Advances in Machine Learning to Forecast Incoming Cohort Size, Characteristics, & Course Enrollments

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

Texas Association for Institutional Research
February 25, 2026

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

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

Learning Outcomes

- Understand Tarleton's EDW history, movement to cloud technologies, and first machine learning (ML) model

- Evaluate your existing systems in conjunction with Tarleton's approach

- Create an action plan to modernize your environment to do ML models and predictive analytics

About Tarleton State University

- Level VI SACSCOC institution
- Carnegie High Research Activity
- NCAA Division I
- Rapidly growing regional institution within TAMUS
- 21K enrollment

About You

Poll

Infrastructure

- Oracle Database / Warehouse
- Cron Jobs build daily tables
- Texan Facts (WebFocus)

- Azure Databricks
- Several Source systems
- Power BI
- Machine Learning in Azure Databricks
- Inherited older research server

Admitted Matriculation Projection (AMP)

- Goal: Forecast course-level enrollment of incoming Fall cohort
- For each (admit, course), estimate probability that admit enrolls in that course
- Sum probabilities to get headcount forecasts for course, campus, college, dept, major, high school quartile, TSI status, etc
- Data Sources
    - Enrollment Management’s weekly “Flags report” of admitted students
        - Current date 2023, 2024, 2025, 2026
    - Course enrollments
        - Current date 2023, 2024, 2025, 2026
        - Stable date 2023, 2024, 2025

stable date = Wednesday 2 weeks after census

Before 2023

2023

2024

2025

2026

Future

Inputs

- ACT/SAT score
- Age
- Application date
- Campus
- College
- Current enrollment (SCH)
- Drivetime home to campus
- Gap score
- Gender
- High school quartile
- Housing application (soon)
- Last ssb login
- Legacy
- Orientation attendance
- Scholarship
- TSI math/reading/writing
- TX residency
- Fee Waiver

Dataset

- 1 row per (admit, course)
- Columns:
  - course code
  - data elements (previous slide)
  - Was student enrolled in this course on current date? (T/F)
  - Was student enrolled in this course on stable date? (T/F)
- Common preprocessing (standard rescaling, one-hot-encoding, etc)
- Discuss missing values later

Outputs

- Granular
  - For each (admit, course), probability that student enrolls in that course
  - Log-loss (error) & Shapley values (explainability)
- Aggregate
  - Headcount forecasts (course, campus, college, dept, major, hs quartile, TSI status)
  - Historical error analysis
  - Prediction intervals

Supervised Machine Learning

- 3 incoming student types: FTIC, transfer, returning
- Open question: create separate models for student type & courses?
  - 2024 & 2025: yes
  - 2026: re-evaluating (new hardware & improved ML tools)

Supervised Machine Learning

- Binary classification task with mixed data types
- Decision tree-based classifiers work best (Random Forest, LightGBM, XGBoost, Histogram Gradient Boosting Trees, CatBoost)
- ~~FLAML: Fast Library for Automated Machine Learning~~
- AutoGluon: Amazon Web Services AI group
    - Optimized hyperparameter tuning without human intervention
    - Adjustable “time limit” to prevent run-away jobs
    - Automatic probability calibration
    - Automatically handles categorical
- “predict_proba_oof” = probability that student enrolls in that course

Missing Data

- ACT/SAT optional → missing values
- Highly predictive → 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 correlates with target and other features
  - Imputing missing via mean/median ACT/SAT overestimates
- MiceForest
  - Advanced imputation of missing values using iterative LightGBM
- AutoGluon also offers automatic imputation, but I haven’t assesed it yet

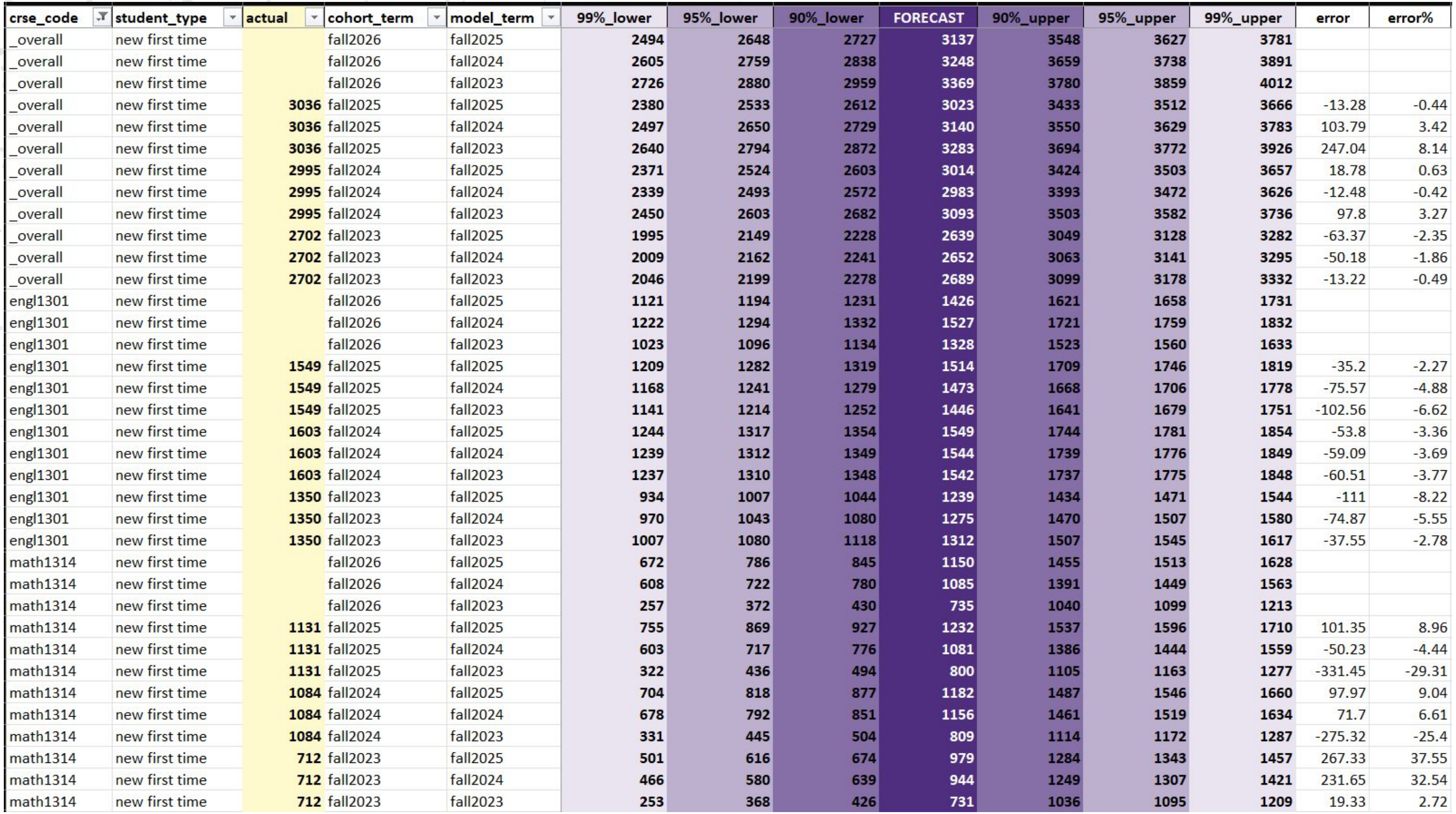
Lagging Applicants

- AMP models students that have already applied (eager)
- How to model 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)
    - Run AMP on eager applicants
    - Inflate headcount using historical lagging-eager ratios
    - Vulnerable to year-over-year changes (ex: earlier admission, different orientation cadence, FAFSA disruption, policy changes, etc)

Single-batch or Blended?

- Training data from 3 Fall cohorts (2023, 2024, 2025). Should we create 3 separate models (single-batch) or 1 combined (blended) model
- Trial & error → single-batch is better. Why?
  - Courses created, destroyed, moved in/out of core, added/removed from majors, restructured, etc
  - Majors & departments move to different colleges
  - Local/transient effects (staffing shortages, advising patterns, etc)
- Can dramatically change underlying patterns & cause model instability
- With single-batch models, leaders with institutional knowledge may be able to “make sense” of year-over-year differences (if they remember the transient effect)
- Impossible with blended models

Fall 2026 Forecasts - Feb 18 data



crse_code	student_type	actual	cohort_term	model_term	99%_lower	95%_lower	90%_lower	FORECAST	90%_upper	95%_upper	99%_upper	error	error%
_overall	new first time		fall2026	fall2025	2494	2648	2727	3137	3548	3627	3781		
_overall	new first time		fall2026	fall2024	2605	2759	2838	3248	3659	3738	3891		
_overall	new first time		fall2026	fall2023	2726	2880	2959	3369	3780	3859	4012		
_overall	new first time	3036	fall2025	fall2025	2380	2533	2612	3023	3433	3512	3666	-13.28	-0.44
_overall	new first time	3036	fall2025	fall2024	2497	2650	2729	3140	3550	3629	3783	103.79	3.42
_overall	new first time	3036	fall2025	fall2023	2640	2794	2872	3283	3694	3772	3926	247.04	8.14
_overall	new first time	2995	fall2024	fall2025	2371	2524	2603	3014	3424	3503	3657	18.78	0.63
_overall	new first time	2995	fall2024	fall2024	2339	2493	2572	2983	3393	3472	3626	-12.48	-0.42
_overall	new first time	2995	fall2024	fall2023	2450	2603	2682	3093	3503	3582	3736	97.8	3.27
_overall	new first time	2702	fall2023	fall2025	1995	2149	2228	2639	3049	3128	3282	-63.37	-2.35
_overall	new first time	2702	fall2023	fall2024	2009	2162	2241	2652	3063	3141	3295	-50.18	-1.86
_overall	new first time	2702	fall2023	fall2023	2046	2199	2278	2689	3099	3178	3332	-13.22	-0.49
engl1301	new first time		fall2026	fall2025	1121	1194	1231	1426	1621	1658	1731		
engl1301	new first time		fall2026	fall2024	1222	1294	1332	1527	1721	1759	1832		
engl1301	new first time		fall2026	fall2023	1023	1096	1134	1328	1523	1560	1633		
engl1301	new first time	1549	fall2025	fall2025	1209	1282	1319	1514	1709	1746	1819	-35.2	-2.27
engl1301	new first time	1549	fall2025	fall2024	1168	1241	1279	1473	1668	1706	1778	-75.57	-4.88
engl1301	new first time	1549	fall2025	fall2023	1141	1214	1252	1446	1641	1679	1751	-102.56	-6.62
engl1301	new first time	1603	fall2024	fall2025	1244	1317	1354	1549	1744	1781	1854	-53.8	-3.36
engl1301	new first time	1603	fall2024	fall2024	1239	1312	1349	1544	1739	1776	1849	-59.09	-3.69
engl1301	new first time	1603	fall2024	fall2023	1237	1310	1348	1542	1737	1775	1848	-60.51	-3.77
engl1301	new first time	1350	fall2023	fall2025	934	1007	1044	1239	1434	1471	1544	-111	-8.22
engl1301	new first time	1350	fall2023	fall2024	970	1043	1080	1275	1470	1507	1580	-74.87	-5.55
engl1301	new first time	1350	fall2023	fall2023	1007	1080	1118	1312	1507	1545	1617	-37.55	-2.78
math1314	new first time		fall2026	fall2025	672	786	845	1150	1455	1513	1628		
math1314	new first time		fall2026	fall2024	608	722	780	1085	1391	1449	1563		
math1314	new first time		fall2026	fall2023	257	372	430	735	1040	1099	1213		
math1314	new first time	1131	fall2025	fall2025	755	869	927	1232	1537	1596	1710	101.35	8.96
math1314	new first time	1131	fall2025	fall2024	603	717	776	1081	1386	1444	1559	-50.23	-4.44
math1314	new first time	1131	fall2025	fall2023	322	436	494	800	1105	1163	1277	-331.45	-29.31
math1314	new first time	1084	fall2024	fall2025	704	818	877	1182	1487	1546	1660	97.97	9.04
math1314	new first time	1084	fall2024	fall2024	678	792	851	1156	1461	1519	1634	71.7	6.61
math1314	new first time	1084	fall2024	fall2023	331	445	504	809	1114	1172	1287	-275.32	-25.4
math1314	new first time	712	fall2023	fall2025	501	616	674	979	1284	1343	1457	267.33	37.55
math1314	new first time	712	fall2023	fall2024	466	580	639	944	1249	1307	1421	231.65	32.54
math1314	new first time	712	fall2023	fall2023	253	368	426	731	1036	1095	1209	19.33	2.72

Prediction Intervals

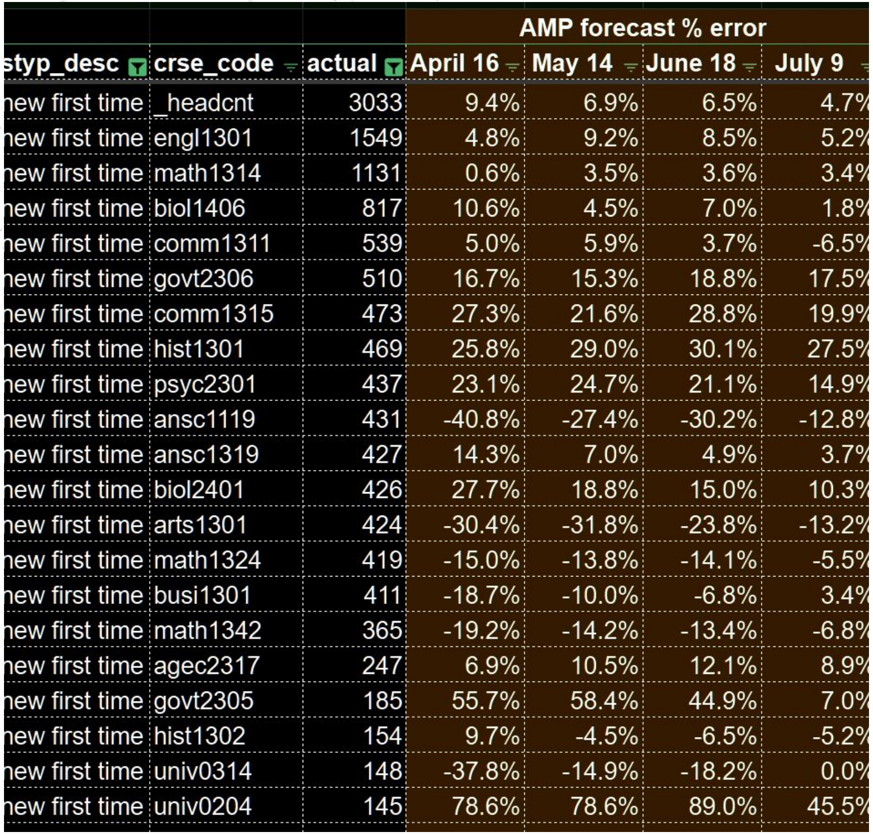
For a given course and cohort_term, let

- $p_{i,j}$ = probability that admit i enrolls in course for cohort_term as predicted by the model trained on data from model_term j
- $\Sigma_i p_{i,j} = f_j$ = headcount forecast
- $p_{i,j} * (1 - p_{i,j}) = v_{i,j}$ (binomial variance)
- $\Sigma_i v_{i,j}/n$ = model_term variance (within-group)
- $\text{var}_j(f_j)$ = forecast variance across model_terms (between-group)
- SE = sqrt(within + between)
- Prediction interval = $f_j \pm SE * z^*$

Estimating all student headcounts

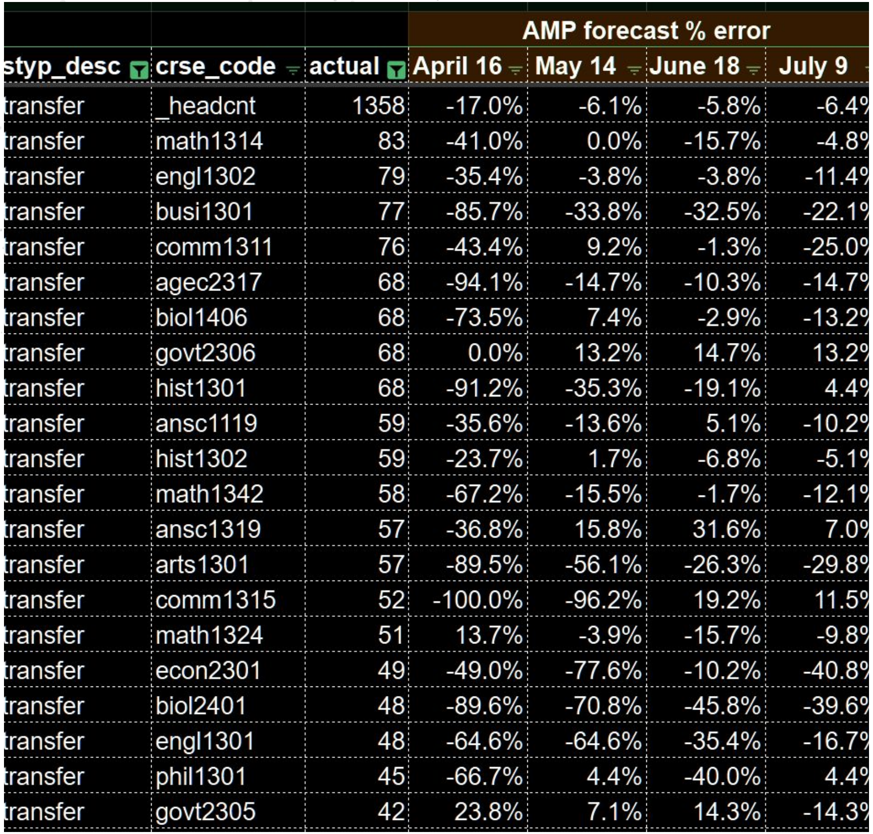
- “Thanks for forecasting incoming headcounts. How about continuing?”
- AMP FTIC forecast / true prior FTIC headcount * true prior actual total headcount
- Reasonable iff total / FTIC ratio stable

Fall 2025 Results



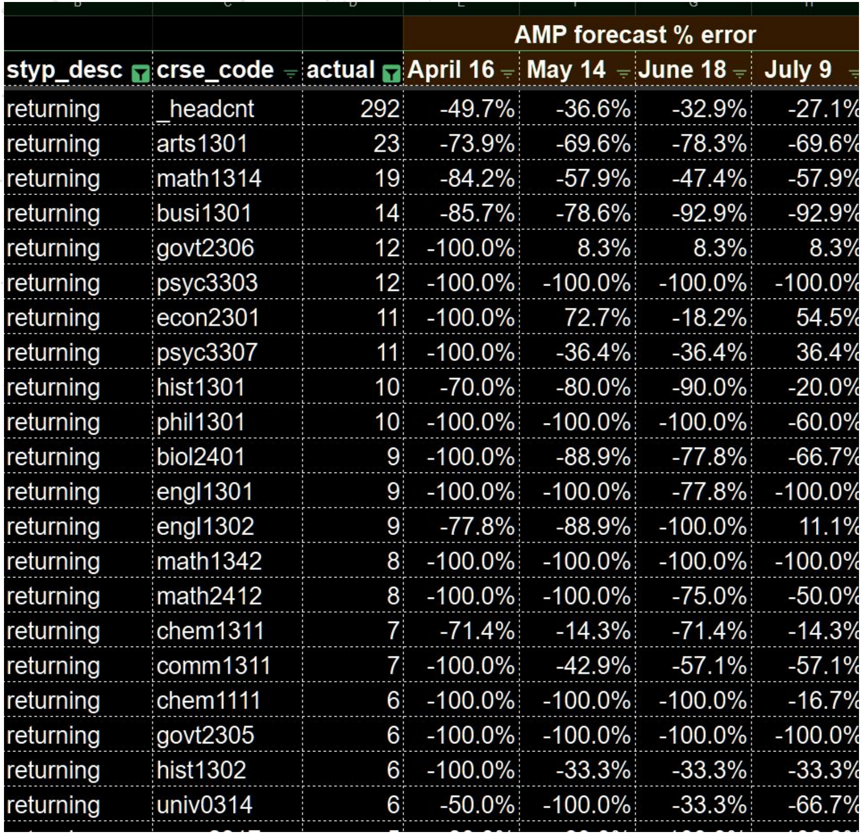
			AMP forecast % error			
styp_desc	crse_code	actual	April 16	May 14	June 18	July 9
new first time	_headcnt	3033	9.4%	6.9%	6.5%	4.7%
new first time	engl1301	1549	4.8%	9.2%	8.5%	5.2%
new first time	math1314	1131	0.6%	3.5%	3.6%	3.4%
new first time	biol1406	817	10.6%	4.5%	7.0%	1.8%
new first time	comm1311	539	5.0%	5.9%	3.7%	-6.5%
new first time	govt2306	510	16.7%	15.3%	18.8%	17.5%
new first time	comm1315	473	27.3%	21.6%	28.8%	19.9%
new first time	hist1301	469	25.8%	29.0%	30.1%	27.5%
new first time	psyc2301	437	23.1%	24.7%	21.1%	14.9%
new first time	ansc1119	431	-40.8%	-27.4%	-30.2%	-12.8%
new first time	ansc1319	427	14.3%	7.0%	4.9%	3.7%
new first time	biol2401	426	27.7%	18.8%	15.0%	10.3%
new first time	arts1301	424	-30.4%	-31.8%	-23.8%	-13.2%
new first time	math1324	419	-15.0%	-13.8%	-14.1%	-5.5%
new first time	busi1301	411	-18.7%	-10.0%	-6.8%	3.4%
new first time	math1342	365	-19.2%	-14.2%	-13.4%	-6.8%
new first time	agec2317	247	6.9%	10.5%	12.1%	8.9%
new first time	govt2305	185	55.7%	58.4%	44.9%	7.0%
new first time	hist1302	154	9.7%	-4.5%	-6.5%	-5.2%
new first time	univ0314	148	-37.8%	-14.9%	-18.2%	0.0%
new first time	univ0204	145	78.6%	78.6%	89.0%	45.5%

Fall 2025 Results

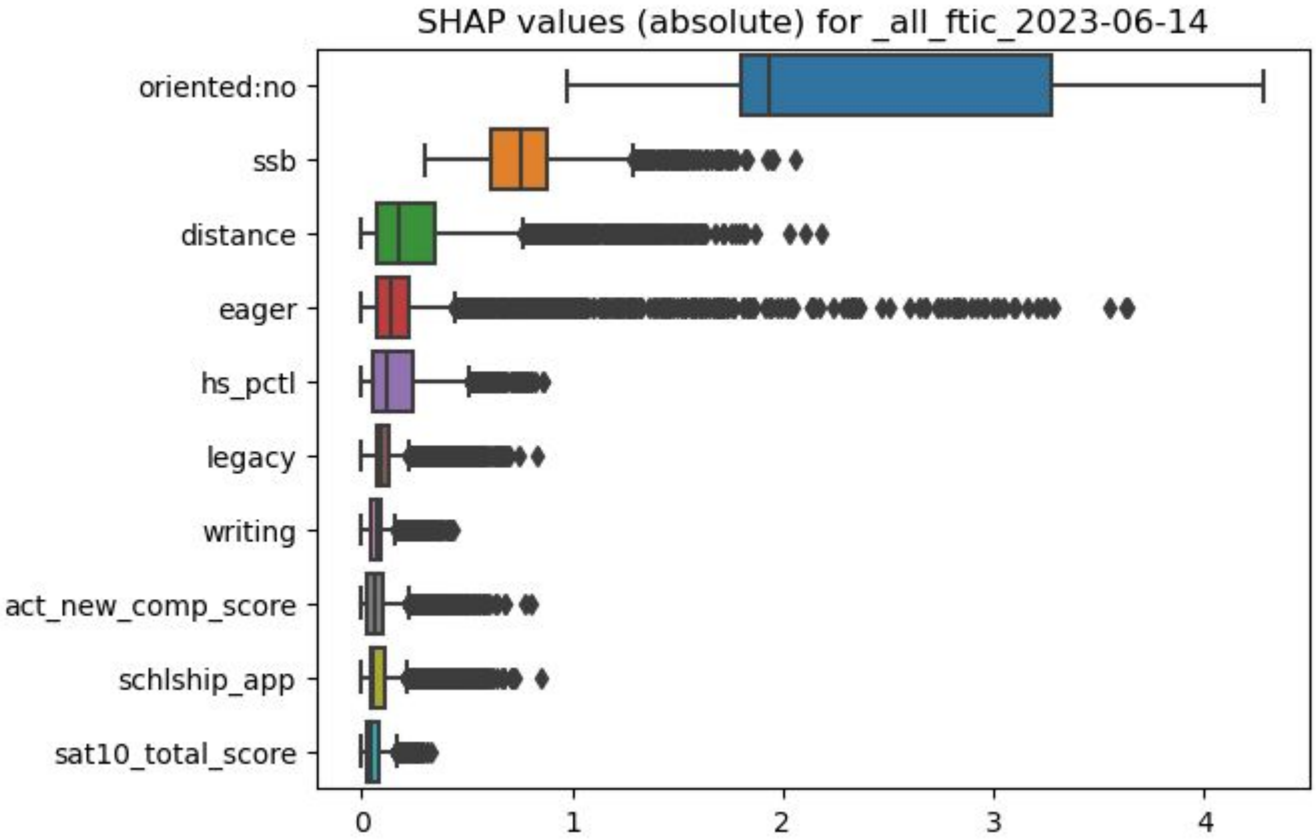


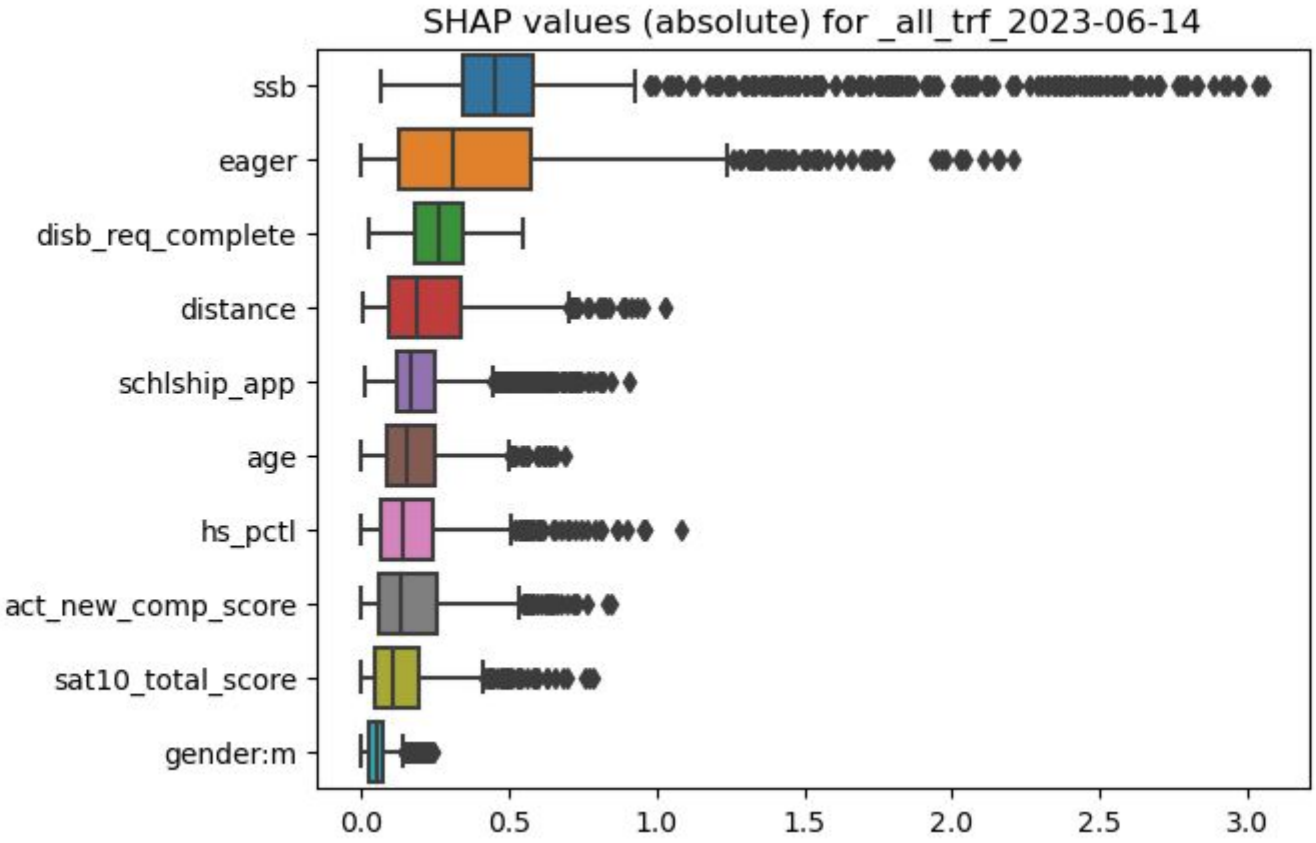
			AMP forecast % error			
styp_desc	crse_code	actual	April 16	May 14	June 18	July 9
transfer	_headcnt	1358	-17.0%	-6.1%	-5.8%	-6.4%
transfer	math1314	83	-41.0%	0.0%	-15.7%	-4.8%
transfer	engl1302	79	-35.4%	-3.8%	-3.8%	-11.4%
transfer	busi1301	77	-85.7%	-33.8%	-32.5%	-22.1%
transfer	comm1311	76	-43.4%	9.2%	-1.3%	-25.0%
transfer	agec2317	68	-94.1%	-14.7%	-10.3%	-14.7%
transfer	biol1406	68	-73.5%	7.4%	-2.9%	-13.2%
transfer	govt2306	68	0.0%	13.2%	14.7%	13.2%
transfer	hist1301	68	-91.2%	-35.3%	-19.1%	4.4%
transfer	ansc1119	59	-35.6%	-13.6%	5.1%	-10.2%
transfer	hist1302	59	-23.7%	1.7%	-6.8%	-5.1%
transfer	math1342	58	-67.2%	-15.5%	-1.7%	-12.1%
transfer	ansc1319	57	-36.8%	15.8%	31.6%	7.0%
transfer	arts1301	57	-89.5%	-56.1%	-26.3%	-29.8%
transfer	comm1315	52	-100.0%	-96.2%	19.2%	11.5%
transfer	math1324	51	13.7%	-3.9%	-15.7%	-9.8%
transfer	econ2301	49	-49.0%	-77.6%	-10.2%	-40.8%
transfer	biol2401	48	-89.6%	-70.8%	-45.8%	-39.6%
transfer	engl1301	48	-64.6%	-64.6%	-35.4%	-16.7%
transfer	phil1301	45	-66.7%	4.4%	-40.0%	4.4%
transfer	govt2305	42	23.8%	7.1%	14.3%	-14.3%

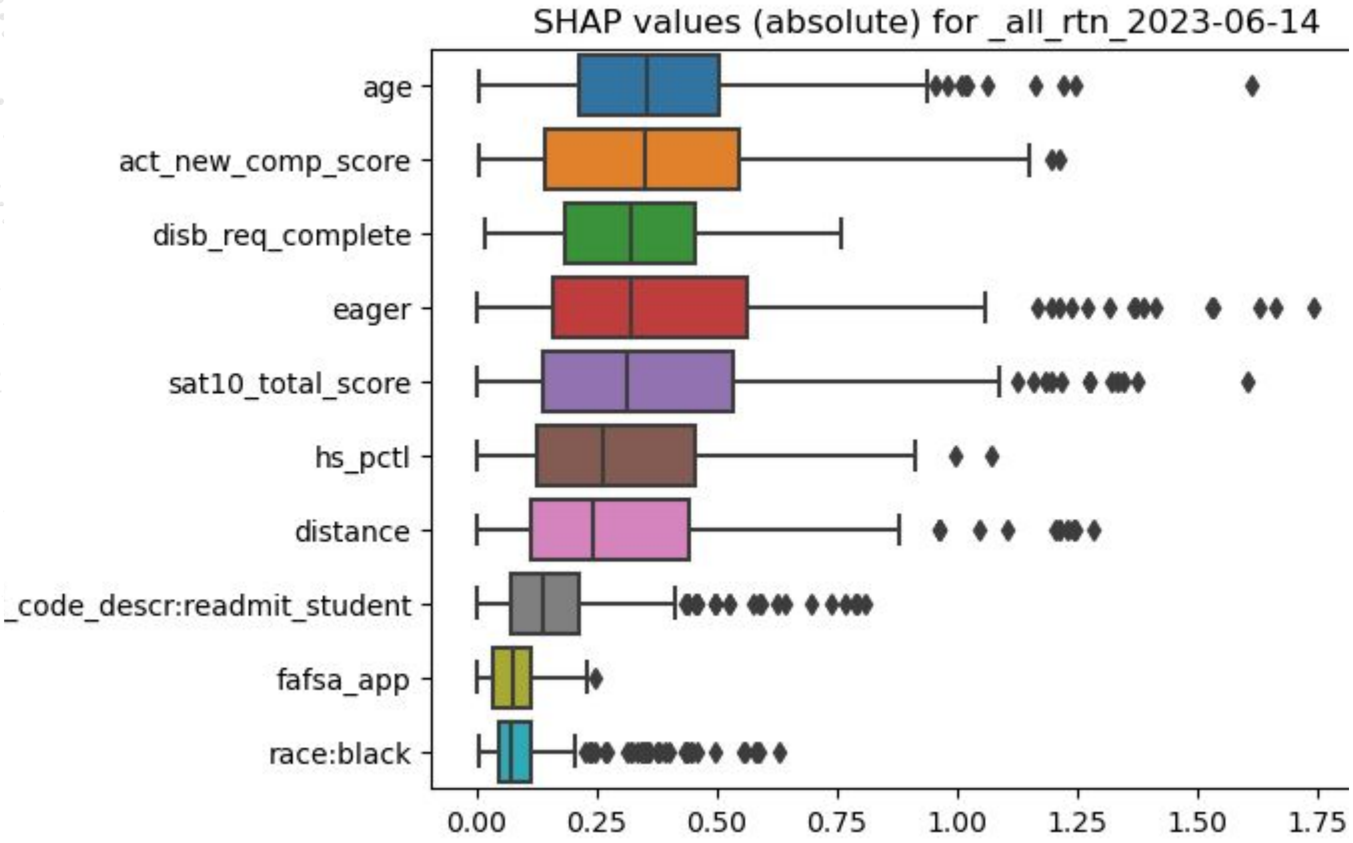
Fall 2025 Results



			AMP forecast % error			
styp_desc	crse_code	actual	April 16	May 14	June 18	July 9
returning	_headcnt	292	-49.7%	-36.6%	-32.9%	-27.1%
returning	arts1301	23	-73.9%	-69.6%	-78.3%	-69.6%
returning	math1314	19	-84.2%	-57.9%	-47.4%	-57.9%
returning	busi1301	14	-85.7%	-78.6%	-92.9%	-92.9%
returning	govt2306	12	-100.0%	8.3%	8.3%	8.3%
returning	psyc3303	12	-100.0%	-100.0%	-100.0%	-100.0%
returning	econ2301	11	-100.0%	72.7%	-18.2%	54.5%
returning	psyc3307	11	-100.0%	-36.4%	-36.4%	36.4%
returning	hist1301	10	-70.0%	-80.0%	-90.0%	-20.0%
returning	phil1301	10	-100.0%	-100.0%	-100.0%	-60.0%
returning	biol2401	9	-100.0%	-88.9%	-77.8%	-66.7%
returning	engl1301	9	-100.0%	-100.0%	-77.8%	-100.0%
returning	engl1302	9	-77.8%	-88.9%	-100.0%	11.1%
returning	math1342	8	-100.0%	-100.0%	-100.0%	-100.0%
returning	math2412	8	-100.0%	-100.0%	-75.0%	-50.0%
returning	chem1311	7	-71.4%	-14.3%	-71.4%	-14.3%
returning	comm1311	7	-100.0%	-42.9%	-57.1%	-57.1%
returning	chem1111	6	-100.0%	-100.0%	-100.0%	-16.7%
returning	govt2305	6	-100.0%	-100.0%	-100.0%	-100.0%
returning	hist1302	6	-100.0%	-33.3%	-33.3%	-33.3%
returning	univ0314	6	-50.0%	-100.0%	-33.3%	-66.7%


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
- Forecast housing demand / Housing data
- Dashboard
- Causal Machine Learning

Scan the QR code to complete the session survey.

2026