# Can We Predict Which Students Won't Show Up?

A Data Science Approach to College Enrollment

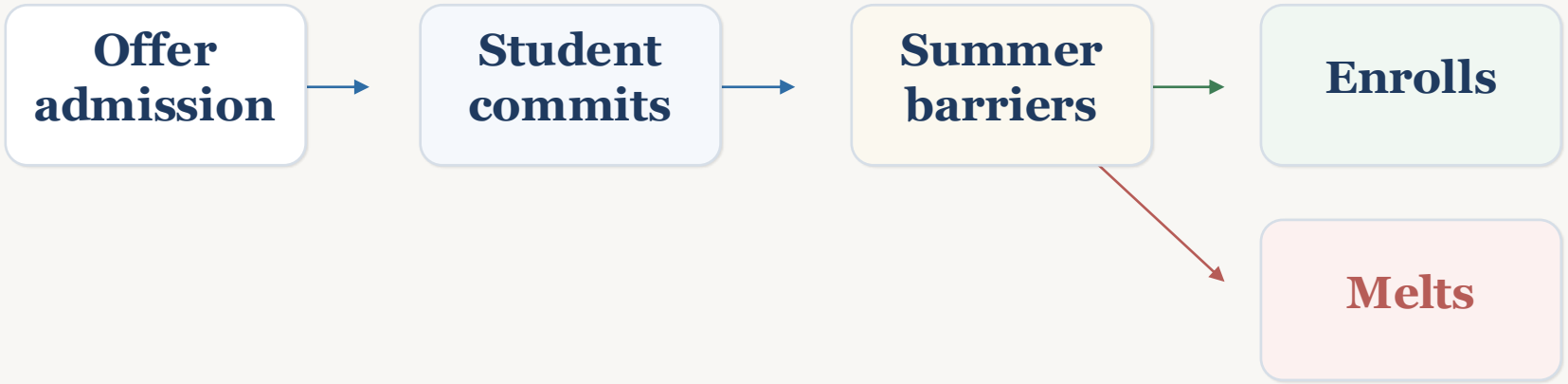
**Applied Data Science • Higher Education**

Osei Tweneboah, Ph.D.
*Assistant Professor of Data Science*

Ramapo College of New Jersey
Faculty Scholarship Symposium

# What is “Summer Melt”?

Students who say “yes” in the spring but do not arrive in the fall.




## Common reasons students melt

- financial aid gaps
- missed paperwork or deadlines
- low social support
- housing / orientation hurdles
- a competing offer
- personal circumstances

Data science matters because institutions can intervene during the summer window rather than discover the loss only after students do not arrive.

# Why this problem matters

Summer melt is not just an admissions issue; it affects planning, equity, and student support.

## Three numbers to keep in view

**10%–20%** estimated national summer-melt range

**40%** Students from low-income households

**1/3** High school seniors defer or cancel admission offer

## Institutional stakes

### Budget & Forecasting

Tuition revenue and class-size planning become less predictable when committed students do not enroll.

### Operations

Housing, dining, course sections, and advising capacity all depend on a realistic incoming-class estimate.

### Student Success & Equity

If at-risk students can be identified early, institutions can intervene with support rather than react after the fact.

# Study design and data

A six-year admissions dataset and a three-way prediction task.

## Dataset snapshot

**2014–2019** six admission cycles

**14,226** admitted students

**2,443** students who committed

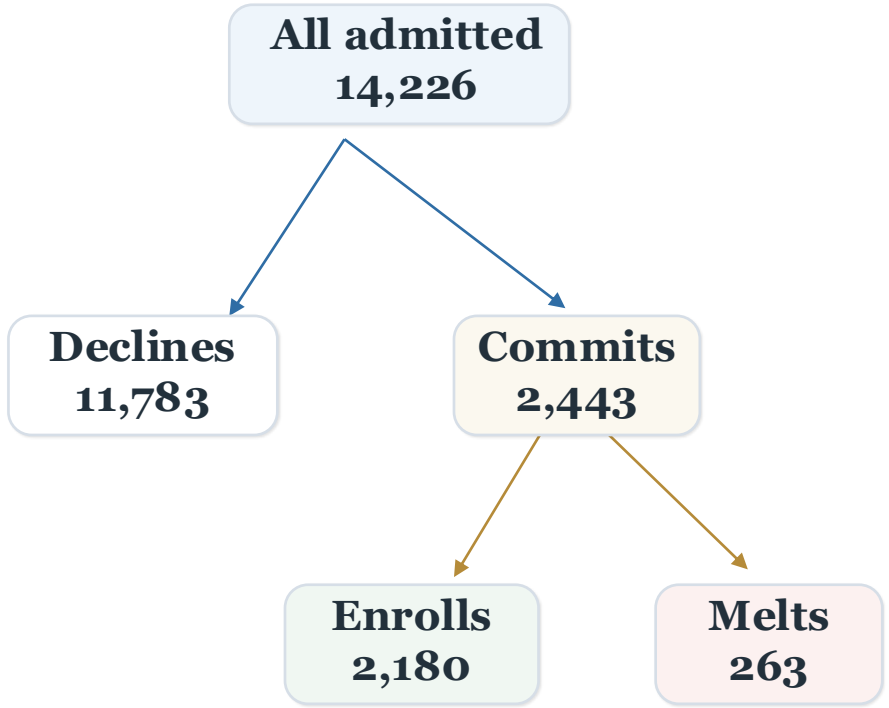
**15** admissions variables / features

**3 outcomes** decline, enroll, melt

### Examples of model inputs

- GPA
- financial need
- campus visit
- first-gen status
- Legacy
- academic interest

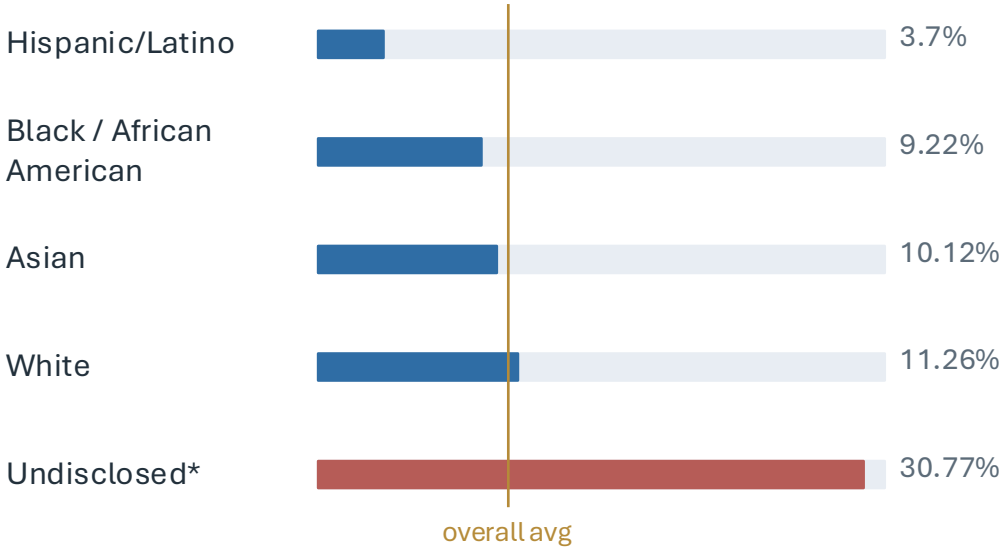
## Prediction structure (Hierarchical)



# What patterns appear in the data?

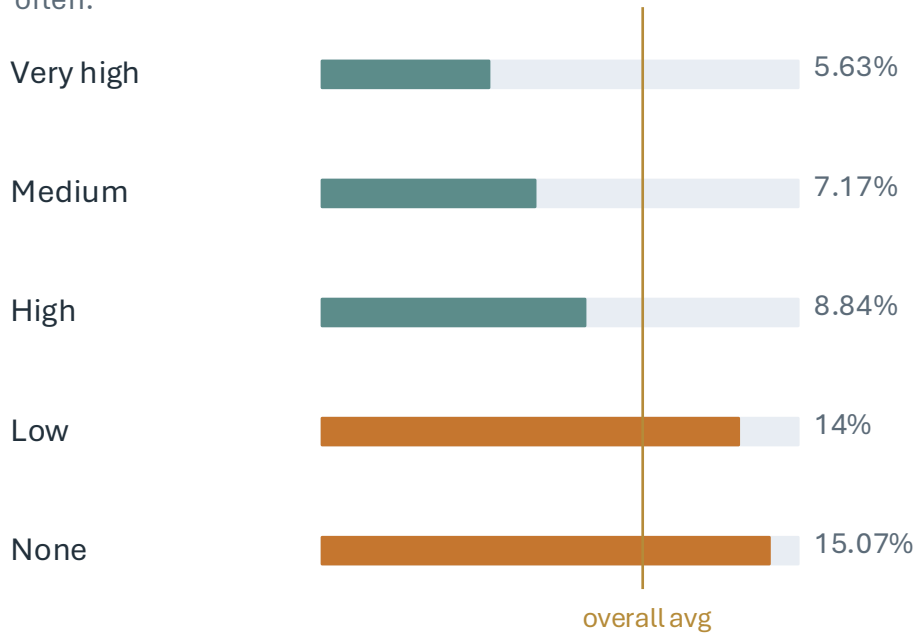
## Melt rate by ethnicity

Selected categories shown; overall average = 10.77%




## Melt rate by financial need

At Occidental, low / no need students melted more often.




**Takeaway:** descriptive patterns differ by institution, so local predictive modeling can be more informative than national generalizations alone.

# From admissions records to early-warning flags

### Assemble admissions data

1 Use variables already collected in the admissions process: GPA, need, campus visit, reader ratings, source type, and more.

### Clean and structure the problem

2 Treat each admitted student as one observation and assign one of three outcomes: decline, enroll, or melt.

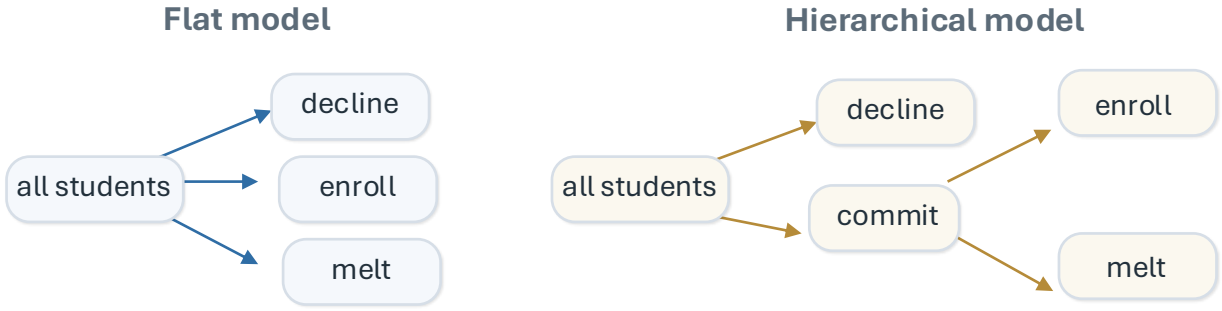
### Train predictive models

3 Compare a flat multiclass classifier with a hierarchical approach that separates committed students before predicting enroll vs. melt.

### Use the output for support

4 Flag students for outreach, not punishment. The model is a triage tool that tells staff where to look first.

## Two structures compared




## What makes the problem hard?

- Only about 10% of committed students melt, so the data are highly imbalanced.
- A model can look “accurate” overall while still missing the students we most care about.
- That is why the study tests oversampling and undersampling strategies, not just one default model.

# Why recall matters

For this problem, missing at-risk students is more costly than casting a wide net.

## Target metric

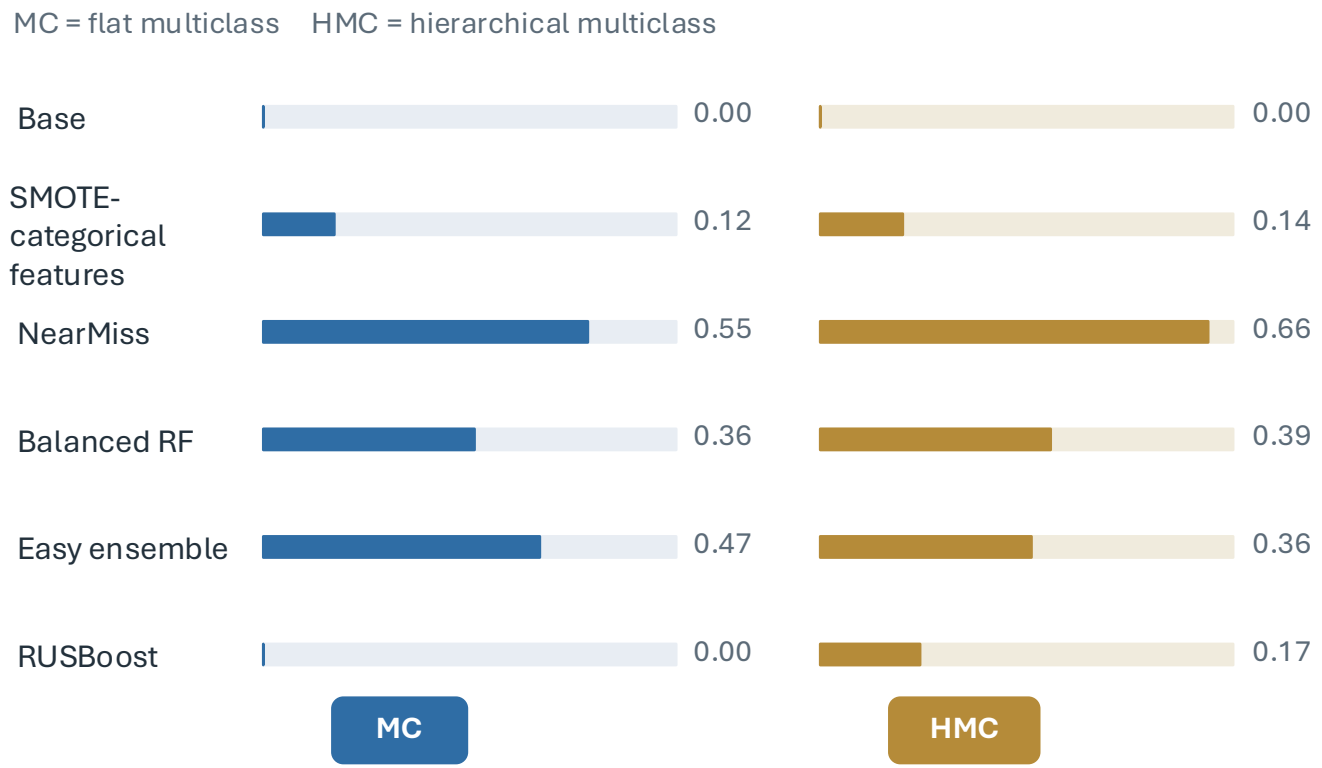
$$\text{Recall}_{\text{melt}} = \frac{\text{students correctly flagged as melt}}{\text{all students who actually melted}}$$

The research focuses on recall because the institution wants to catch as many likely melt cases as possible, even if that means some false alarms.

**High recall = better triage**

*Our goal is not just accuracy - it's intervention*

## Melt-class recall by strategy




**NearMiss** gives the best recall for melt, and the hierarchical version performs best overall.

# Key empirical takeaway

The best-performing model catches many more melt cases than a naive baseline, but it is still a screening tool.

**66%**

melt recall from the hierarchical model using NearMiss undersampling

In the study, that corresponds to detecting about 50 of the 77 students in the test set who actually melted.

best result in the study

## What this does — and does not — mean

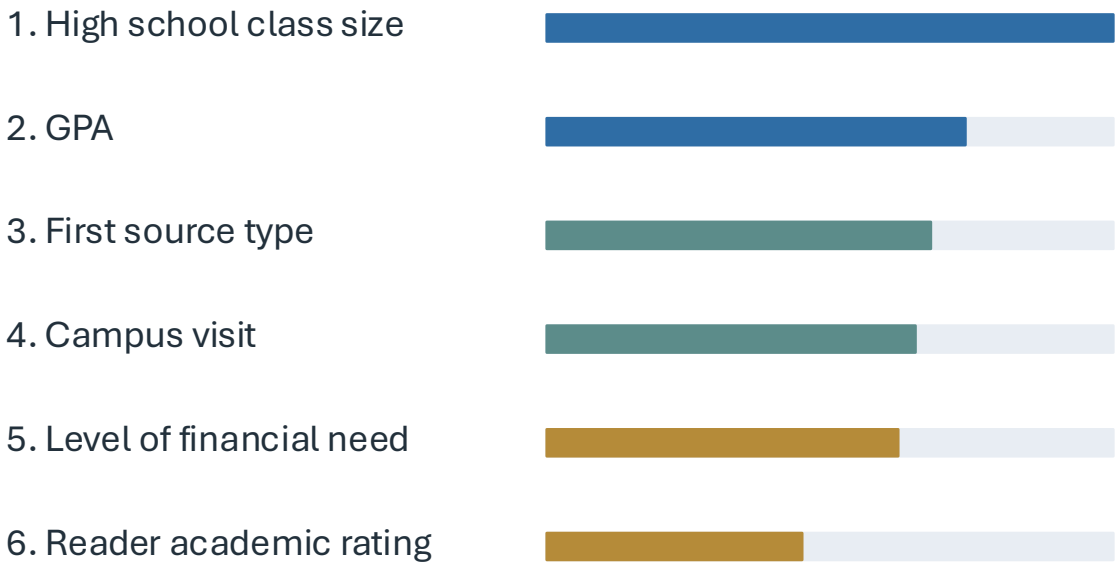
- The model can surface a manageable list of students who deserve extra attention during the summer.
- It does not explain every individual case; student circumstances can change quickly and for many reasons.
- Precision is low in the best recall model, so the output should support advising judgment, not replace it.

**Best use case: a high-recall alert list paired with human follow-up.**

# Which factors mattered most?

The strongest predictors combine academic preparation, engagement, and context.

## Top features in the tree-based models




## Interpretation

**Preparation**

Academic profile variables remain important, but they are not the whole story.

**Engagement**

Campus visits and recruitment pathway may capture how connected a student feels to the institution.

**Affordability context**

Financial need still matters, but its effect can differ across institutions.

# What institutions can do with this

Prediction becomes useful only when it is linked to concrete summer support.

## Targeted outreach

- prioritize text / email nudges
- review aid and paperwork bottlenecks
- route students to peer or staff support

## Operational planning

- improve enrollment forecasting
- adjust advising and course capacity
- plan housing, dining, and orientation resources

## Equity lens

- check who is being flagged
- audit for bias and false positives
- use alerts to add support, not deny opportunity

Data Science becomes persuasive when it improves real decisions for real students.

# A Ramapo pilot opportunity

The framework is portable if the process stays transparent and human-centered.

## Possible pilot roadmap

1

### Audit the data

What admissions variables already exist, and which are reliable enough to use?

2

### Build a baseline

Start with interpretable models and compare them against simple business-as-usual heuristics.

3

### Validate with staff

Work with admissions and student-success teams to review flags and refine the workflow.

4

### Use as decision support

Generate outreach lists, monitor outcomes, and keep a human-in-the-loop process.

## Guardrails

- Check for bias across student groups.
- Do not treat predictions as facts.
- Use the model to offer support, not to withhold opportunity.
- Reassess the model as policies, aid, and applicant pools change.

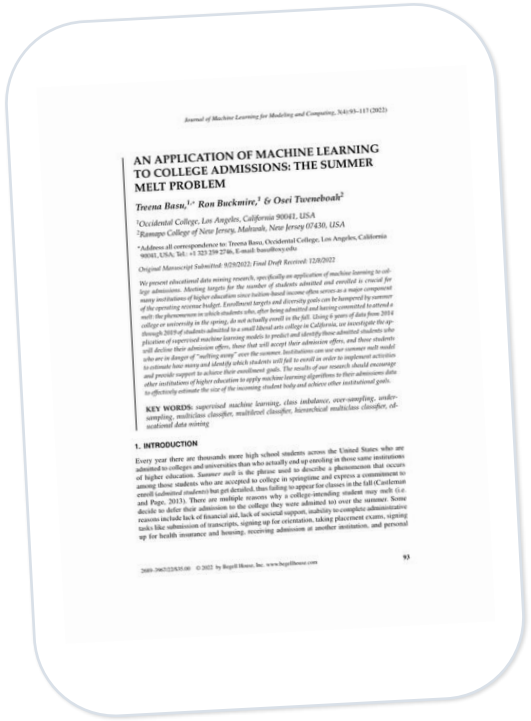
## Closing takeaway

# Applied data science is most valuable when it helps institutions act earlier and support students more effectively.

In this case, the contribution is simple but powerful: use admissions data to identify students who may need help before summer melt becomes a fall surprise.

**Questions?**

Thank you



AN APPLICATION OF MACHINE LEARNING TO COLLEGE ADMISSIONS: THE SUMMER MELT PROBLEM

Joint published work
with Treena Basu and Ron Buckmire