Projecting Incoming Cohort Size, Characteristics, and Course Enrollments via Machine Learning

TARLETON STATE UNIVERSITY
Member of The Texas A&M University System

Texas Assoc of Institutional Researchers
February 26, 2025

Dr. Scott Cook
Chief Data Scientist - University Strategy
Associate Professor - Mathematics
scook@tarleton.edu

Dr. Morgan Carter
Assoc Vice President of Institutional Data & Analytics
TAIR President
mcarter@tarleton.edu

Admitted Matriculation Projection (AMP)

- Goal: Project Fall 2024 course-level enrollment of incoming admitted students
- For each incoming admitted student, AMP predicts probability to:
    - Matriculate (take any course)
    - Enroll in each individual course (~75 specified by provost)
- Aggregate projections for course, college, gender, campus*, hs pctl, state, county, etc
- Data Sources
    - Enrollment Management's weekly "Flags report" of admitted students
        - Current
        - Same date 2021, 2022, 2023
    - Course enrollments
        - Current
        - Same date 2021, 2022, 2023
        - Census date 2021, 2022, 2023

Before 2023

2023

2024

Future?

Inputs

- Age
- Application date
- TX residency
- ACT/SAT score
- High school quartile
- Gender
- Race/Ethnicity*
- Major college
- Gap score

- Driving distance home to campus
- Legacy
- Attended orientation
- TSI scores (math/reading/writing)
- Campus
- Scholarship
- ~~*Submitted FAFSA*~~
- Fee waiver
- Logged in student info system (ssb)

*SB 17, SFFA, & DCL compliant

Outputs

- Individual student level
  - Probability to enroll in course X
  - Shapley scores (influence from each input on each prediction)
- Aggregate projections for university, college, major, course
  - Enrollment of FTIC, transfer, and returning students
  - Historical errors analysis
  - Prediction intervals (in progress)

Dataset

- 1 row per admitted student
- 1 column per student data element (previous slide)
- 2 columns per course
  - enroll_current: Was student enrolled in this course on this day? (T/F)
  - enroll_census: Was student enrolled in this course at census? (T/F)
- Common preprocessing (standard rescaling, one-hot-encoding, etc)
- Discuss missing values later

Supervised Machine Learning

- 3 student types: FTIC, transfer, returning (not continuing)
- Train separate models for each (course, student type) using:
  - Rows for that student type
  - Features: student data + enroll_current for this course
  - Target: enroll_census for this course

Supervised Machine Learning

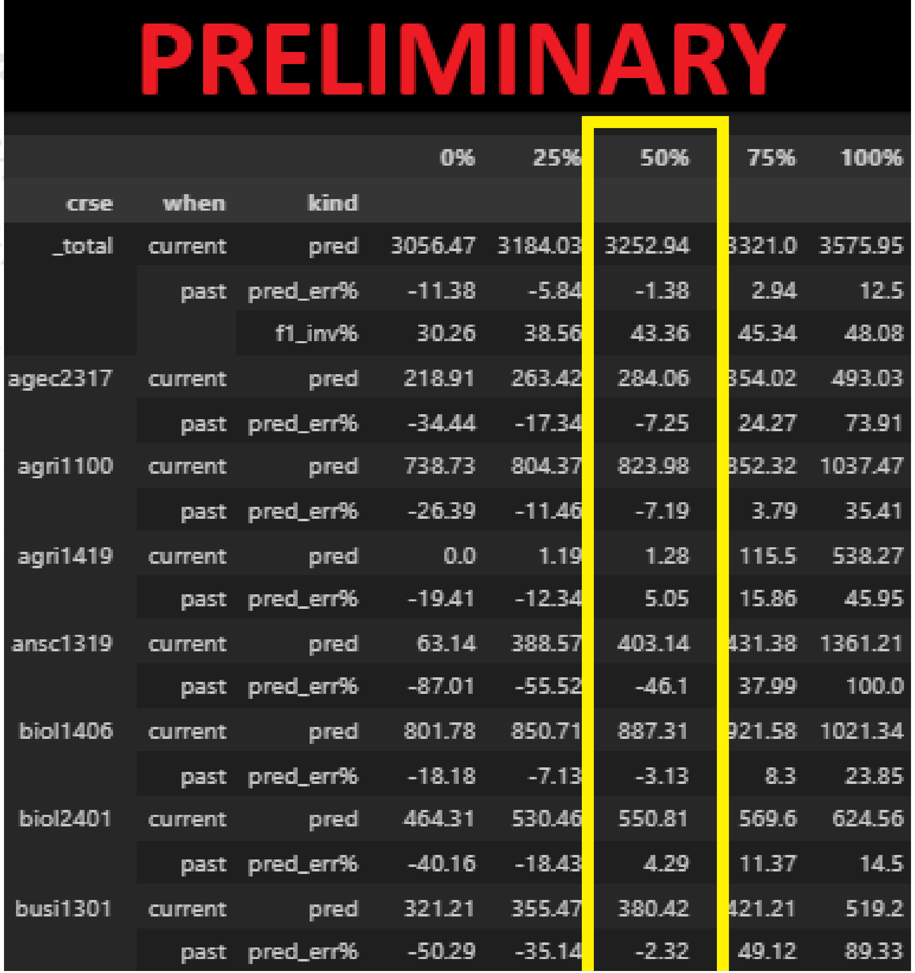
- Binary classification task with mixed data types
- Decision tree-based classifiers work best (Random Forest, LightGBM, XGBoost, Histogram Gradient Boosting Trees)
- FLAML: Fast Library for Automated Machine Learning
  - Microsoft Research open-source Python automated machine learning (2021)
  - Optimized hyperparameter tuning without human intervention
  - Adjustable “time budget” to prevent run-away jobs
- “predict_proba” estimates probability that each student will be enrolled in specified course at census
- May need probability calibration for accurate aggregations (course, college, etc)

Missing Data

- Data mostly complete except ACT/SAT (~⅓ missing)
- Highly predictive (see Shapley) → do not want to drop → impute missing values
- Not missing at random - motivated, well-prepared students submit ACT/SAT at higher rates AND have higher scores AND are more likely to matriculate
  - Missingness correlated with target
  - Imputing missing ACT/SAT with mean ACT/SAT would overestimate
- MiceForest
  - Advanced imputation of missing values using iterative LightGBM
  - Multiple imputation → prediction intervals (turned out too narrow - tweaking)

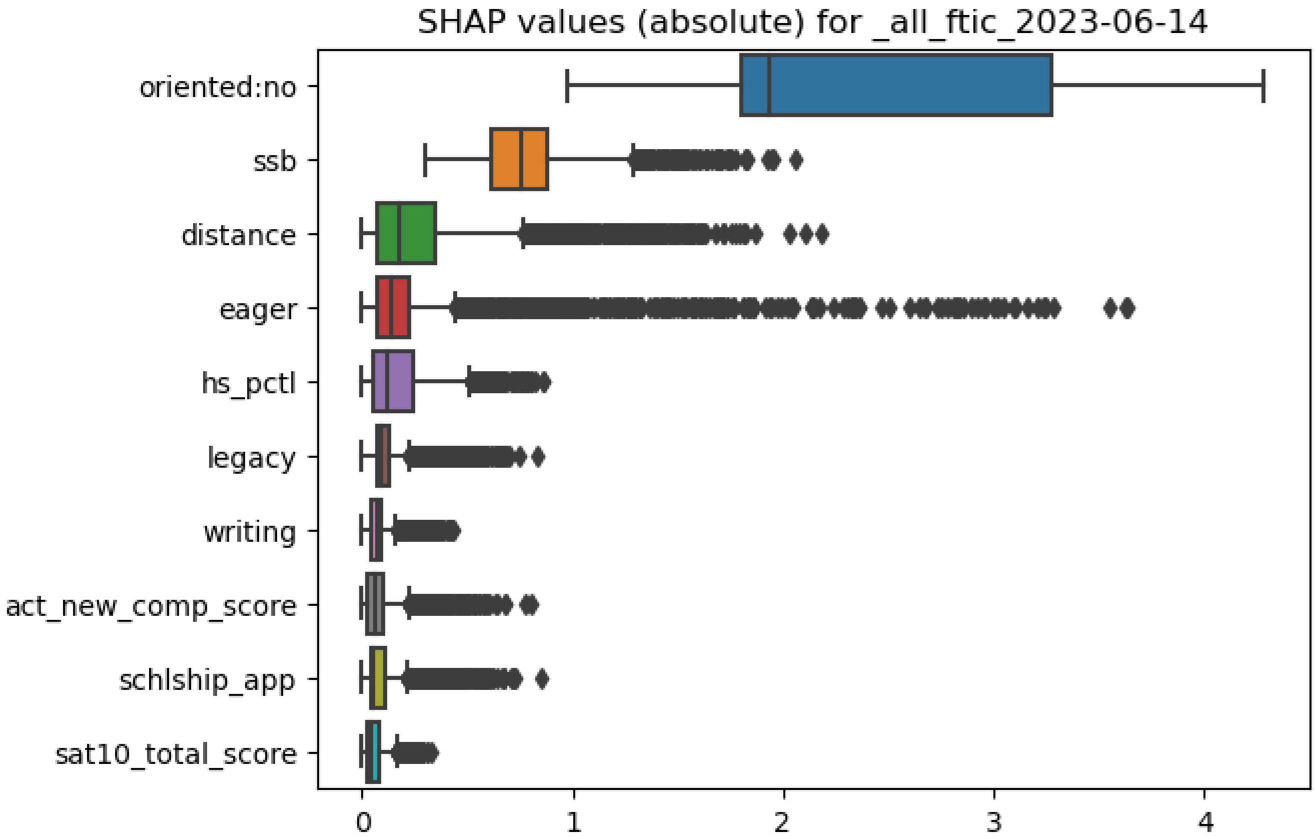
Lagging Applicants

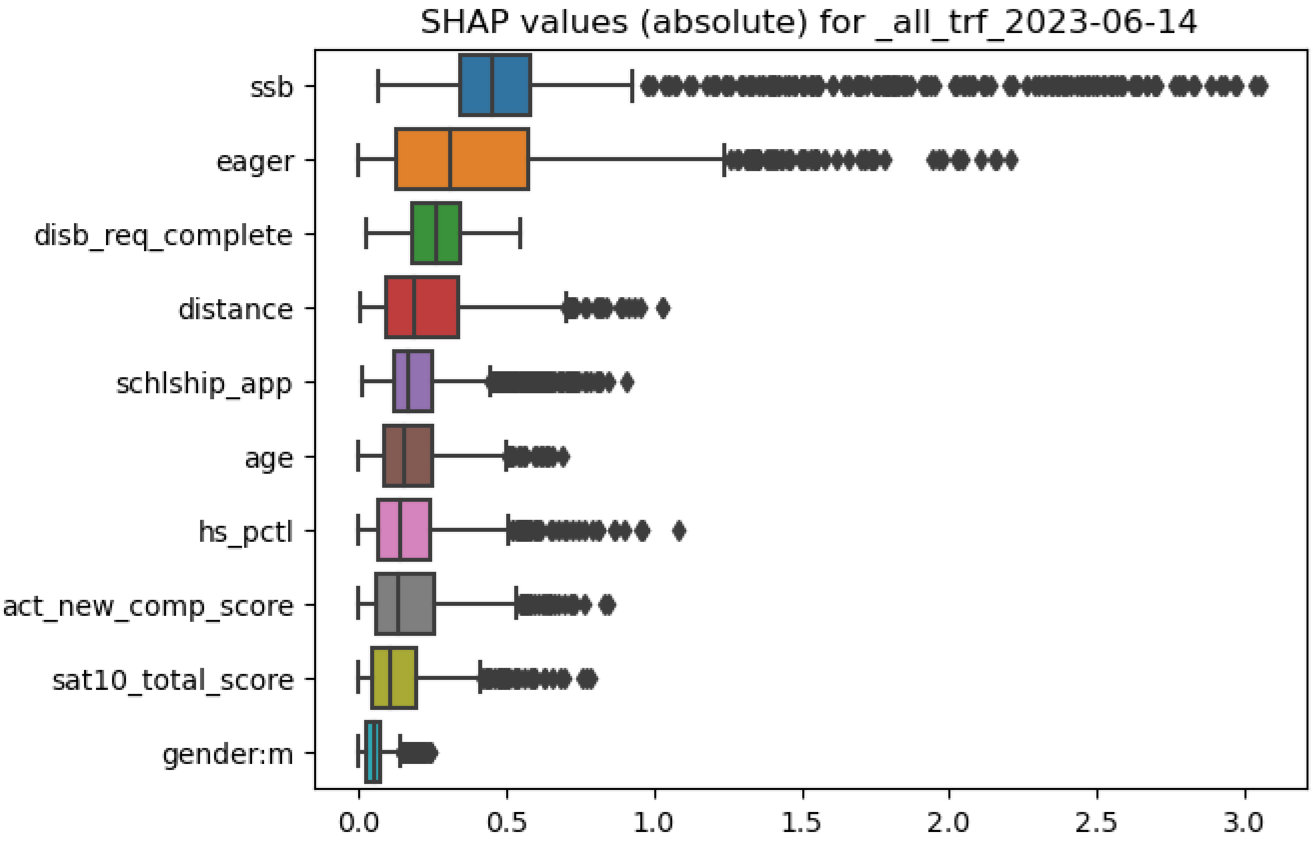
- AMP models students that have already applied (eager)
- What about students that will apply between now and Fall? (lagging)
- Key assumption: Rate & characteristics of this year’s lagging applicants will be similar to same period in prior years
  - Compute lagging/eager ratio for prior years (remarkably stable for FTIC, transfer, returning separately)
  - Project 2024 based on eager applicants
  - Inflate using prior lagging-eager ratios → models 2024 lagging applicants
  - Vulnerable to year-over-year changes (ex: earlier admission, different orientation cadence, FAFSA disruption, policy changes, etc)

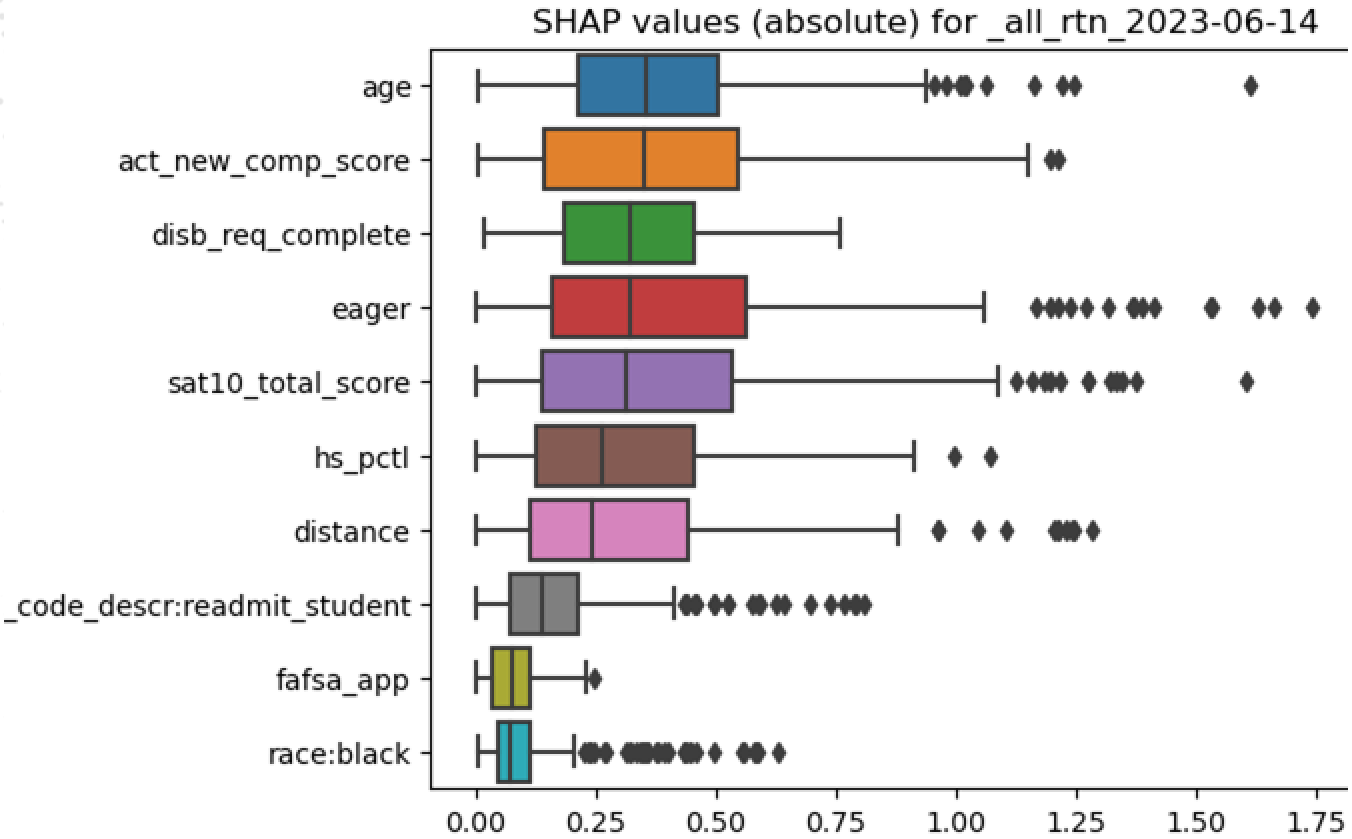


PRELIMINARY

			0%	25%	50%	75%	100%
crse	when	kind					
_total	current	pred	3056.47	3184.03	3252.94	3321.0	3575.95
	past	pred_err%	-11.38	-5.84	-1.38	2.94	12.5
		f1_inv%	30.26	38.56	43.36	45.34	48.08
agec2317	current	pred	218.91	263.42	284.06	354.02	493.03
	past	pred_err%	-34.44	-17.34	-7.25	24.27	73.91
agri1100	current	pred	738.73	804.37	823.98	852.32	1037.47
	past	pred_err%	-26.39	-11.46	-7.19	3.79	35.41
agri1419	current	pred	0.0	1.19	1.28	115.5	538.27
	past	pred_err%	-19.41	-12.34	5.05	15.86	45.95
ansc1319	current	pred	63.14	388.57	403.14	431.38	1361.21
	past	pred_err%	-87.01	-55.52	-46.1	37.99	100.0
biol1406	current	pred	801.78	850.71	887.31	921.58	1021.34
	past	pred_err%	-18.18	-7.13	-3.13	8.3	23.85
biol2401	current	pred	464.31	530.46	550.81	569.6	624.56
	past	pred_err%	-40.16	-18.43	4.29	11.37	14.5
busi1301	current	pred	321.21	355.47	380.42	421.21	519.2
	past	pred_err%	-50.29	-35.14	-2.32	49.12	89.33


SHAP values (absolute) for _all_ftic_2023-06-14
oriented:no
ssb
distance
eager
hs_pctl
legacy
writing
act_new_comp_score
schlship_app
sat10_total_score
0
1
2
3
4


SHAP values (absolute) for _all_trf_2023-06-14
ssb
eager
disb_req_complete
distance
schlship_app
age
hs_pctl
act_new_comp_score
sat10_total_score
gender:m
0.0
0.5
1.0
1.5
2.0
2.5
3.0


SHAP values (absolute) for _all_rtn_2023-06-14
age
act_new_comp_score
disb_req_complete
eager
sat10_total_score
hs_pctl
distance
_code_descr:readmit_student
fafsa_app
race:black
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75

Results

Dr. Javier Garza, Vice President for Enrollment Management:

- In Fall 2024, FTIC headcount was up 11% but FTIC semester credit hours were up 14%. Historically, these are equal.
- He believes AMP is the only salient difference & credits it with the extra 3% SCH (approx $350,000)
- He believes AMP gave dept heads better estimates for course demand early enough to create sections & hire instructors.
- This gave advisors more options to put students into additional courses, generating SCH growth independently of headcount growth.

Additions & Improvements

- Incorporate high school course grades via new transcript OCR
- Project housing demand
- Prediction intervals
- Dashboard
- Train on single year or multiple?
- Adjust training process to handle course-specific year-over-year-changes
- Course-specific vs university-wide inflation factors
- Lower-level course demand at Ft. Worth campus
- Causal Machine Learning

Causal ML

- Be careful when interpreting predictive models in search of causal insights — SHAP latest documentation

Scan the QR code to complete the session survey.

Texas Association for Institutional Research

Annual Conference: February 25-28, 2025
Omni Hotel in Corpus Christi, TX