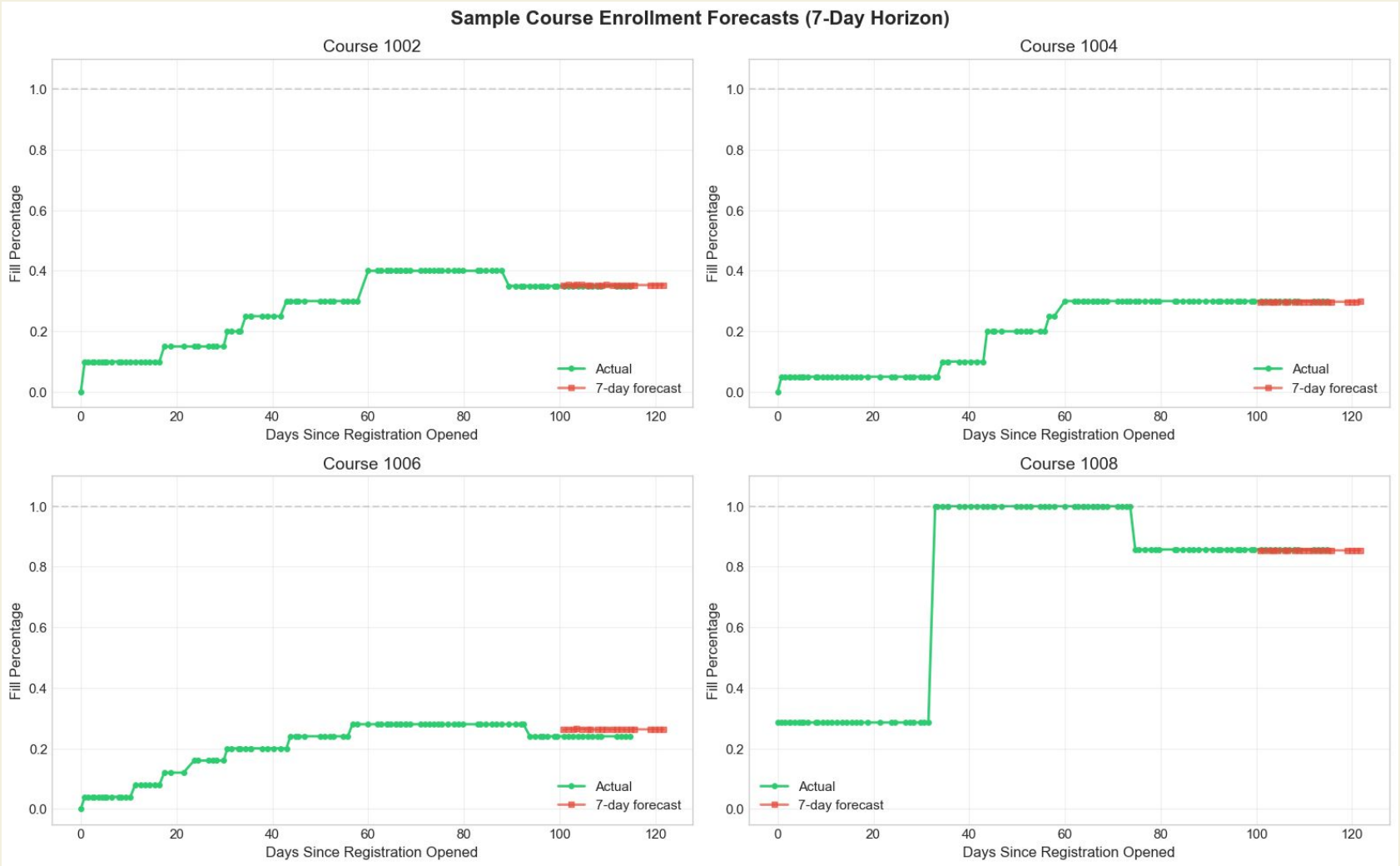
Ranking / Scoring





Architecture (ARC) and Psychology (PSY) courses are popular predictors.

Online courses also seem to contribute well to popularity prediction


Sample Course Enrollment Forecasts (7-Day Horizon)
Course 1002
Course 1004
Course 1006
Course 1008
Fill Percentage
Days Since Registration Opened
Actual
7-day forecast
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120

Course Popularity Classification by Levels

Popularity metrics for ranking courses

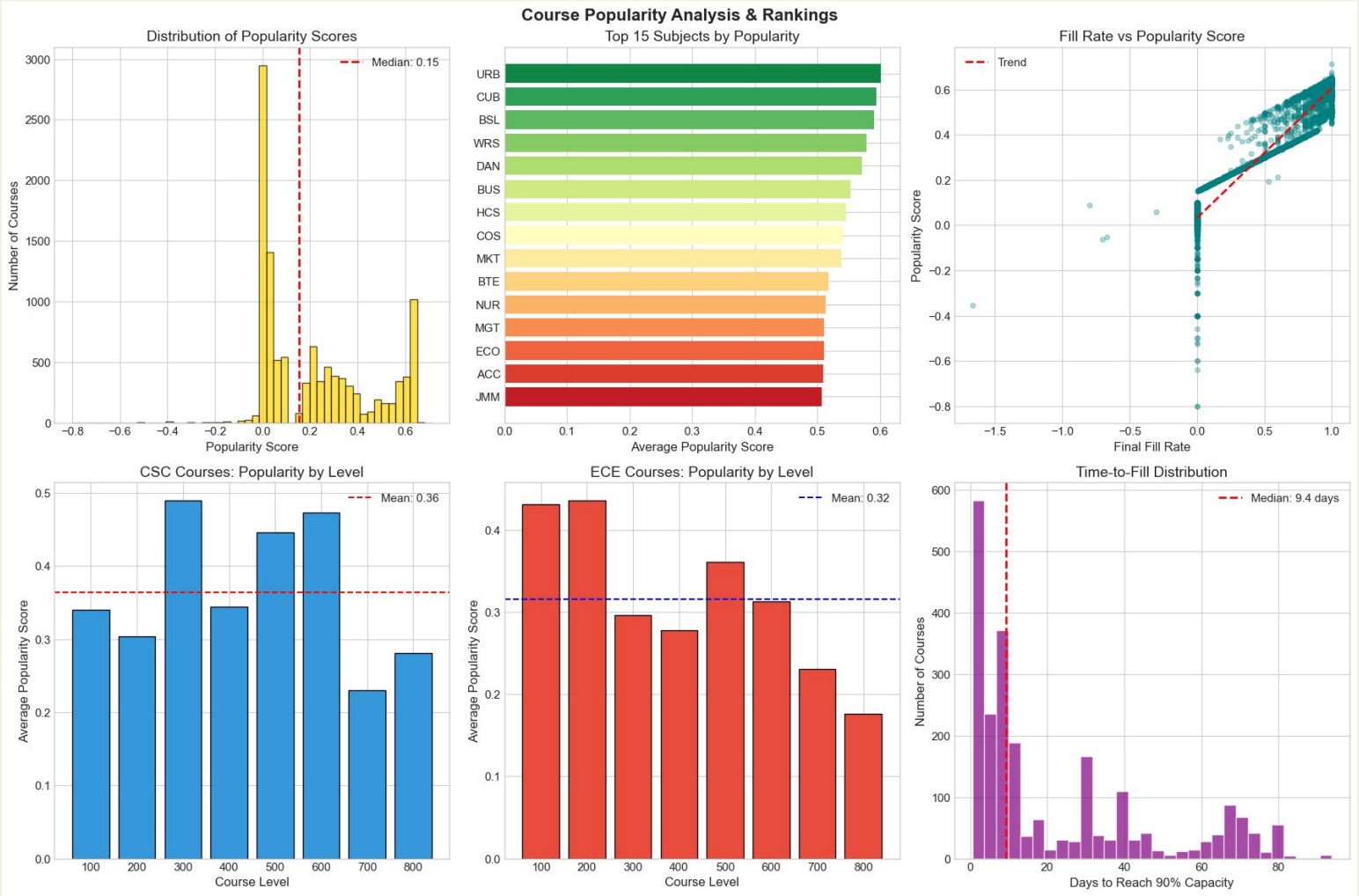
Created a composite popularity index for courses based on how quickly and how fully they fill, including waitlist and early rush signals.

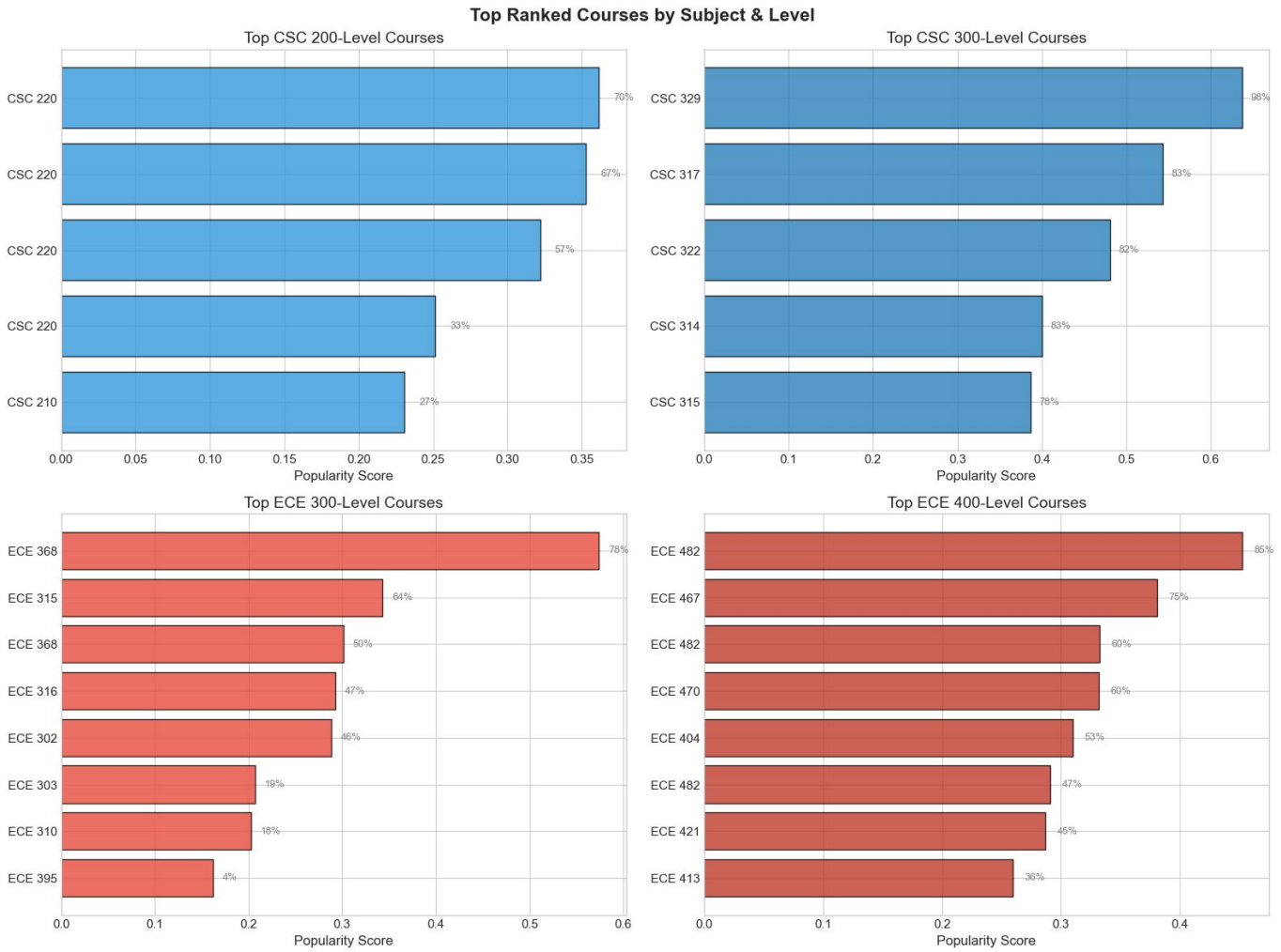
Calculate Time-to-Fill Score (days to 90% capacity)

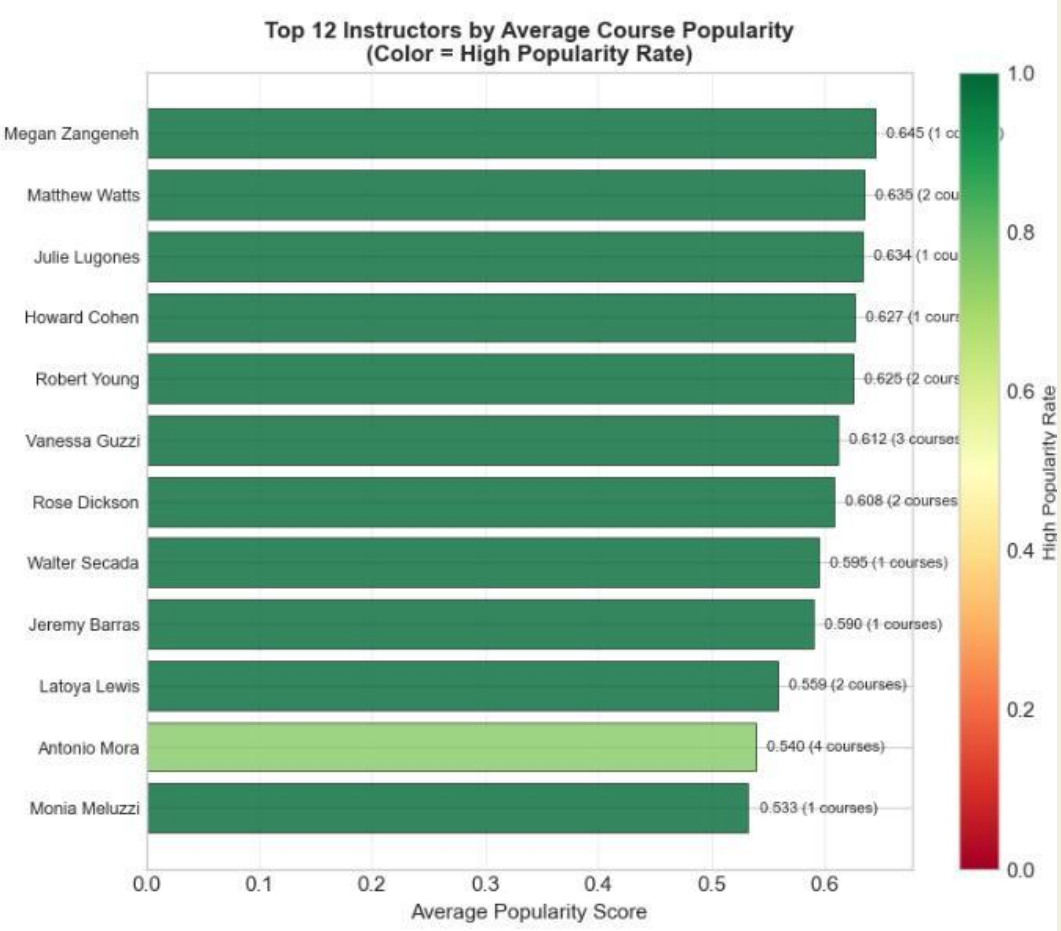
Measure how many days each course takes to reach 90% capacity and convert this into a score where faster-filling courses are more popular.

Popularity metrics summary

Calculate these metrics for 11,203 courses, providing a detailed popularity profile that includes final fill rate, velocity, time-to-fill, waitlist demand, early rush, and an overall composite score.


Course Popularity Analysis & Rankings
Distribution of Popularity Scores
Median: 0.15
Number of Courses
Popularity Score
Top 15 Subjects by Popularity
URB
CUB
BSL
WRS
DAN
BUS
HCS
COS
MKT
BTE
NUR
MGT
ECO
ACC
JMM
Average Popularity Score
Fill Rate vs Popularity Score
Trend
Popularity Score
Final Fill Rate
CSC Courses: Popularity by Level
Mean: 0.36
Average Popularity Score
Course Level
ECE Courses: Popularity by Level
Mean: 0.32
Average Popularity Score
Course Level
Time-to-Fill Distribution
Median: 9.4 days
Number of Courses
Days to Reach 90% Capacity


Top Ranked Courses by Subject & Level
Top CSC 200-Level Courses
CSC 220
CSC 220
CSC 220
CSC 220
CSC 210
70%
67%
57%
33%
27%
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Popularity Score
Top CSC 300-Level Courses
CSC 329
CSC 317
CSC 322
CSC 314
CSC 315
98%
83%
82%
83%
78%
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Popularity Score
Top ECE 300-Level Courses
ECE 368
ECE 315
ECE 368
ECE 316
ECE 302
ECE 303
ECE 310
ECE 395
78%
64%
50%
47%
46%
19%
18%
4%
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Popularity Score
Top ECE 400-Level Courses
ECE 482
ECE 467
ECE 482
ECE 470
ECE 404
ECE 482
ECE 421
ECE 413
85%
75%
60%
60%
53%
47%
45%
36%
0.0
0.1
0.2
0.3
0.4
Popularity Score


Top 12 Instructors by Average Course Popularity
(Color = High Popularity Rate)
Megan Zangeneh
Matthew Watts
Julie Lugones
Howard Cohen
Robert Young
Vanessa Guzzi
Rose Dickson
Walter Secada
Jeremy Barras
Latoya Lewis
Antonio Mora
Monia Meluzzi
0.645 (1 co
0.635 (2 cou
0.634 (1 cou
0.627 (1 cours
0.625 (2 cours
0.612 (3 courses
0.608 (2 courses
0.595 (1 courses)
0.590 (1 courses)
0.559 (2 courses)
0.540 (4 courses)
0.533 (1 courses)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Average Popularity Score
1.0
0.8
0.6
0.4
0.2
0.0
High Popularity Rate

1. Introduction & Motivation

2. Cleaning & Preprocessing

3. Feature Engineering

4. Modeling

5. Conclusion

Conclusion + Future Directions

Our analysis provides metrics that can guide resource allocation and improve course selection planning for student with the potential to reduce burden on advisors due to better planning.

Potential Future Avenues:

→ Add notifications for courses with unusually low or high enrollment

→ Flag courses that may benefit from targeted marketing and outreach

References

Advisor to student ratio/caseload resources. NACADA. (n.d.). https://nacada.ksu.edu/Resources/Clearinghouse/View-Articles/Advisor-to-Student-Ratio-Caseload-Resources.aspx

Amos, D. (2024, March 28). A practical introduction to web scraping in Python. Real Python. https://realpython.com/python-web-scraping-practical-introduction/

Digest of Education Statistics, 2021. National Center for Education Statistics (NCES) Home Page, a part of the U.S. Department of Education. (n.d.). https://nces.ed.gov/programs/digest/d21/tables/dt21_326.10.asp

University of Miami (2021). CaneLink. https://canelink.miami.edu/