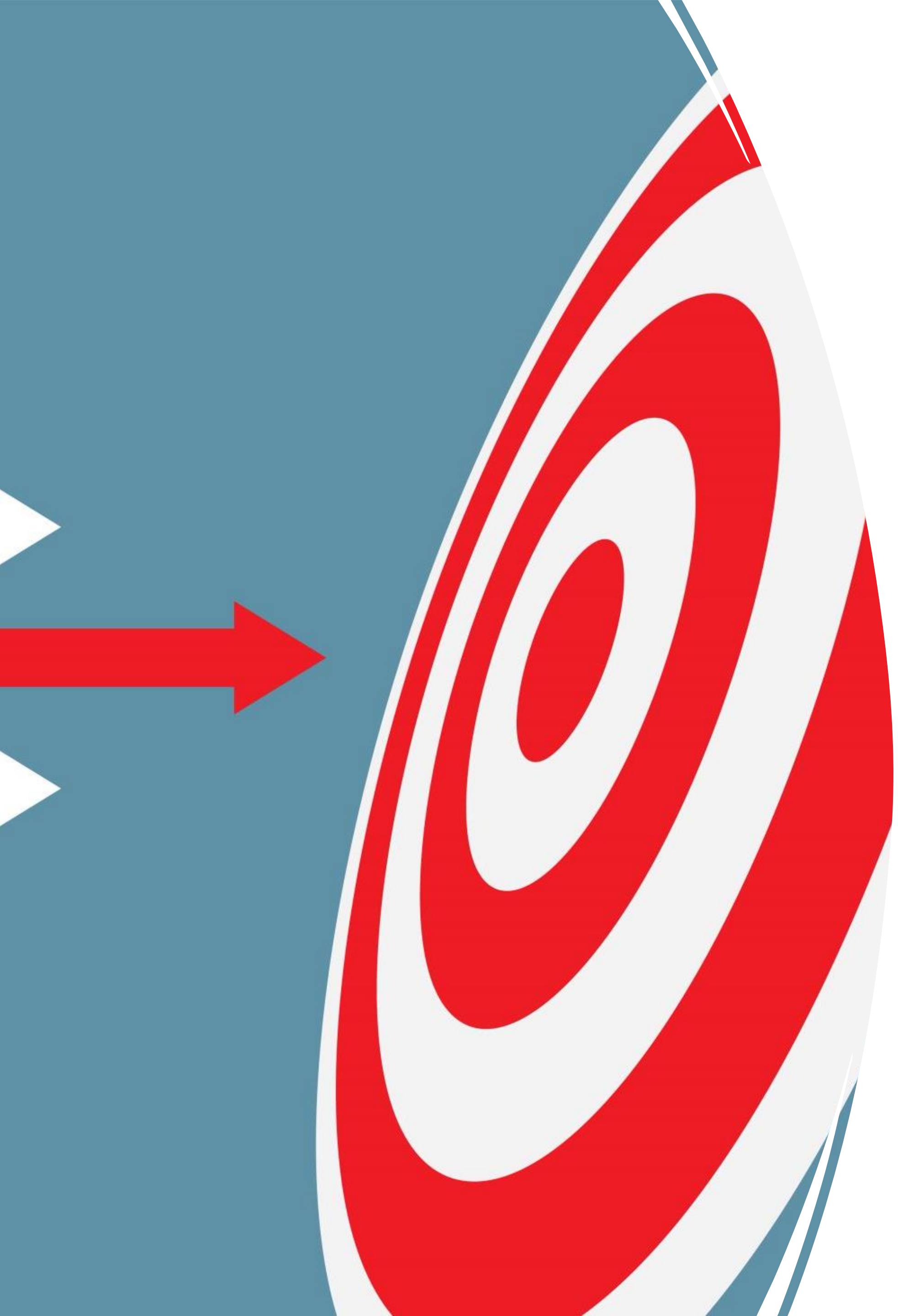


AI for Student Success: Targeting Dropout Risk Before It Happens

From Silent Attrition to Smart Prevention

A Mid Point Presentation

By Denis & Dhruv

The Case

- **Every 1% increase in retention = $2–3M in retained tuition**
- SNHU's first-year retention: **61%**
- 6-year graduation rate: **39%**
- Thousands of students lost silently each year
- AI identifies at-risk students before disengagement escalates
- Source – College Navigator by NCES.GOV

The Background

- SNHU leads in access and innovation — but retention remains a blind spot.
- 1 in 4 U.S. students drops out before sophomore year — often without warning.
- Each dropout means lost tuition, lost trust, and lost mission impact.
- At SNHU, thousands of LMS signals go unused — a silent crisis in plain sight.
- Traditional support reacts too late — we need radar, not rearview mirrors.

The Problem

- **Silent Dropout Crisis = Silent Revenue Loss**
- Every student lost = lost revenue, momentum, and mission alignment.
- SNHU’s LMS captures 100K+ behavioral signals per week — but risk patterns go unnoticed.
- Faculty can't scale 1:1 support — struggling students go unseen.
- Dropouts lead to lost revenue ($10K–$25K/student), lower graduation rates, and decreased rankings.
- Traditional models act **too late** — only after academic damage is done.

Business & Human Impact of Student Failure

Problem Beyond Academics

- When students fail or drop out, it is not only a *grade issue* — it becomes a human problem.
- Professors may experience stress, frustration, burnout when many students underperform.
- Negative emotional climate in class can impact other students, reducing motivation and engagement.
- High failure/dropout rate creates a negative reputation for the university and affects business decisions.
- Professors face increasing pressure to maintain academic standards.
- Students who struggle often avoid communication, leading to ineffective teaching-feedback loops.

Opportunity for Innovation

- Predict at-risk students early using LMS behavior patterns.
- Empower advisors with real-time alerts to act proactively.
- Model is ethical, scalable, and built for SNHU's systems.
- Proactive > Reactive: intervene before students disengage.
- **Why Now?**

We already collect the data — we just haven't been using it for retention strategy.

The Proposed Solution

- Predictive models:
- Classification: Pass/Fail likelihood
- Regression: Final grade estimation **AI-Powered Early Warning System for At-Risk Students**
- Built on student activity, demographic, and course performance data
- Transparent and explainable — shows *why* risk is triggered
- Weekly scans + advisor dashboards

Project Objectives

- Build two predictive models using LMS + demographic risk signals. Evaluate model accuracy, interpretability, and bias
- Recommend data-driven intervention paths
- Support scalable student success — **at every level of performance**

Strategic & Competitive Edge

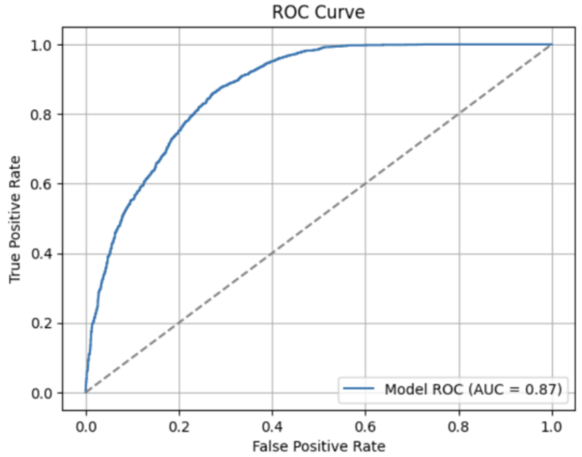
- Retaining 100 students preserves $1.5M+ in tuition revenue.
- Avoid re-enrollment & refund losses.
- Boosts rankings, satisfaction, and compliance outcomes.
- Fully in-house, explainable, and tailored to SNHU’s advising ecosystem.
- Strengthens SNHU’s mission: accessible, supported learning at scale.

Predicting Student Risk and Performance

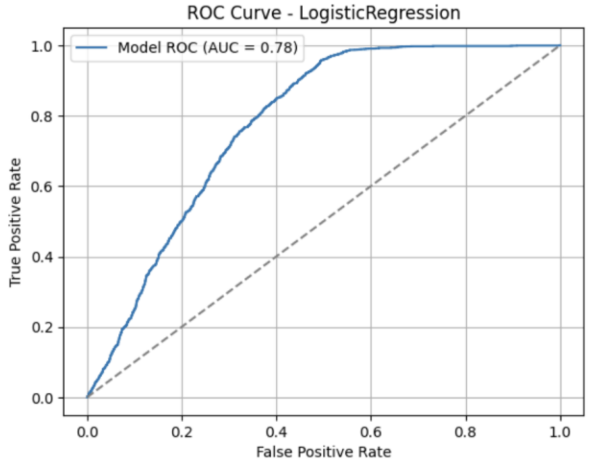
- Beyond pass/fail, we wanted to **predict exact assessment scores**.
- This helps personalize interventions and monitor student progress.
- Used **OLS for interpretability** and **Linear Regression for prediction**.

Pass/Fail Model Performance

Accuracy: 0.7871556072953046
ROC AUC: 0.8719236128636512



Accuracy: 0.7050834303453628
ROC AUC: 0.7773847056678457



Random Forest:

•Accuracy: 78.7%

•AUC: 0.872

- **Logistic Regression:**
- Accuracy: 70.5%
- AUC: 0.777

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     avg_assessment_score   R-squared:                       0.157
Model:                              OLS   Adj. R-squared:                  0.157
Method:                   Least Squares   F-statistic:                     960.6
Date:                  Tue, 18 Nov 2025   Prob (F-statistic):               0.00
Time:                          15:58:29   Log-Likelihood:            -1.0640e+05
No. Observations:                 25770   AIC:                         2.128e+05
Df Residuals:                     25764   BIC:                         2.129e+05
Df Model:                             5
Covariance Type:              nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                    64.1152      0.275    232.929      0.000      63.576      64.655
assessments_attempted     0.9119      0.029     31.744      0.000       0.856       0.968
total_clicks              0.0002   9.11e-05      1.843      0.065   -1.07e-05       0.000
active_days               0.0682      0.003     21.700      0.000       0.062       0.074
studied_credits          -0.0303      0.002    -12.173      0.000      -0.035      -0.025
num_of_prev_attempts     -0.9077      0.206     -4.413      0.000      -1.311      -0.505
==============================================================================
Omnibus:                     3669.467   Durbin-Watson:                   1.770
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6843.894
Skew:                          -0.911   Prob(JB):                         0.00
Kurtosis:                       4.747   Cond. No.                     6.90e+03
==============================================================================
```

Focus on Numerical Predictors

- assessments_attempted, active_days, and studied_credits showed strong significance
- Allowed us to understand direct behavioral drivers of success

Statistical Significance – P-Values

num_of_prev_attempts	-0.2523	0.197	-1.283	0.199	-0.638	0.133
studied_credits	-0.0043	0.003	-1.669	0.095	-0.009	0.001
assessments_attempted	0.9261	0.036	25.831	0.000	0.856	0.996
total_clicks	-0.0004	9.85e-05	-3.914	0.000	-0.001	-0.000
active_days	0.0887	0.003	26.390	0.000	0.082	0.095
target_pass	-0.6922	0.213	-3.250	0.001	-1.110	-0.275
gender_M	-0.0703	0.224	-0.314	0.753	-0.509	0.368
region_East Midlands Region	0.0099	0.432	0.023	0.982	-0.838	0.858
region_Ireland	-0.2042	0.540	-0.378	0.705	-1.262	0.854
region_London Region	-0.7450	0.408	-1.825	0.068	-1.545	0.055
region_North Region	-0.3458	0.534	-0.647	0.517	-1.393	0.701
region_North Western Region	-0.0906	0.420	-0.216	0.829	-0.914	0.732
region_Scotland	1.0474	0.393	2.665	0.008	0.277	1.818
region_South East Region	1.2263	0.442	2.776	0.006	0.360	2.092
region_South Region	-0.1118	0.397	-0.282	0.778	-0.889	0.666
region_South West Region	0.0987	0.424	0.233	0.816	-0.732	0.929
region_Wales	0.2294	0.443	0.517	0.605	-0.640	1.098
region_West Midlands Region	0.5078	0.428	1.186	0.236	-0.332	1.347
region_Yorkshire Region	-0.4248	0.459	-0.926	0.354	-1.324	0.474
highest_education_HE Qualification	1.0357	0.273	3.787	0.000	0.500	1.572
highest_education_Lower Than A Level	-2.6578	0.199	-13.353	0.000	-3.048	-2.268
highest_education_No Formal quals	-6.5049	0.954	-6.817	0.000	-8.375	-4.634
highest_education_Post Graduate Qualification	6.2992	0.893	7.057	0.000	4.550	8.049
imd_band_10-20	0.1829	0.404	0.453	0.651	-0.609	0.974
imd_band_20-30%	1.5264	0.400	3.821	0.000	0.743	2.309
imd_band_30-40%	1.8957	0.402	4.720	0.000	1.108	2.683
imd_band_40-50%	2.5975	0.411	6.323	0.000	1.792	3.403
imd_band_50-60%	2.2781	0.412	5.529	0.000	1.471	3.086
imd_band_60-70%	2.5808	0.420	6.141	0.000	1.757	3.405
imd_band_70-80%	2.7386	0.421	6.499	0.000	1.913	3.564
imd_band_80-90%	3.9175	0.430	9.120	0.000	3.076	4.760
imd_band_90-100%	4.0899	0.445	9.199	0.000	3.218	4.961
imd_band_?	4.8391	0.636	7.610	0.000	3.593	6.085
age_band_35-55	0.1870	0.206	0.909	0.363	-0.216	0.590
age_band_55<=	1.9892	1.069	1.861	0.063	-0.106	4.084
disability_Y	-1.2177	0.309	-3.946	0.000	-1.822	-0.613
module_BBB	7.9591	0.609	13.080	0.000	6.766	9.152
module_CCC	0.8667	0.618	1.402	0.161	-0.345	2.079
module_DDD	-0.1212	0.591	-0.205	0.837	-1.280	1.037
module_EEE	14.7225	0.627	23.483	0.000	13.494	15.951
module_FFF	5.8541	0.614	9.538	0.000	4.651	7.057
module_GGG	13.2315	0.676	19.560	0.000	11.906	14.557
presentation_2013J	1.1926	0.298	3.998	0.000	0.608	1.777
presentation_2014B	0.5349	0.314	1.706	0.088	-0.080	1.150
presentation_2014J	0.8367	0.300	2.787	0.005	0.248	1.425

- Only variables with **$p < 0.05$** retained for modeling
- P-value filtering helped clean and reduce overfitting

Feature Groups Used

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     avg_assessment_score   R-squared:                       0.253
Model:                              OLS   Adj. R-squared:                  0.252
Method:                   Least Squares   F-statistic:                     189.5
Date:                  Tue, 18 Nov 2025   Prob (F-statistic):               0.00
Time:                          16:48:38   Log-Likelihood:            -1.0485e+05
No. Observations:                 25770   AIC:                         2.098e+05
Df Residuals:                     25723   BIC:                         2.102e+05
Df Model:                            46
Covariance Type:              nonrobust
------------------------------------------------------------------------------
```

•Combined three main types:

•Numerical: Clicks, active days, assessments

•OLS analysis

•Categorical: Education level, modules

•Socio-demographic: Region, IMD band, disability

•Feature engineering informed by

OLS Results – Initial Stats

	coef	std err	t	P>|t|	[0.025	0.975]
const	54.9728	0.303	181.572	0.000	54.379	55.566
assessments_attempted	0.8846	0.032	27.997	0.000	0.823	0.947
active_days	0.0797	0.002	37.754	0.000	0.076	0.084
imd_band_30-40%	1.0207	0.310	3.288	0.001	0.412	1.629
imd_band_40-50%	1.7809	0.321	5.555	0.000	1.153	2.409
imd_band_50-60%	1.5032	0.321	4.683	0.000	0.874	2.132
imd_band_60-70%	1.8486	0.329	5.613	0.000	1.203	2.494
imd_band_70-80%	1.9956	0.330	6.046	0.000	1.349	2.643
imd_band_80-90%	3.1833	0.337	9.456	0.000	2.523	3.843
imd_band_90-100%	3.3204	0.345	9.613	0.000	2.643	3.997
highest_education_Post Graduate Qualification	8.4746	0.869	9.758	0.000	6.772	10.177
highest_education_HE Qualification	1.6717	0.263	6.357	0.000	1.156	2.187
highest_education_Lower Than A Level	-2.4741	0.198	-12.516	0.000	-2.861	-2.087
disability_Y	-1.1874	0.307	-3.869	0.000	-1.789	-0.586
module_GGG	12.8067	0.367	34.892	0.000	12.087	13.526
module_BBB	7.3949	0.251	29.434	0.000	6.902	7.887
module_EEE	14.2372	0.344	41.417	0.000	13.563	14.911
module_FFF	5.0158	0.249	20.179	0.000	4.529	5.503
presentation_2013J	0.7800	0.225	3.473	0.001	0.340	1.220
presentation_2014J	0.5467	0.213	2.562	0.010	0.129	0.965

OLS Regression Results

Dep. Variable:	avg_assessment_score	R-squared:	0.246
Model:	OLS	Adj. R-squared:	0.246
Method:	Least Squares	F-statistic:	443.0
Date:	Tue, 18 Nov 2025	Prob (F-statistic):	0.00
Time:	19:51:17	Log-Likelihood:	-1.0496e+05
No. Observations:	25770	AIC:	2.100e+05
Df Residuals:	25750	BIC:	2.101e+05
Df Model:	19		
Covariance Type:	nonrobust		

- interpreted feature importance via OLS regression
- **Key negative predictors**: disability_Y, lower education level

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     avg_assessment_score   R-squared:                       0.246
Model:                              OLS   Adj. R-squared:                  0.246
Method:                   Least Squares   F-statistic:                     443.0
Date:                  Tue, 18 Nov 2025   Prob (F-statistic):               0.00
Time:                          19:51:17   Log-Likelihood:            -1.0496e+05
No. Observations:                 25770   AIC:                         2.100e+05
Df Residuals:                     25750   BIC:                         2.101e+05
Df Model:                            19
Covariance Type:              nonrobust
==================================================================================================================
                                                coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------
const                                        54.9728      0.303    181.572      0.000      54.379      55.566
assessments_attempted                         0.8846      0.032     27.997      0.000       0.823       0.947
active_days                                   0.0797      0.002     37.754      0.000       0.076       0.084
imd_band_30-40%                               1.0207      0.310      3.288      0.001       0.412       1.629
imd_band_40-50%                               1.7809      0.321      5.555      0.000       1.153       2.409
imd_band_50-60%                               1.5032      0.321      4.683      0.000       0.874       2.132
imd_band_60-70%                               1.8486      0.329      5.613      0.000       1.203       2.494
imd_band_70-80%                               1.9956      0.330      6.046      0.000       1.349       2.643
imd_band_80-90%                               3.1833      0.337      9.456      0.000       2.523       3.843
imd_band_90-100%                              3.3204      0.345      9.613      0.000       2.643       3.997
highest_education_Post Graduate Qualification 8.4746      0.869      9.758      0.000       6.772      10.177
highest_education_HE Qualification            1.6717      0.263      6.357      0.000       1.156       2.187
highest_education_Lower Than A Level         -2.4741      0.198    -12.516      0.000      -2.861      -2.087
disability_Y                                 -1.1874      0.307     -3.869      0.000      -1.789      -0.586
module_GGG                                   12.8067      0.367     34.892      0.000      12.087      13.526
module_BBB                                    7.3949      0.251     29.434      0.000       6.902       7.887
module_EEE                                   14.2372      0.344     41.417      0.000      13.563      14.911
module_FFF                                    5.0158      0.249     20.179      0.000       4.529       5.503
presentation_2013J                            0.7800      0.225      3.473      0.001       0.340       1.220
presentation_2014J                            0.5467      0.213      2.562      0.010       0.129       0.965
==============================================================================
Omnibus:                     3344.182   Durbin-Watson:                   1.932
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7009.327
Skew:                          -0.801   Prob(JB):                         0.00
Kurtosis:                       4.991   Cond. No.                         862.
==============================================================================
```

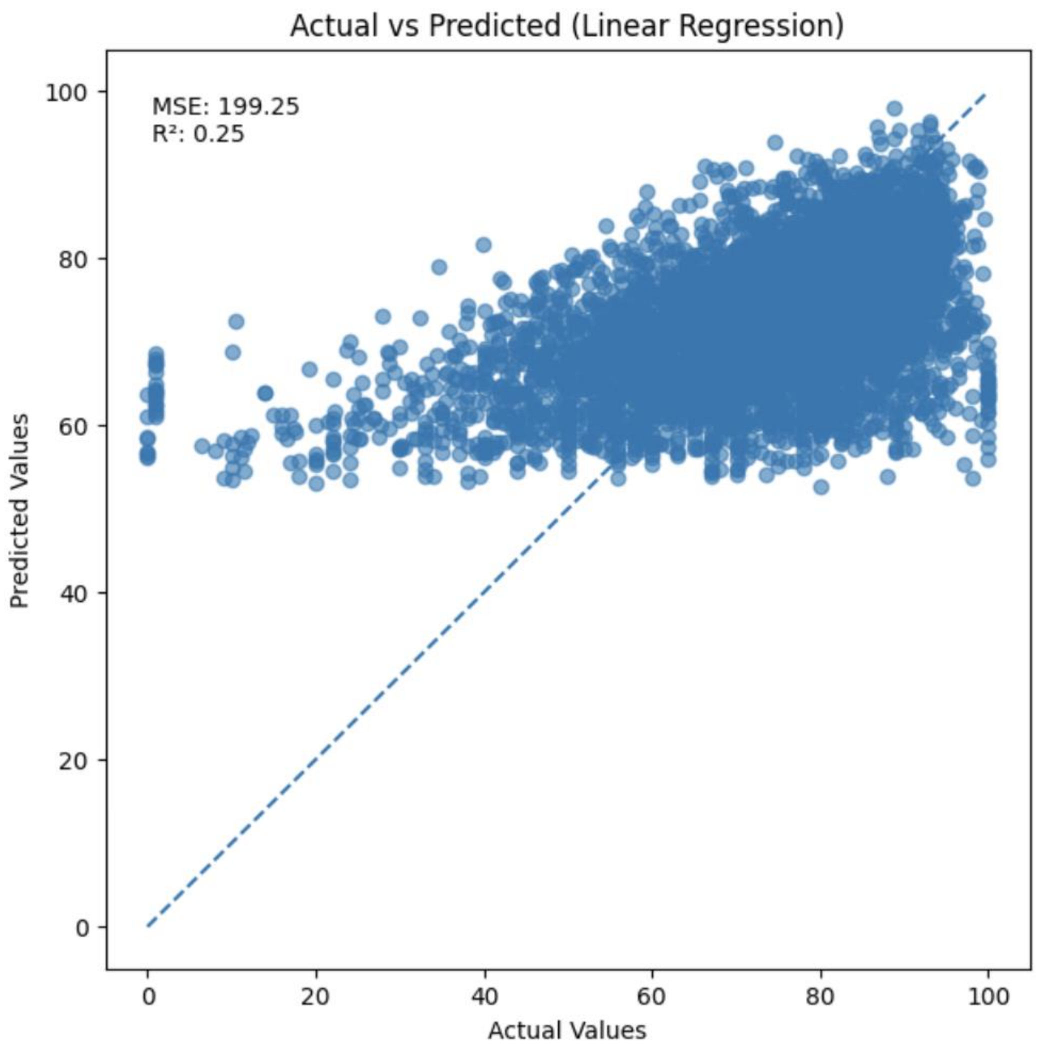
Feature Refinement Process

- Refined feature set after removing low-significance variables
- Focused on predictors with low p-values ($p < 0.05$)
- Improved model generalization and reduced noise

OLS Summary Table

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     avg_assessment_score   R-squared:                       0.246
Model:                              OLS   Adj. R-squared:                  0.246
Method:                   Least Squares   F-statistic:                     443.0
Date:                  Tue, 18 Nov 2025   Prob (F-statistic):               0.00
Time:                          19:51:17   Log-Likelihood:            -1.0496e+05
No. Observations:                 25770   AIC:                         2.100e+05
Df Residuals:                     25750   BIC:                         2.101e+05
Df Model:                            19
Covariance Type:              nonrobust
```

- Final model shows R^2 = **0.246**
- Indicates moderate explanatory power



Model Performance – Linear Regression

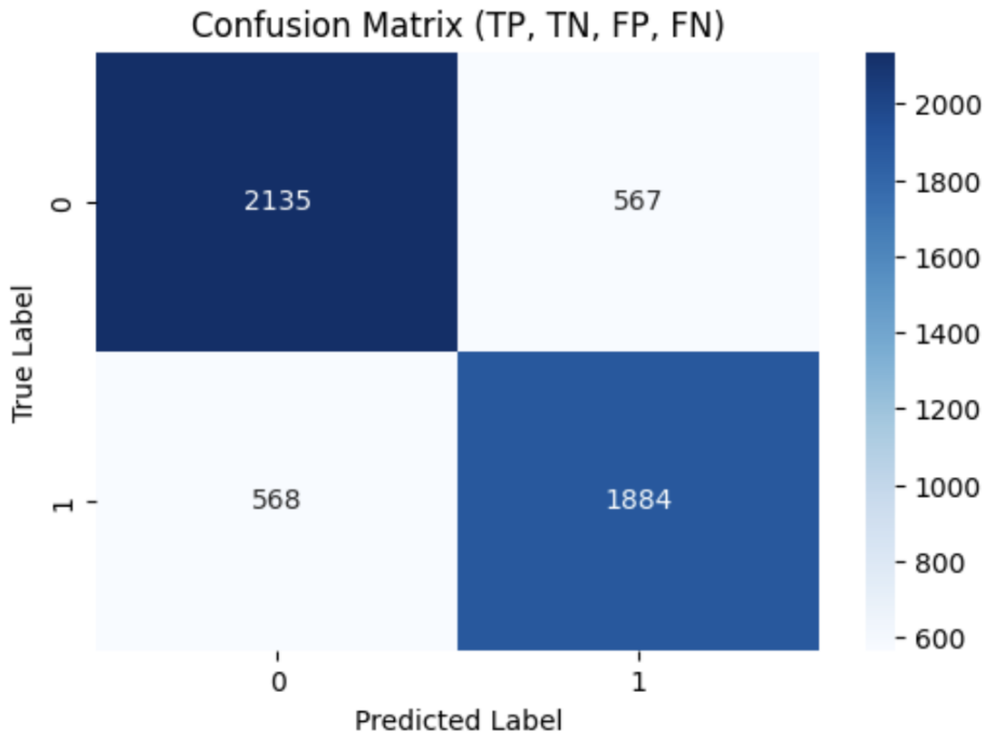
- **R^2 = 0.25, MSE = 199.25**

Where We Got Stuck — and How to Move Forward

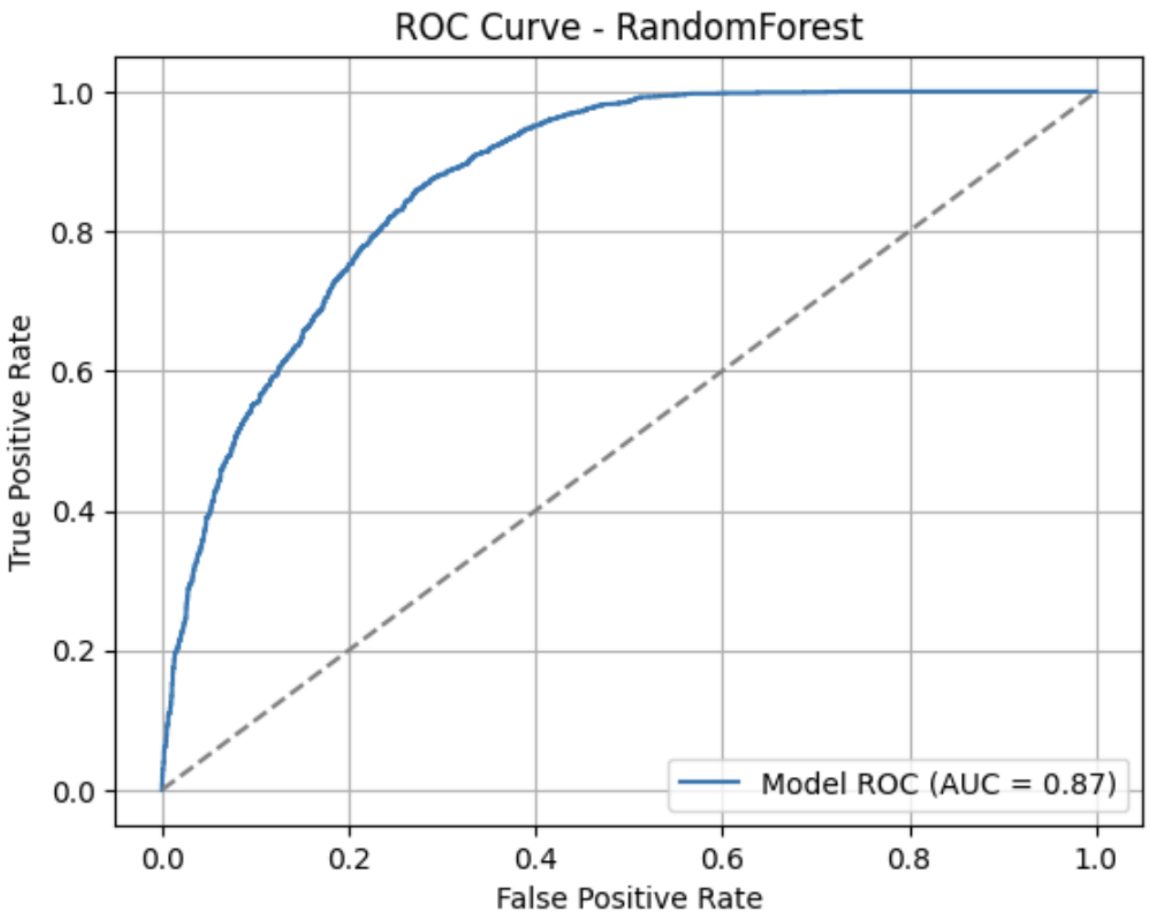
- What We Achieved
  - Built accurate model to predict Pass/Fail
  - Early risk alerts for struggling students
  - Helps advisors focus support faster

- What Limited Us
  - Score prediction model (Linear Regression) had low accuracy
  - Could explain only 25% of variation in scores ($R^2 = 0.25$)
  - Missing key behavioral data — shallow feature set

Pass/Fail Model Performance



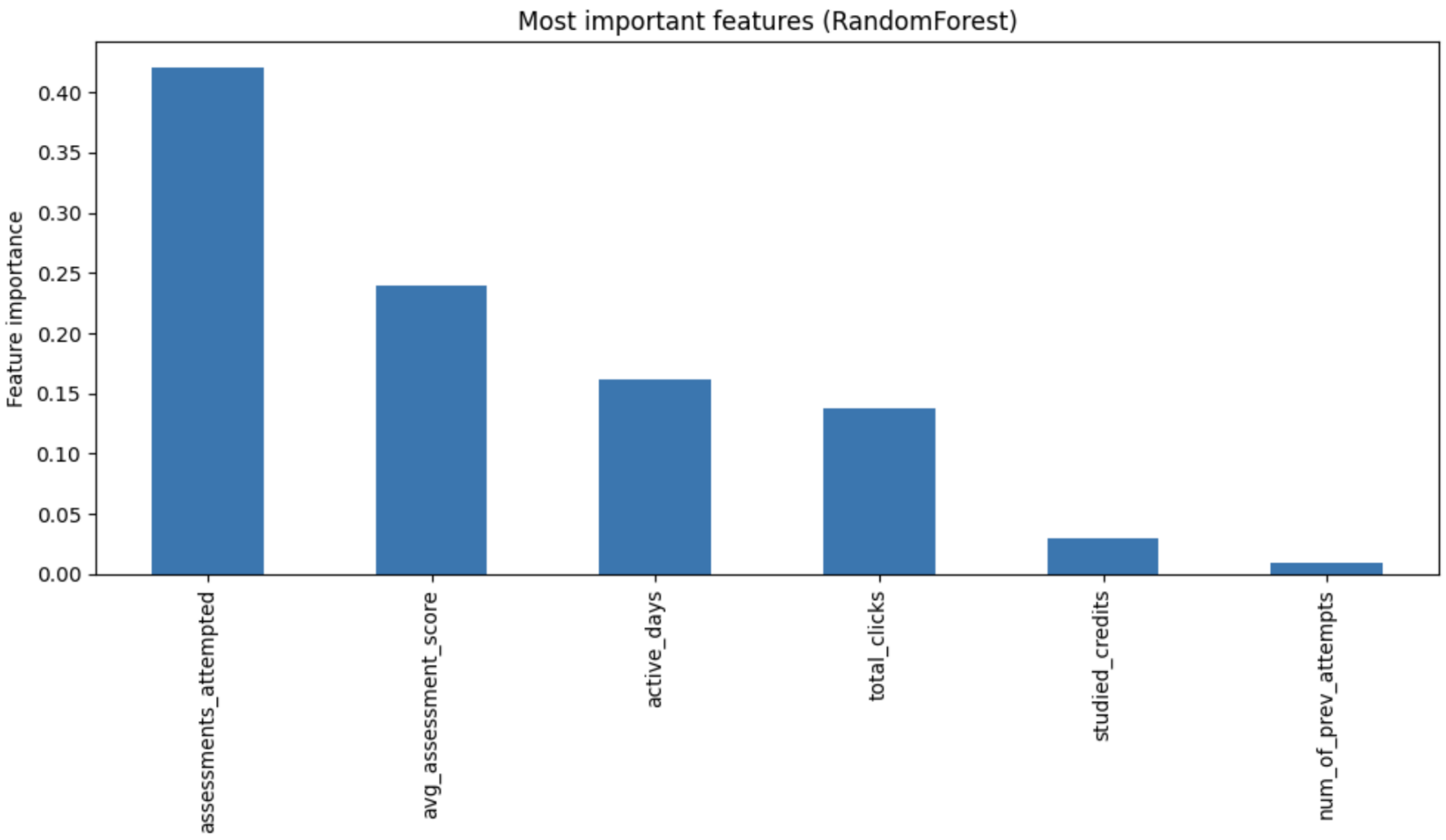

```
Accuracy: 0.7871556072953046
ROC AUC: 0.8719236128636512
```



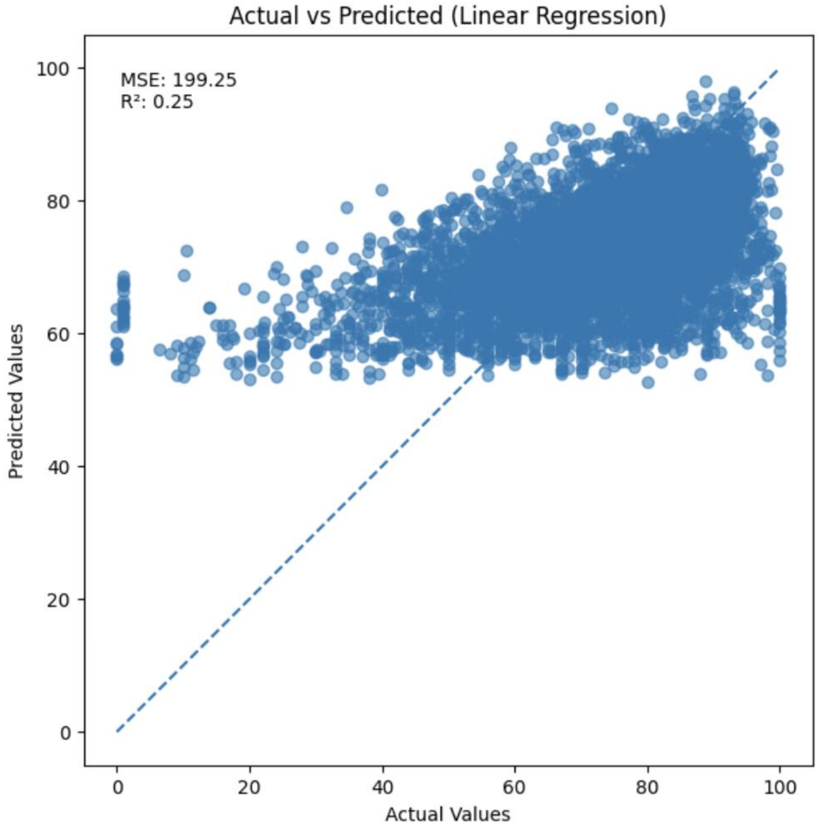
Key Predictors of Student Success

Using the original dataset and raw behavioral features, the classification model demonstrates solid performance.

The Random Forest classifier achieves reasonable accuracy and maintains a good balance between false positives and false negatives, making it suitable for early risk detection.



Baseline Model Performance Using Raw Features

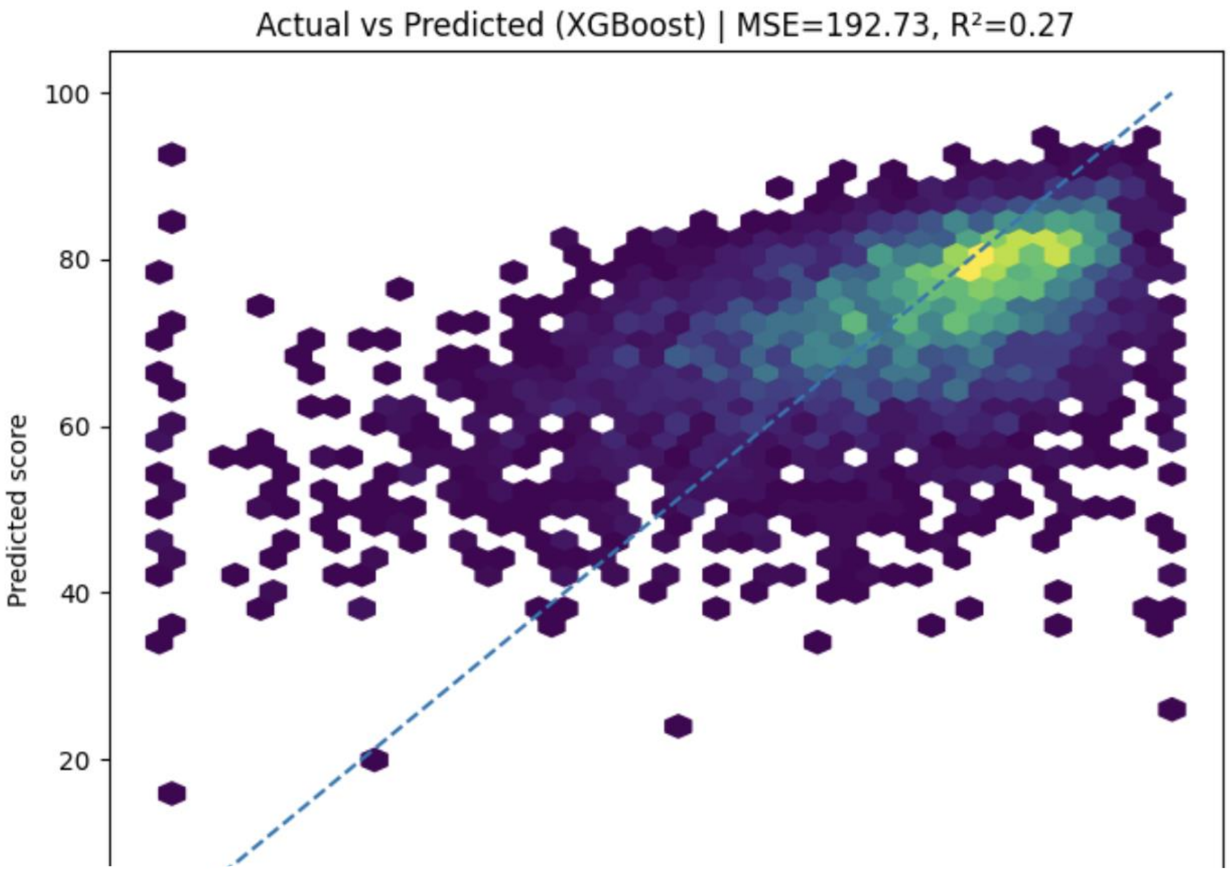


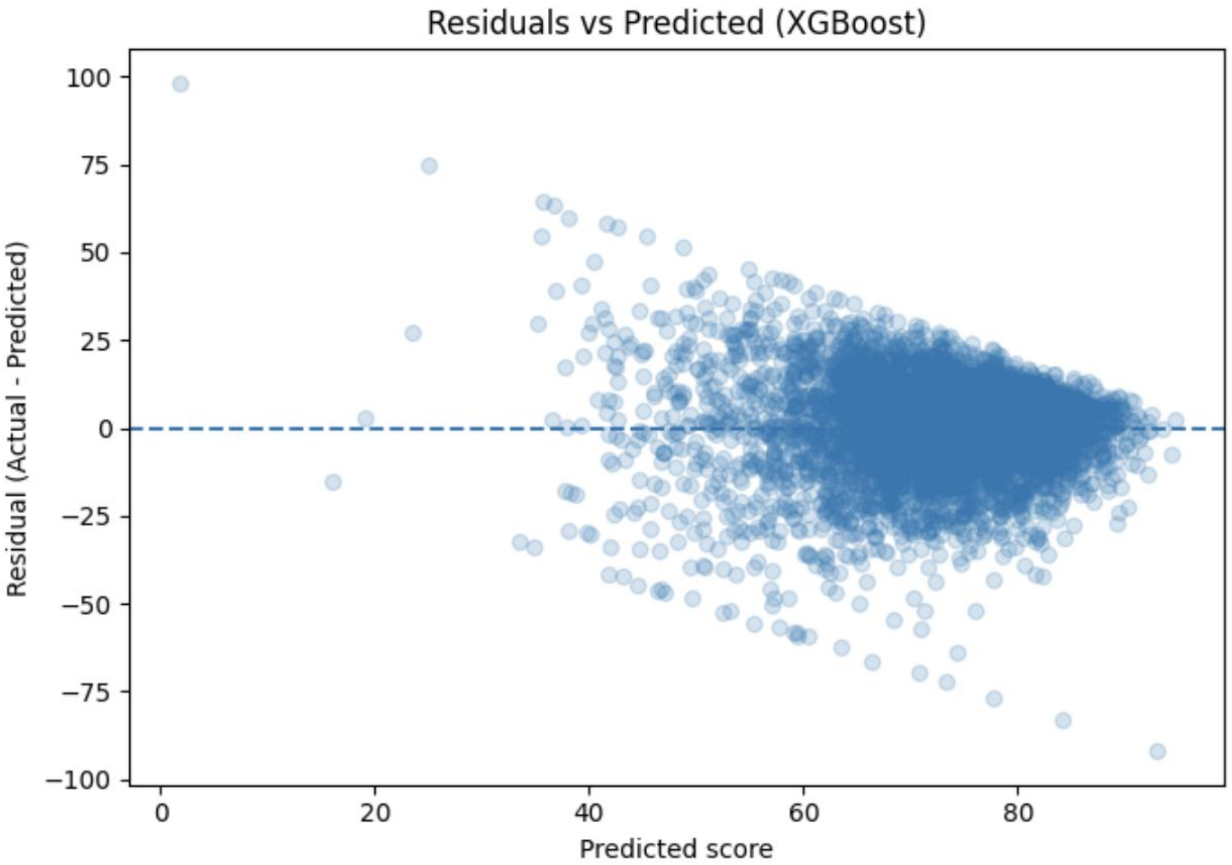
Input Features

1. Behavioral
   - assessments_attempted
   - active_days
2. Educational background
   - highest_education_Post Graduate Qualification
   - highest_education_HE Qualification
   - highest_education_Lower Than A Level
3. Course-related
   - module_GGG
   - module_BBB
   - module_EEE
   - module_FFF
4. Presentation (semester)
   - presentation_2013J
   - presentation_2014J

Baseline Performance and Limitations of Raw Feature Models

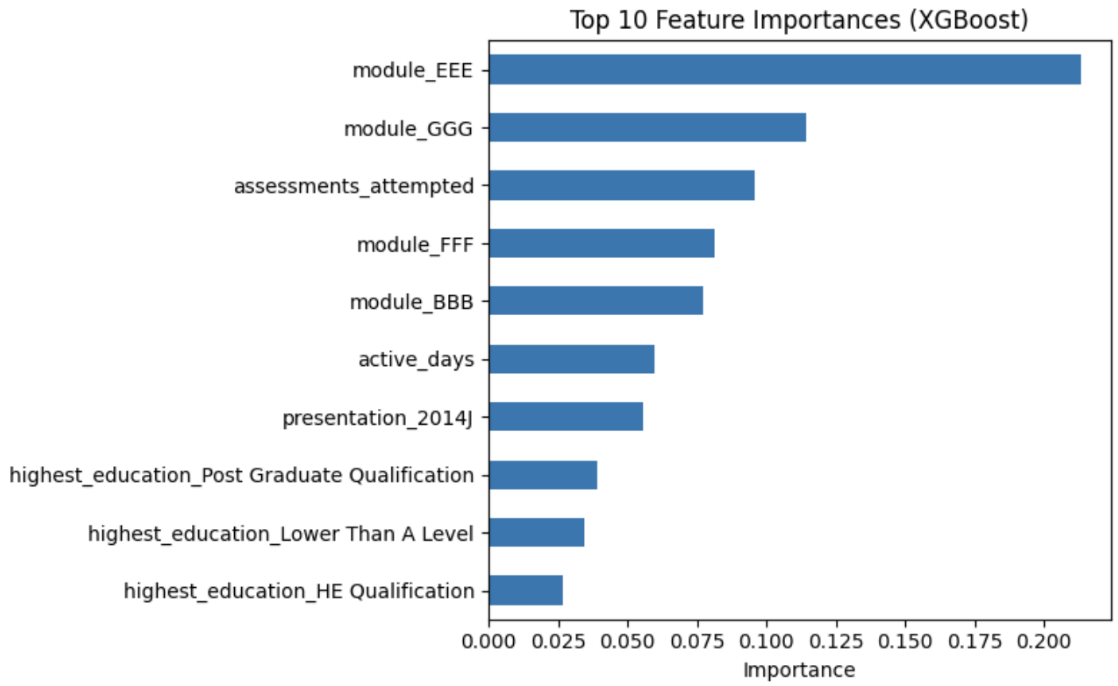
- The baseline linear regression model shows limited predictive performance when using the original feature set.
- Most input variables describe behavioral engagement (e.g., assessments attempted, active days) rather than direct academic outcomes.
- Educational background features capture prior qualifications, but do not reflect current course performance.
- Course-related features (modules and presentation periods) provide contextual information, not learning quality.
- As a result, a large portion of variance in the average assessment score remains unexplained.
- This limitation motivates both feature expansion and the use of more expressive models to improve prediction accuracy.

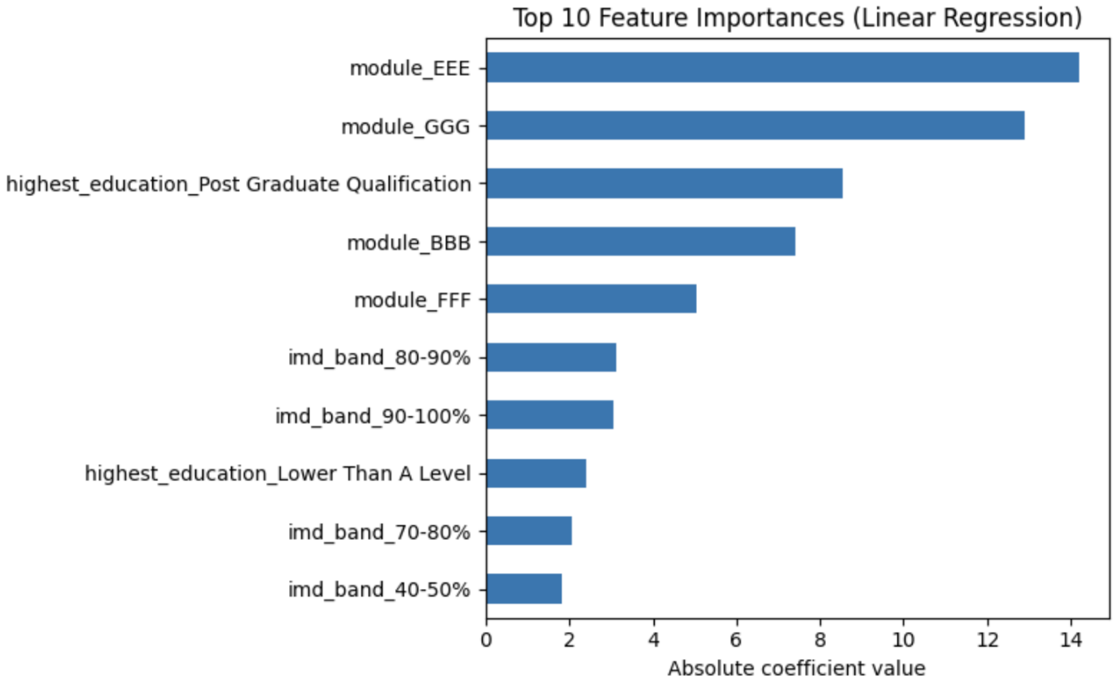




XGBoost Performance Using Raw Features

- Replacing the linear model with XGBoost does not significantly improve performance when using raw features. This suggests that feature quality, rather than model choice, is the primary limiting factor.





Feature Importance Comparison: Linear Regression vs XGBoost

Both models rely on the same raw features. Changing the model shifts importance weights, but does not unlock new predictive signals.

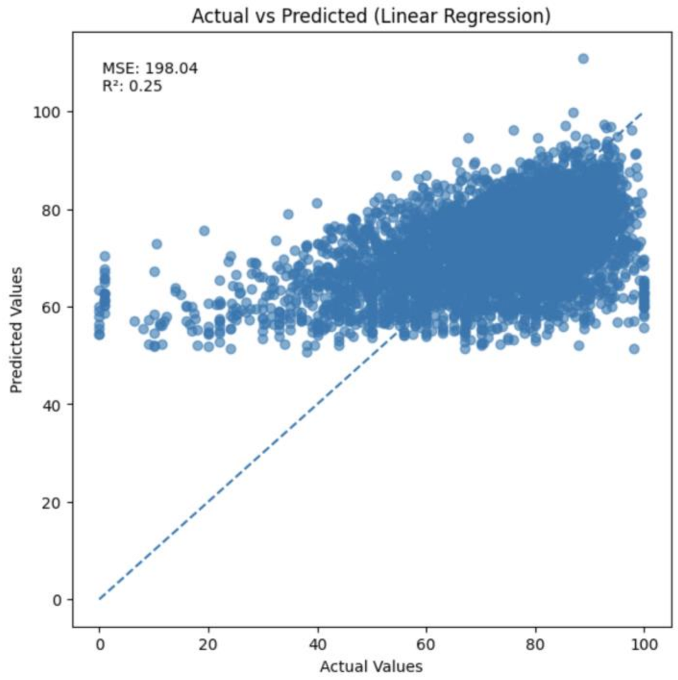
Feature Engineering

```
['assessments_attempted',
 'active_days',
 'clicks_per_day',
 'attempts_per_day',
 'clicks_per_credit',
 'clicks_per_attempt',
 'attempts_per_active_day',
 'active_days_ratio',
 'engagement_score',
 'assessment_pressure',
 'imd_band_30-40%',
 'imd_band_40-50%',
 'imd_band_50-60%',
 'imd_band_60-70%',
 'imd_band_70-80%',
 'imd_band_80-90%',
 'imd_band_90-100%',
 'highest_education_Post Graduate Qualification',
 'highest_education_HE Qualification',
 'highest_education_Lower Than A Level',
 'disability_Y',
 'module_GGG',
 'module_BBB',
 'module_EEE',
 'module_FFF',
 'presentation_2013J',
 'presentation_2014J']
```

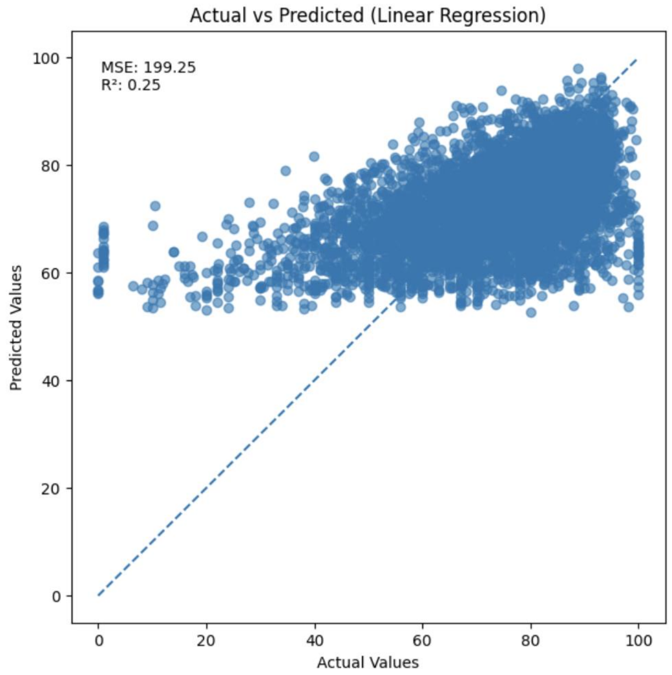
These engineered features transform raw activity counts into normalized and interaction-based signals, allowing the model to better distinguish between passive participation and meaningful engagement.

- clicks_per_day — average daily platform activity, capturing engagement intensity
- attempts_per_day — frequency of assessment attempts over time
- clicks_per_credit — learning effort normalized by course load
- clicks_per_attempt — engagement efficiency per assessment attempt
- attempts_per_active_day — assessment pressure during active study days
- active_days_ratio — consistency of participation relative to total credits
- engagement_score — combined measure of activity volume and duration
- assessment_pressure — interaction between attempt frequency and total attempts

Model Performance Improvement: Linear Regression

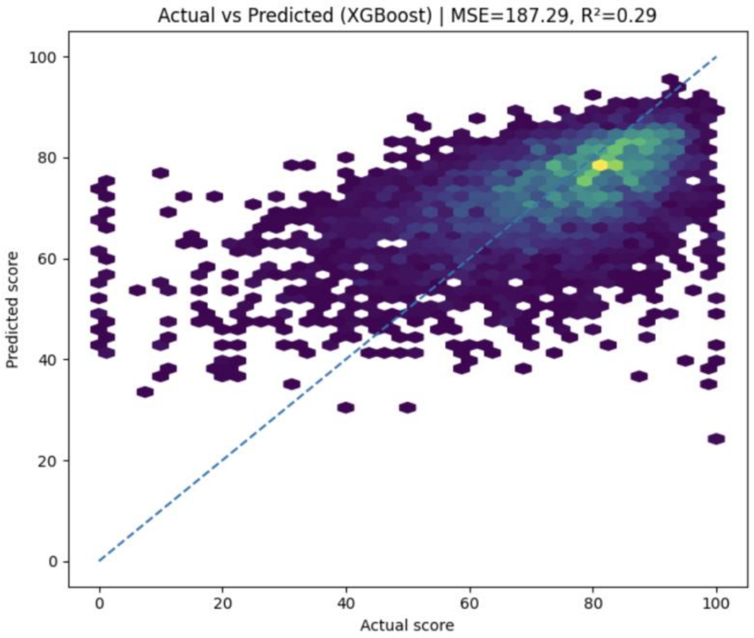


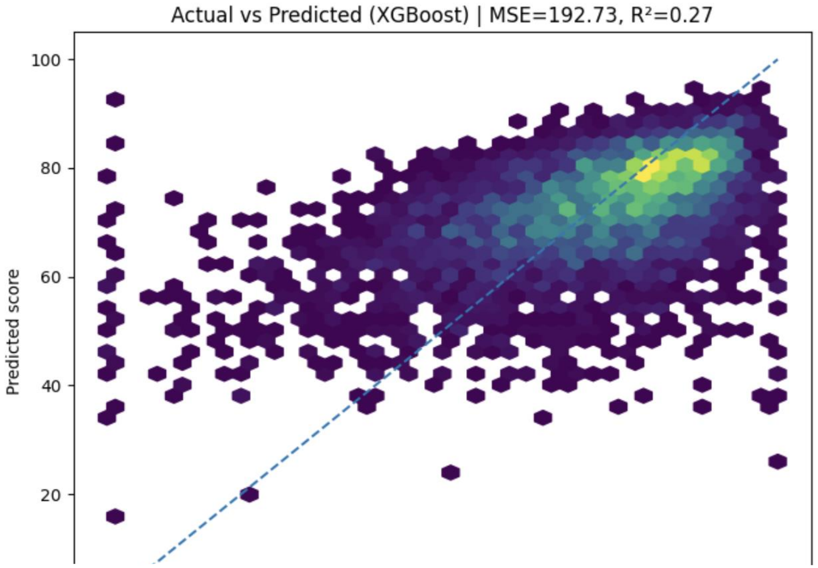
After Feature Expansion
MSE=198, R^2=0.25



Before Feature Expansion
MSE=199, R^2=0.25

Model Performance Improvement : XGBoost

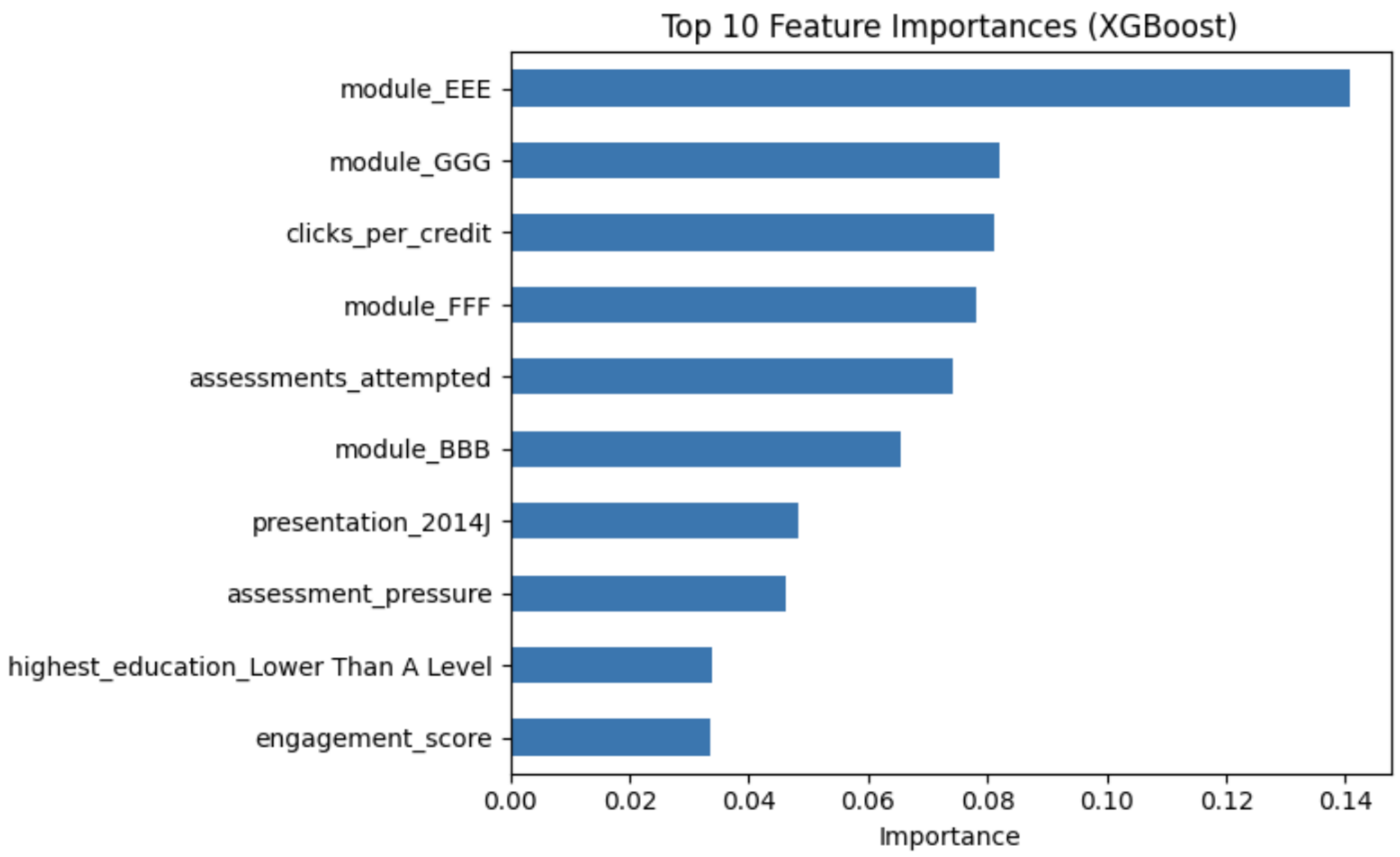




After Feature Expansion
MSE=187, R^2=0.29

Before Feature Expansion
MSE=192, R^2=0.27

Performance Gains with XGBoost and Engineered Features



Conclusion

- Adding new engineered engagement and assessment-based features led to minor improvements in model performance, but did not fundamentally change predictive accuracy.
- Switching from Linear Regression to XGBoost resulted in a small gain (best result: R^2 increased from ~0.27 to ~0.29, MSE decreased from ~192 to ~187), indicating that model choice alone is not the main limiting factor.
- Linear Regression performance remained largely unchanged ($R^2 \approx 0.25$ before and after feature expansion), confirming its limited ability to capture complex patterns in the data.
- Overall, both models reached a performance plateau, suggesting that data limitations, rather than modeling approach or feature engineering, constrain prediction quality.
- Meaningful improvements in predicting student scores would require richer and more informative data, particularly detailed assessment-level and academic performance indicators.