The Impact of Financial Aid and Academic Pathways on Student Success

(Predictive Analytics and Interactive Tools for Graduation Outcomes)

Course: DATA 606 Capstone in Data Science

Student: *Chaturya Yarradoddi*
Faculty Advisor: *Dr. Anupam Joshi*

UMBC

Motivation & Problem Statement

Motivation

- Universities must improve graduation outcomes while managing financial and accountability pressures.
- Student success depends on **financial aid**, **academic readiness**, and **entry pathways**, but these factors are rarely analyzed together.
- Institutions need tools that move beyond static reporting toward **predictive, actionable insights**.

Problem Statement

- Existing systems do not integrate high school data, placement tests, AP records, demographics, and multi-year financial aid into one modeling framework.
- Without an integrated predictive model, universities cannot **identify at-risk students early**, **optimize aid allocation**, or **test policy scenarios**.
- There is a need for a **unified, leakage-free, operational predictive system** that produces individualized graduation probabilities and supports institutional decision-making

Data Sources

DataSet

FA Stdnt Success *25,960 rows × 131 columns*
Complete student records: demographics, high-school metrics, placement tests, SAT scores, and six-year financial aid history.
AP Courses *58,246 rows × 13 columns*
AP exam-level data: test components, scores, credit earned, and transfer credits.

Key columns

AP Courses

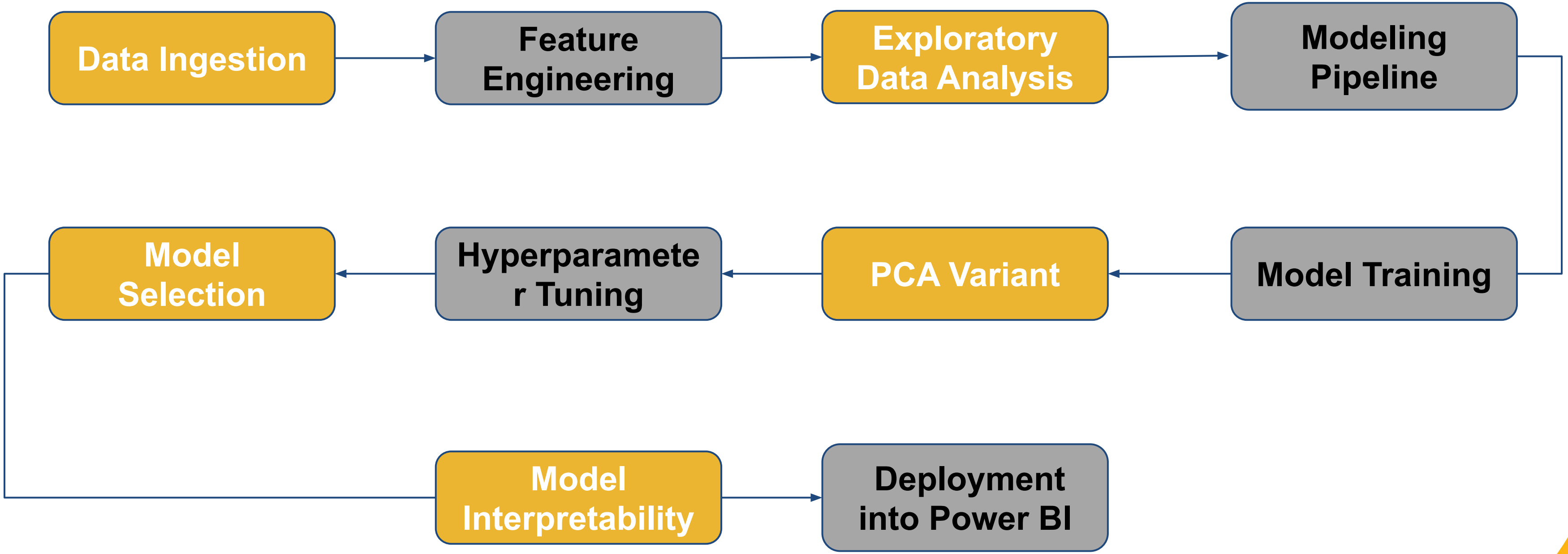
- **EmployeeID:** unique student identifier
- **TEST_ID, TEST_COMPONENT**: AP exam code
- **TestDescription**: official AP exam title
- **SCORE**: student's AP score
- **EARN_CREDIT**: whether UMBC awarded credit (Y/N)
- **UNT_TRNSFR**: number of credits transferred

FA Student Success

- **Demographics:** gender, ethnicity, residency, ZIP, housing
- **Matriculation details:** entry term, transfer/freshman status
- **High-school data:** GPA, GPA bands, percentile ranks
- **Placement scores:** ALEKS, English, Algebra, Calculus
- **SAT scores:** Math, Reading/Writing, combined scores
- **Financial aid for 6 years:** Y1–Y6 Need indicators, Grant amounts, Merit amount, Scholarship program amounts
- **Enrollment behavior:** full-time/part-time across multiple semesters
- **Degree completion fields:** years to degree, terms to degree

UMBC

Model Architecture



Feature Engineering

INITIAL CLEANING OF AP DATA

The AP dataset was cleaned by standardizing TEST_ID and TEST_COMPONENT fields, constructing a unified AP_Code, and normalizing TestDescription text. Each exam was then mapped into one of seven subject categories social, language, English, STEM, math, computer, or art/music using a standardized AP subject dictionary.

AP FEATURE ENGINEERING

We aggregated all AP rows per student to compute total tests, unique codes, AP credits, and score summaries (avg/max). Subject-level AP counts were generated for seven domains, along with composite STEM features such as AP_ct_STEM_like and AP_STEM_ratio. The final engineered AP dataset was saved as CSV and Parquet files.

FINANCIAL AID FEATURE ENGINEERING

Financial variables were created by summing six years of grant and merit amounts into TotalSupport and generating Supported/NeedStatus indicators. Students were grouped into SupportBin ranges (NoSup to >20K) to capture practical funding levels, and the engineered aid dataset was stored in CSV and Parquet formats.

MERGING AP + AID DATA

EmployeeID values were standardized across datasets and merged using a left join, ensuring all aid records remained. The merge was validated through match-rate checks to confirm proper alignment of AP features with student profiles.

Feature Engineering

HANDLING MISSING DATA

All numeric fields (scores, credits, counts, support totals) were imputed with zero, and all categorical fields were filled with “Unknown.” After cleaning, the merged dataset contained **zero missing values**, ensuring full compatibility with machine-learning pipelines

DUPLICATE HANDLING

We detected 34 duplicated student IDs mostly due to AP test repetitions or administrative inconsistencies. Duplicates were resolved by taking maximum values for numeric fields and merging unique values for categorical fields, resulting in **25,926 unique students**.

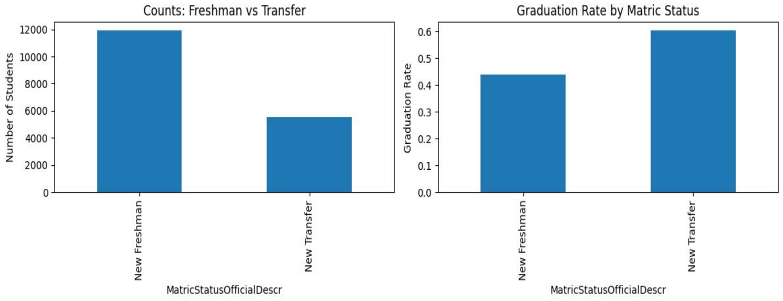
FEATURE SELECTION & LEAKAGE REMOVAL

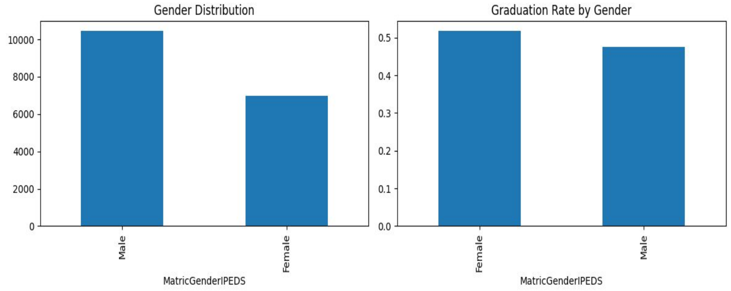
To prevent leakage, all post-enrollment, future-term, multi-year aid, and degree-progress columns were removed. Redundant indicators (HasScore flags), AP aggregates, SAT1600Score, and high-cardinality identifiers were dropped, leaving a clean schema of **31 predictive features** (20 numeric, 11 categorical).

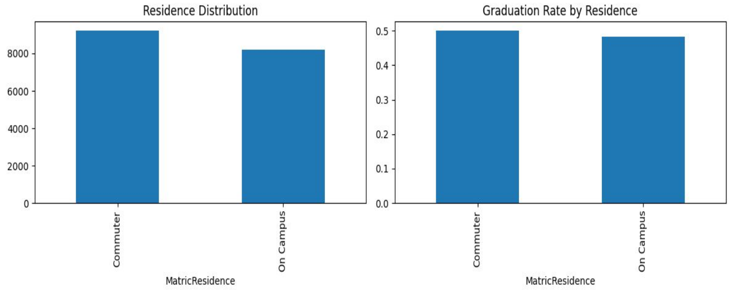
FINAL DATASET

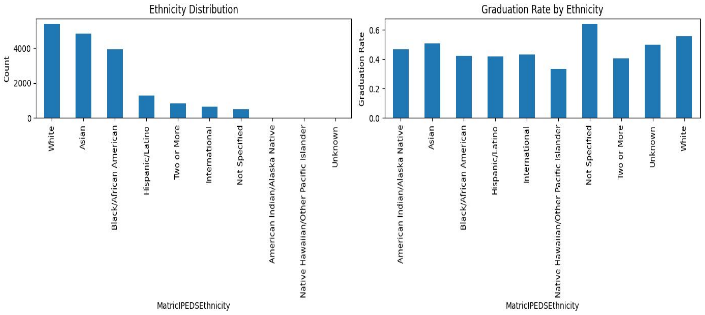
The final modeling dataset contains **25,926 students** and **31 fully engineered, leakage-free features**. These include high-school metrics, placement scores, AP subject indicators, financial support features, and matriculation characteristics forming the structured foundation for all predictive modeling.

EDA


Counts: Freshman vs Transfer
Number of Students
0
2000
4000
6000
8000
10000
12000
New Freshman
New Transfer
MatricStatusOfficialDescr
Graduation Rate by Matric Status
Graduation Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
New Freshman
New Transfer
MatricStatusOfficialDescr


Gender Distribution
0
2000
4000
6000
8000
10000
Male
Female
MatricGenderIPEDS
Graduation Rate by Gender
0.0
0.1
0.2
0.3
0.4
0.5
Female
Male
MatricGenderIPEDS


Residence Distribution
0
2000
4000
6000
8000
Commuter
On Campus
MatricResidence
Graduation Rate by Residence
0.0
0.1
0.2
0.3
0.4
0.5
Commuter
On Campus
MatricResidence


Ethnicity Distribution
Count
0
2000
4000
White
Asian
Black/African American
Hispanic/Latino
Two or More
International
Not Specified
American Indian/Alaska Native
Native Hawaiian/Other Pacific Islander
Unknown
MatricIPEDSEthnicity
Graduation Rate by Ethnicity
Graduation Rate
0.0
0.2
0.4
0.6
American Indian/Alaska Native
Asian
Black/African American
Hispanic/Latino
International
Native Hawaiian/Other Pacific Islander
Not Specified
Two or More
Unknown
White
MatricIPEDSEthnicity

EDA

- The dataset shows clear differences between freshmen and transfer students. First-time freshmen make up most of the population, but transfer students graduate at a much higher rate, reflecting stronger prior academic experience.

- Gender patterns indicate that female students consistently graduate at higher rates than male students. Although the difference is moderate, it aligns with national trends showing stronger academic persistence among women.

- Residence status has only a small effect on success. Commuter students show slightly higher graduation rates than on-campus students, suggesting that living arrangements do not significantly alter academic outcomes.

- Graduation outcomes vary across ethnic groups. White and Asian students show the highest completion rates among major categories, while Black/African American and Hispanic/Latino students show lower rates, indicating persistent equity gaps.

- The “Not Specified” ethnicity group shows unexpectedly high graduation rates because this subgroup contains many transfer students, high-aid recipients, full-time enrollees, and students with transfer credits factors that collectively drive strong outcomes.

- Financial support levels show one of the strongest relationships with graduation. Students receiving high institutional support (>$20K) graduate at far higher rates than those receiving minimal or no support, demonstrating that financial resources play a major role in degree completion.

Model Performance

The modeling dataset used 33 pre-enrollment and first-term features, including academic readiness, AP indicators, demographics, residency, enrollment status, and financial support. Graduation was encoded as 1 (Yes) and 0 (No). Data was split using a stratified 70/15/15 train–validation–test approach. A unified scikit-learn pipeline handled preprocessing, including median imputation, categorical mode imputation, z-score scaling, and one-hot encoding. Seven models were tested: Logistic Regression, Random Forest, Gradient Boosting, XGBoost, Linear SVM, RBF SVM, and Gaussian Naïve Bayes. Performance was evaluated using accuracy, recall, F1, and primarily AUC. PCA experiments were conducted and Hyperparameter tuning using RandomizedSearchCV significantly improved tree-based models. Boosting models captured complex interactions between academic and financial features.

Model	Accuracy	Precision	Recall	F1	AUC
XGBoost (Tuned)	0.7578	0.7400	0.7869	0.7627	0.8456
XGBoost (Base)	0.7555	0.7412	0.7770	0.7587	0.8454
GB (Tuned)	0.7511	0.7388	0.7687	0.7534	0.8449
GB (Base)	0.7501	0.7392	0.7646	0.7517	0.8411
RF (Tuned)	0.7503	0.7288	0.7890	0.7577	0.8388
XGBoost (PCA)	0.7490	0.7354	0.7698	0.7522	0.8348
GB (PCA)	0.7372	0.7294	0.7453	0.7373	0.8259

Model	Accuracy	Precision	Recall	F1	AUC
RF (Base)	0.7418	0.7268	0.7661	0.7460	0.8251
RF (PCA)	0.7377	0.7209	0.7666	0.7431	0.8149
LogReg (Tuned)	0.7303	0.7277	0.7266	0.7272	0.8099
LogReg (Base)	0.7308	0.7292	0.7251	0.7271	0.8096
LogReg (PCA)	0.7298	0.7270	0.7266	0.7268	0.8075
GaussianNB (PCA)	0.4983	0.4964	0.9693	0.6566	0.5280
GaussianNB (Base)	0.5014	0.4980	0.9756	0.6594	0.5118

Model Performance

- Model evaluation used **Accuracy, Precision, Recall, F1 Score, and AUC** on the test dataset.
- **Tuned XGBoost** achieved the **best performance** with **AUC = 0.8456**, **Accuracy = 0.7578**, and the **highest F1 score**.
- **XGBoost Base** and **Gradient Boosting Tuned** also performed strongly with AUC values **above 0.844**.
- **Random Forest** models performed moderately well but were slightly weaker than boosting methods.
- **Logistic Regression** models remained stable with **AUC ≈ 0.81** but underperformed compared to tree-based models.
- **PCA-based models performed worse**, showing reduced accuracy and AUC due to loss of feature information.
- **Gaussian Naïve Bayes** performed the worst (**AUC ≈ 0.51**) because it cannot model complex feature interactions.
- Boosting methods (especially **Tuned XGBoost**) provided the **highest predictive power** and were selected for further interpretability and Power BI deployment.

Feature Importance

Feature	XGB Base	XGB Tuned	GB Tuned
ALEKSScore	0.051	0.047	0.218
AP_ct_computer	0.022	0.020	0.017
AP_ct_math	–	–	0.006
AP_ct_social	–	–	0.005
AP_max_score	0.019	0.019	0.022
AP_total_transfer_credits	–	0.016	0.008
AlgSCORE	0.016	0.016	0.011
CalScore	–	–	0.010
EngSCORE	0.115	0.149	0.171
HS_Percentile_Unknown	0.019	0.015	–
HS_GPA_Band 2.5–2.99	0.016	–	–
HS_GPA_Band Unknown	0.017	0.017	–
Gender: Male	–	–	0.005

Feature	XGB Base	XGB Tuned	GB Tuned
Ethnicity: Black/African American	0.018	0.014	0.007
Ethnicity: Hispanic/Latino	0.016	–	–
Residency: Out of State	0.016	–	0.006
Status: New Transfer	0.053	0.062	0.043
SATMathScore	0.018	0.019	0.033
SATReadingWriting	0.021	0.021	0.051
Sem1_FTPT_PT	0.025	0.021	0.013
SupportBin 15K–20K	–	0.019	–
SupportBin 5K–10K	0.029	0.027	–
SupportBin <5K	0.042	0.034	0.014
SupportBin >20K	0.061	0.099	–
Supported Yes	0.018	0.017	–
TotalSupport	0.030	0.026	0.235

Feature Importance

- All three models identify the same core drivers of graduation. Placement test performance especially **EngSCORE**, followed by **ALEKS**, **SAT Reading/Writing**, and **SAT Math** shows the strongest and most stable relationship with student success.

- Financial support emerges as one of the most influential predictors. Students receiving **more than $20,000** in institutional aid consistently show higher persistence, confirming that **SupportBin_>20K** and **TotalSupport** heavily contribute to model predictions.

- **Transfer status** is a major structural factor across all models. Students entering as transfer students display significantly higher predicted graduation probabilities compared with new freshmen.

- High school preparation, including **GPA**, **AP coursework**, and **class rank**, provides moderate but meaningful predictive power. These variables reinforce long-term academic readiness.

- First-term enrollment intensity (**Sem1_FTPT**) adds smaller but consistent value, with full-time students generally achieving stronger outcomes.

- Overall, the models agree that **academic readiness + financial stability + entry pathway** form the three pillars that shape graduation predictions.

Categorical Gap Analysis

This table shows how different student groups are expected to graduate based on the model’s predictions. Each student’s predicted graduation probability is generated by the model, and students are then grouped by characteristics such as gender, ethnicity, support level, GPA, and enrollment type. For each group, the table reports the average predicted graduation rate .

Feature	Category	XGB Base	XGB Tuned	GB Tuned
Term Load	FT	3591	0.499	0.498
	PT	298	0.424	0.428
Residence	On Campus	1653	0.505	0.503
	Commuter	2236	0.485	0.486
Need Status	Need Yes	3146	0.503	0.504
	Need No	743	0.449	0.444

Categorical Gap Analysis

Feature	Category	XGB Base	XGB Tuned	GB Tuned
Matric Status	New Transfer	0.548	0.543	0.548
	New Freshman	0.458	0.460	0.456
Gender	Female	0.523	0.519	0.520
	Male	0.476	0.477	0.476
Ethnicity	American Indian/Alaska Native	0.660	0.641	0.612
	Not Specified	0.600	0.589	0.606
	White	0.536	0.536	0.537
	Asian	0.510	0.505	0.510
	Black/African American	0.428	0.432	0.425
	Hispanic/Latino	0.417	0.428	0.416
HS GPA Band	No GPA	0.641	0.559	0.623
	3.5–3.99	0.477	0.474	0.475
	4.0–4.49	0.467	0.466	0.464
	2.5–2.99	0.462	0.463	0.461
HS Percentile	91–100%	0.633	0.629	0.621
	81–90%	0.536	0.526	0.528
	Unknown	0.482	0.482	0.482
SupportBin	>20K	0.682	0.681	0.679
	15K–20K	0.587	0.575	0.586
	10K–15K	0.406	0.407	0.403
	5K–10K	0.312	0.320	0.314
	<5K	0.258	0.266	0.264

Categorical Gap Analysis

Transfer students consistently outperform first-time freshmen, and all three models (XGBoost Base, XGBoost Tuned, Gradient Boosting Tuned) confirm this pattern with high stability.

Female students show slightly higher graduation probabilities than male students, a small but consistent advantage across all models.

Ethnicity patterns remain stable:

- Highest predicted outcomes: *American Indian/Alaska Native* and *Not Specified* groups.
- Lower predicted outcomes: *Black/African American* and *Hispanic/Latino* groups.

Full-time enrollment in the first term strongly improves graduation predictions, while part-time students consistently receive lower predictions.

Residence has a mild effect on-campus students perform slightly better than commuters, but the difference is small.

High school percentile matters: students in the **91–100% range** achieve the strongest predictions, while lower ranges and missing percentile data yield weaker outcomes.

Financial support shows the largest categorical impact:

- Students receiving **> $20K** institutional aid reach ~0.68 predicted success.
- Students receiving **< $5K** drop to ~0.26–0.32.
- This confirms that **institutional aid is one of the strongest predictors of persistence**.

All three models produce nearly identical categorical structures, demonstrating that these relationships are robust, consistent, and model-independent.

Positive Policy Effects

Policy Change	XGBoost Base	XGBoost Tuned	GB Tuned
Force = New Transfer	**+7.62**	**+5.90**	**+8.97**
Ethnicity = International	+4.84	+2.97	+4.62
Ethnicity = Not Specified	+4.16	+2.74	+4.46
NeedStatus = No	+4.14	+2.86	+0.42
HighSchoolGPA +20%	+2.49	+1.71	+1.98
HighSchoolGPA +10%	+2.05	+1.38	+1.65
SupportBin: 15K–20K	+2.64	+2.37	+1.03
SupportBin: >20K	+2.51	+5.42	+0.96
TotalSupport +20%	+1.43	+1.18	+1.59
Matric Gender = Female	+1.92	+0.00–0.27	+1.90

Negative Policy Effects

Policy Change	XGBoost Base	XGBoost Tuned	GB Tuned
Force = New Freshman	**–8.16**	**–7.82**	**–9.32**
Ethnicity = Hispanic/Latino	–5.25	–4.39	–5.54
Ethnicity = Black/African American	–3.98	–3.39	–3.98
Out of State Resident	–5.44	–4.27	–5.18
SupportBin <5K	–2.20	–3.32	–0.16
HighSchoolGPA –20%	–3.03	–2.61	–2.30
TotalSupport –20%	–2.06	–1.77	–2.30
SAT/Eng/ALEKS –20%	–1.3 to –1.7	–1.2 to –1.7	–1.5 to –2.0

Policy Simulation

Transfer status produces the strongest positive policy effect across all three models. Transfer students consistently receive higher predicted graduation probabilities due to prior credits, clearer degree pathways, and demonstrated academic readiness.

Ethnicity effects remain stable and structural, not model-dependent:

- *International* and *Not Specified* groups show higher predicted outcomes.
- *Hispanic/Latino* and *Black/African American* groups show lower outcomes, indicating systemic barriers rather than model bias.

Financial support shows the most policy-actionable impact:

- Students receiving **> $20K** in aid (SupportBin or TotalSupport) see the largest gains in predicted success.
- Lower-support categories (<$10K) show consistently weaker outcomes.

Academic preparation variables add moderate positive influence, with higher GPA, strong placement scores, and AP performance slightly raising predictions across all models.

Negative shifts appear when students are classified as New Freshmen, or when academic readiness and financial aid levels decline this pattern is identical in all models.

Across all models, positive and negative policy effects remain consistent, confirming that the drivers of student success are stable and not dependent on a specific algorithm.

Partial Dependence

All three models show identical partial-dependence patterns, confirming that their predictions respond to features in the same way.

English readiness (EngSCORE) has the strongest nonlinear effect: graduation probability rises sharply once students move out of the lowest readiness band, then levels off at higher scores.

Financial support shows large positive effects:

- Graduation probability increases steeply as funding rises.
- **TotalSupport** grows steadily until mid-range levels, after which gains flatten, indicating diminishing returns.

Math preparation (ALEKS, SAT Math) produces smooth upward trends, showing that stronger quantitative readiness consistently improves graduation predictions.

High-school readiness metrics (GPA, percentile rank) show gradual, stable improvements across their ranges.

Matriculation status behaves like a binary jump: New Transfer students show significantly higher predicted outcomes than New Freshmen across all models.

AP patterns are mostly stable, with AP_max_score showing a positive slope, while the AP_ct_computer dip appears to result from small, specialized subgroups rather than true negative relationships.

Overall, the PDP curves reinforce the key model drivers: financial support, transfer status, English readiness, and math preparation consistently shape graduation predictions in all three algorithms.

Power Bi

Power BI

GitHub

https://github.com/ChaturyaYarradoddi/UMBC_CP/tree/main

Questions | Comments?

UMBC

THANK YOU