Boosted Reteach Planning: Optimizing Student Learning Outcomes

Final Report

Zehra Erden

zbe1949@mavs.uta.edu

Dr. Amir Farbin, Dr. Masoud Rostami

University of Texas at Arlington

May 2024

Introduction

Harmony Public Schools are charter schools with more than 60 campuses around the state of Texas. Harmony conducts two interim assessments in a year to make sure their students are ready for a state assessment called STAAR. Between these interim assessments teachers create reteaching plans based on the fall interim assessment results and reteach certain topics to their students using various resources they personally pick for their students. This research aims to analyze how these resources affect the student's academic performance on Math interim assessment scores. Using data science, we can accomplish this task.

Objective

The objective of this study is to investigate the impact of various educational resources on student performance in interim assessments and to assess how the utilization of these resources affects learning outcomes.

Dataset

The dataset contains reteach plans collected from teachers and student academic performance data from the Harmony administration. 6 different reteach plans received from 4 8th grade math teachers from different campuses as word or pdf documents while student academic performance data is in csv format.

Methodology

Data Understanding

Numerical features:

1. Student_ID

Student ID numbers that are randomly distributed due to FERPA.

2. Percent Score

   Percent score is the rank of the student compared to others in the same grade and campus.

3. Scale Score

   Scale scores the student received on the assessment.

4. Approach Probability

   Probability of student approaching desired scale score.

5. Meet Probability

   Probability of student meeting desired scale score.

6. Master Probability

   Probability of student mastering the topics.

Categorical features:

1. Assessment

   The type of assessment taken (‘Interim Fall’ or ‘Interim Spring’).

2. Course

   The course student took during the semester (‘Math-8’ or ‘Algebra I’)

3. Teacher Name

   Name of the teacher the student took class of.

4. Section

   The section/class of the student.

5. Mastery Projection

The outcome prediction of the student on the actual STAAR test (‘Did Not Meet’, ‘Approach’, ‘Meet’ and ‘Master’)

6. Projected Tier

   Projected tier of the student based on their percent score.

Data Preprocessing

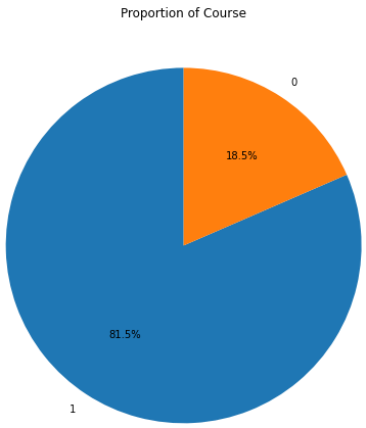
Handling Missing and Invalid Values

There are no missing values and invalid values in the dataset.

Outlier Detection and Treatment

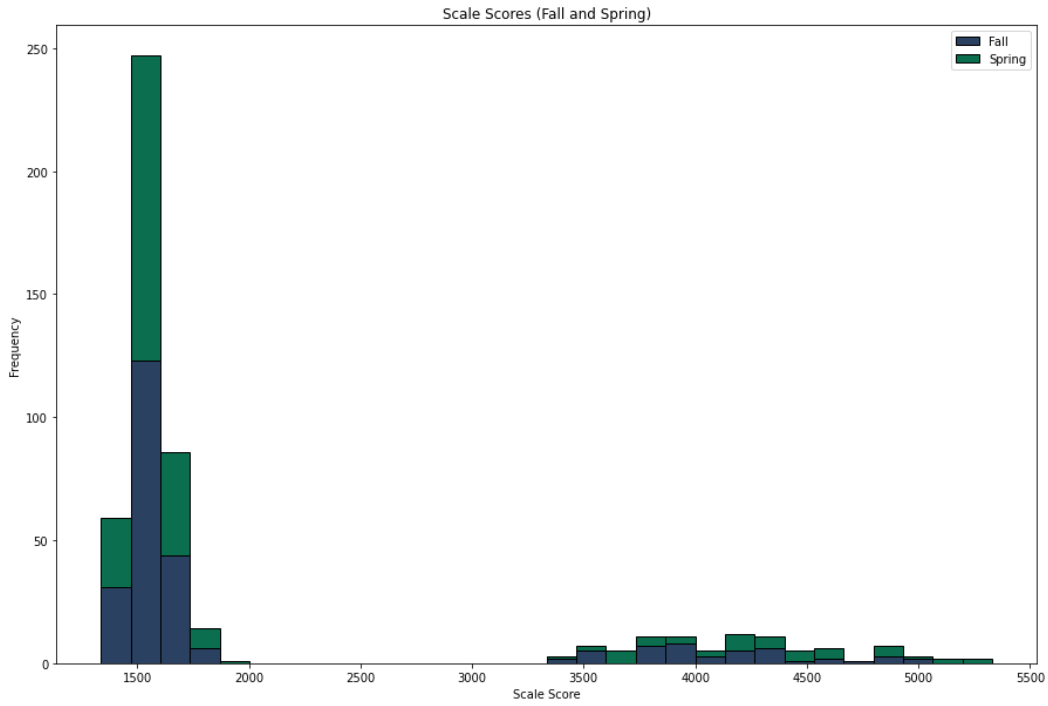
Outliers are identified during the exploratory data analysis. It is imperative to retain outliers since they represent both high-performing and low-performing students, providing valuable insights into the effectiveness of the resources for these students, aligning with the project's goals. That is why the outliers were kept for further analysis.

Single Variable and Multivariate Analysis

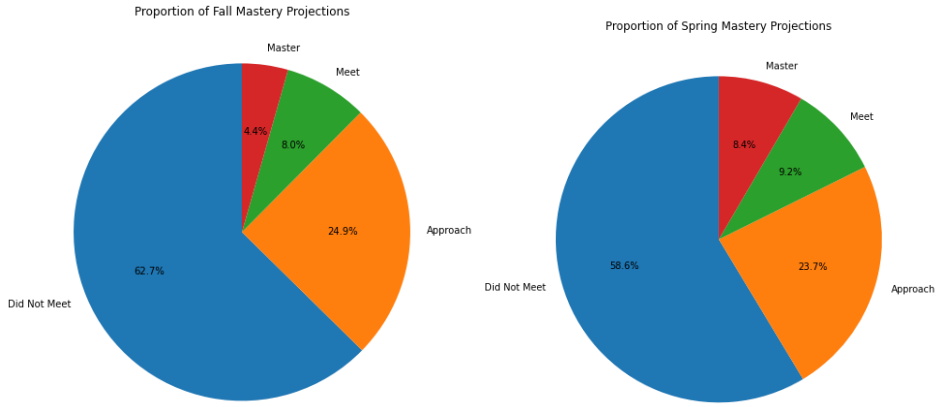


0 indicating Algebra I and 1 indicating Math-8

The course distribution is imbalanced on our sample.



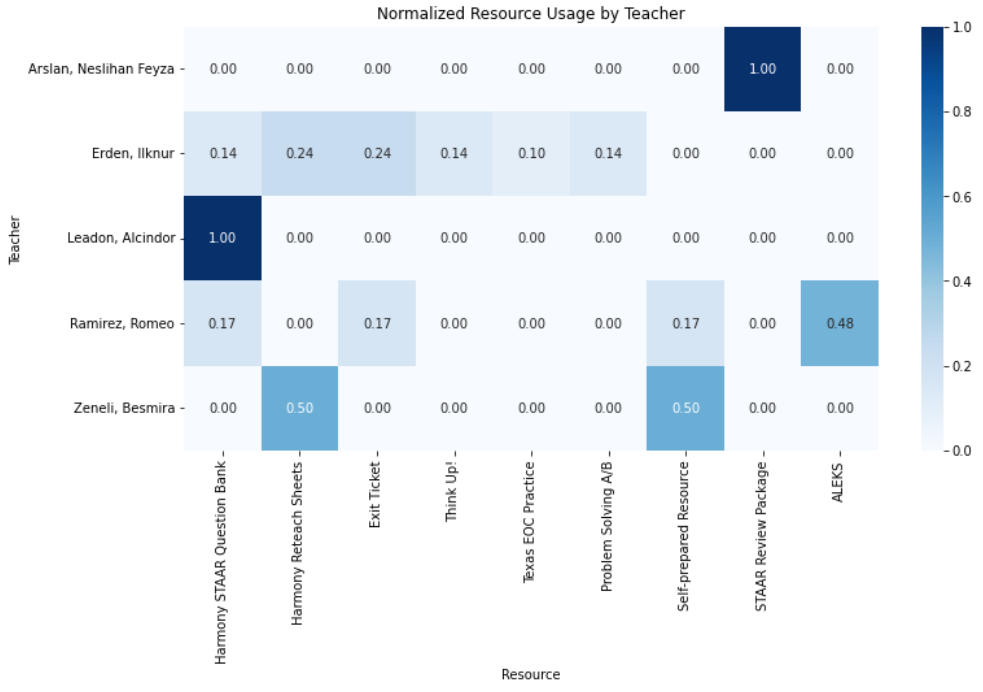
The stacked histogram for fall and spring scores shows that there is an increase in scale scores overall and spring data is a bit more right-skewed compared to fall data.



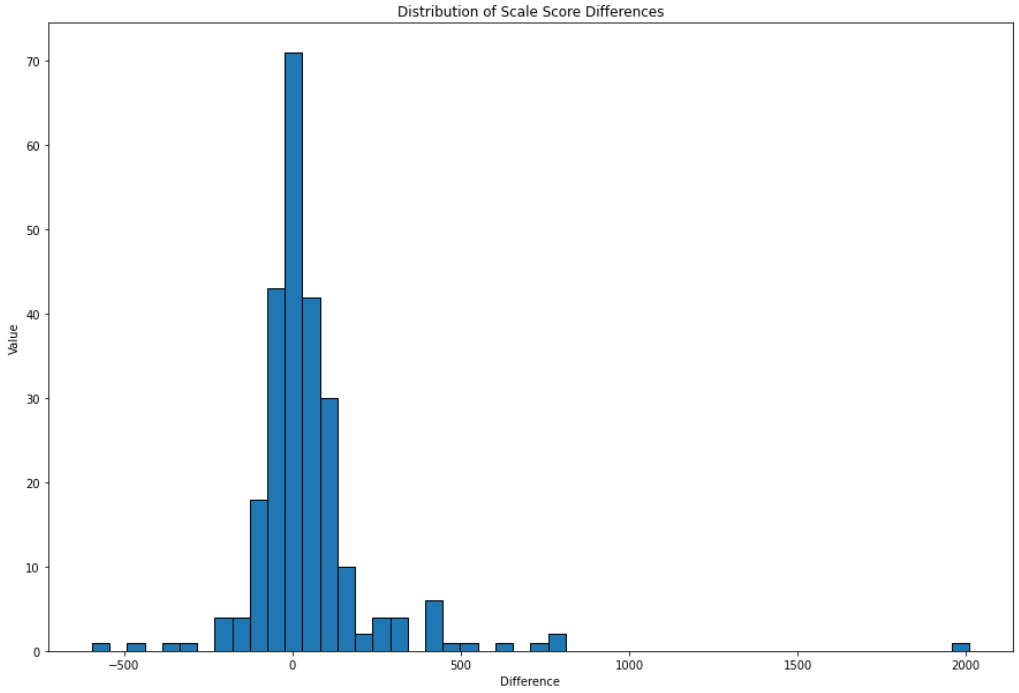
While the ‘Did Not Meet’ and ‘Approach’ projection proportions decreased, ‘Meet’ and ‘Master’ projection proportions increased, indicating the overall better performance in spring interim assessment.

58%	4333	Master	99%	99%	57%	Tier 1
56%	4310	Meet	99%	97%	44%	Tier 1
56%	4101	Meet	99%	74%	7%	Tier 1
18%	1429	Did Not Meet	1%	1%	1%	Tier 3 (RTI)
32%	1528	Did Not Meet	8%	1%	1%	Tier 3 (RTI)
29%	1578	Did Not Meet	52%	4%	1%	Tier 2
41%	1647	Approach	90%	15%	1%	Tier 1
12%	1375	Did Not Meet	1%	1%	1%	Tier 3 (RTI)
65%	1822	Meet	99%	99%	62%	Tier 1

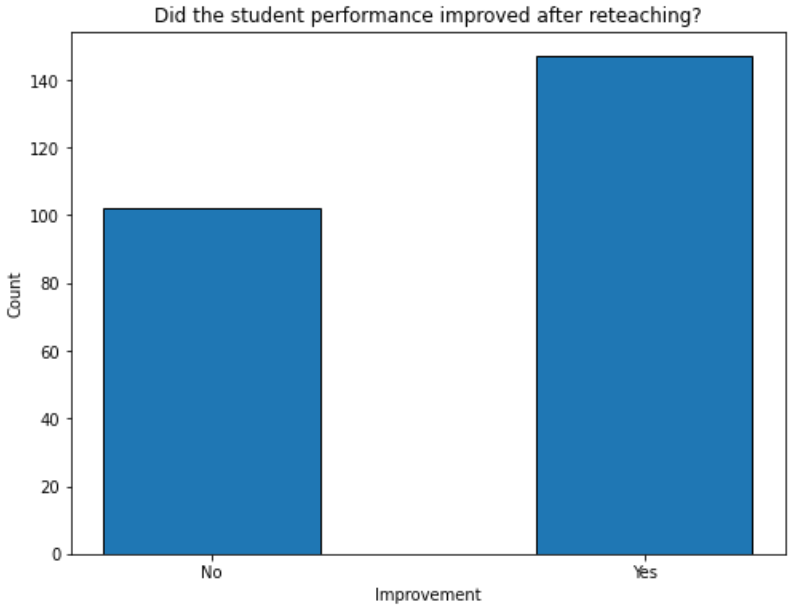
It is concluded that the percent score and projected tier variables is calculated based on the peer and campus comparison. Since we want to see the effects on resources on individual student academic performance, it is decided that they are not useful for analysis and are not used as predictors.



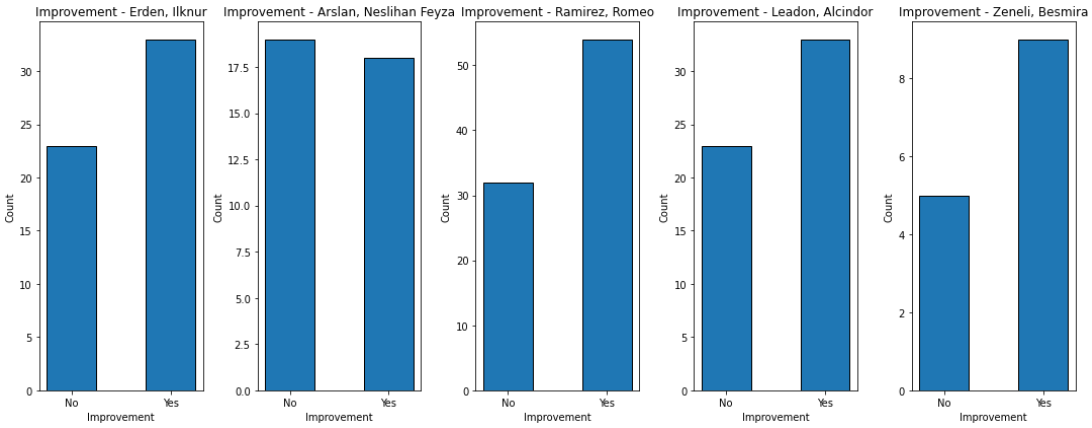
In this visualization we see the proportions of resource usage for each teacher. There are some teachers that used plenty of resources while some of them only used one. It is important to highlight those teachers who taught both Algebra I and Math-8 used more different resources when reteaching.



The ‘Difference’ is the feature that is generated using the Spring and Fall Scale Scores. It indicates the difference between them. The positive values in the ‘Difference’ column show the improvement while the negative values display the otherwise.



‘Improvement’ also is a feature generated using feature engineering. It contains binary values. ‘1’ for positive value in ‘Difference’ columns and ‘0’ for negative, indicating if the student did improve or not. There are more students who improved but the number of students who did not do better in spring is also significant.



Overall teachers did good and the majority of their students improved their academic performance.

Statistical Tests

One-way ANOVA (Course v. Fall Scale Score)

```
    Multiple Comparison of Means - Tukey HSD, FWER=0.05
=============================================================
group1 group2  meandiff  p-adj   lower      upper    reject
-------------------------------------------------------------
     0      1 -2502.6848   0.0 -2582.3714 -2422.9983   True
-------------------------------------------------------------
ANOVA F-value: 3826.5269095179806
ANOVA p-value: 2.428445311522122e-152
```

Null hypothesis of no difference between the groups can be rejected based on the adjusted p-value. "True,” means the null hypothesis is rejected, suggesting that there is a significant difference between the groups.

The mean of group2 (Math-8) is approximately 2502.6848 units lower than the mean of group1 (Algebra I). Students who take Algebra I outperformed the students who take Math-8.

Two-way ANOVA (Teacher, Course v. Difference)

	sum_sq	df	F	PR(>F)
C(Teacher)	1.827545e+06	4.0	13.165178	5.430736e-08
C(Course)	8.921356e+05	1.0	25.706882	7.904064e-07
C(Teacher):C(Course)	5.700087e+05	4.0	4.106200	3.077971e-03
Residual	8.398405e+06	242.0	NaN	NaN

Test results conclude that 'Teacher' and 'Course' factors have a significant effect on the 'Difference' variable indicated by the very low p-values. Interaction between 'Teacher' and 'Course' is also significant.

Data Transformation

After removing the projected tier and percent score variables our dataset had less predictor variables. To come up with a solution more feature engineering is done. I decided to create a "Projected Tier" like feature to robust the model by categorizing the scale scores and labeling them.

```
data['Scaled_Score_Category'] = pd.cut(data['FA Scale Score'], bins=3, labels=['Low', 'Medium', 'High'])
```

Feature Scaling

To get all features in same scale of 0 to 1, min-max scaling (normalization) method is used. It is less sensitive to outliers compared to standard (z-score) normalization and scales all variables in 0 to 1 scale which is very helpful in our case.

Encoding Categorical Variables

Categorical variables with ordinal data are replaced with ordinal numbers. Fall Mastery Projection and Scaled_Score_Category are label encoded.

Modeling

After making sure the dataset is ready for modeling phase, 80% of the data is split for training and the other 20% remaining is used for testing.

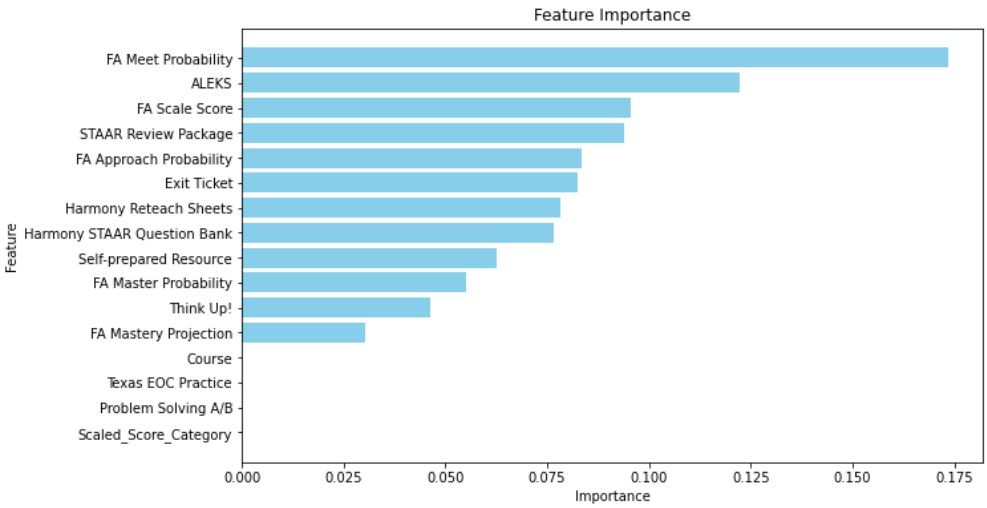
Model Selection

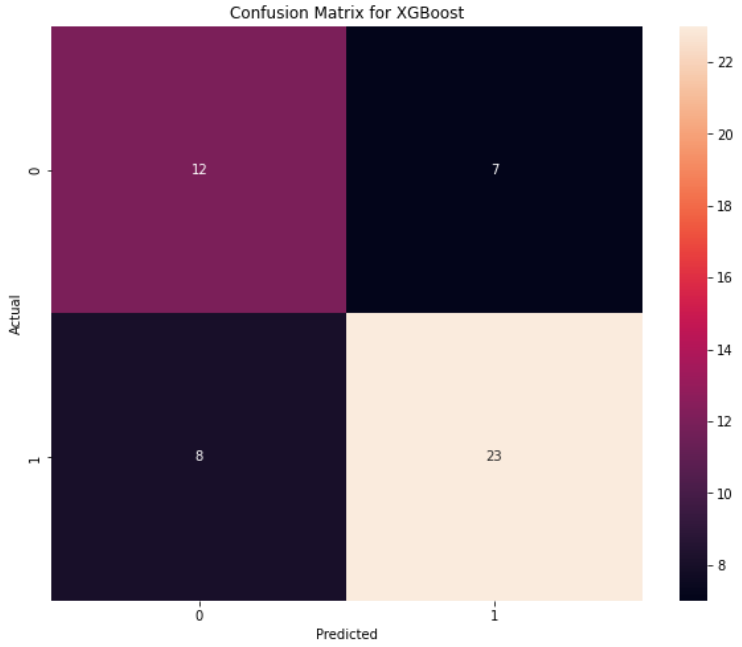
Since many features in dataset are created later manually and there is high correlation between the variables expected, XGBoost and Logistic Regression models are used to see the effects of the resources on academic performance.

Model Evaluation

XGBoost

```
Accuracy: 0.7
Precision: 0.7666666666666667
Recall: 0.7419354838709677
F1 Score: 0.7540983606557377
AUC-ROC: 0.7843803056027165
MAE: 0.3000
MSE: 0.3000
Log Loss: 0.5740
```





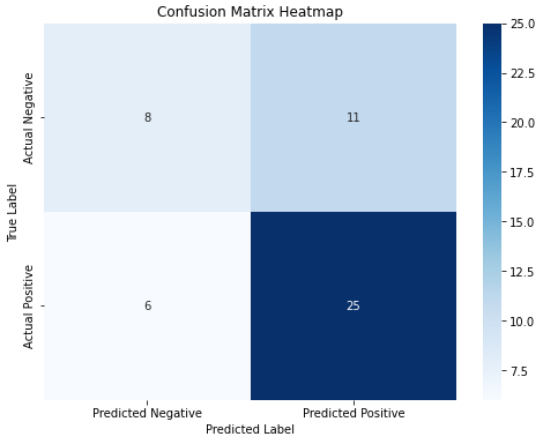
XGBoost is the best performing model with this dataset but it still performs poorly.

Logistic Regression

```
Accuracy: 0.58
Classification Report:
              precision    recall  f1-score   support

           0       0.42      0.44      0.43        18
           1       0.68      0.66      0.67        32

    accuracy                           0.58        50
   macro avg       0.55      0.55      0.55        50
weighted avg       0.59      0.58      0.58        50
```



Logistic regression did not perform better than XGBoost, but it gives us more insights about the predictors/features and their effect on the response variable, ‘Improvement’.

	coef	std err	t	P>|t|	[0.025	0.975]
Harmony STAAR Question Bank	0.5893	0.065	9.005	0.000	0.460	0.718
Harmony Reteach Sheets	0.4239	0.092	4.618	0.000	0.243	0.605
Exit Ticket	-0.0257	0.095	-0.269	0.788	-0.214	0.162
Think Up!	-0.2513	0.047	-5.389	0.000	-0.343	-0.159
Texas EOC Practice	0.3409	0.066	5.144	0.000	0.210	0.471
Problem Solving A/B	-0.2513	0.047	-5.389	0.000	-0.343	-0.159
Self-prepared Resource	0.2190	0.081	2.711	0.007	0.060	0.378
STAAR Review Package	0.4865	0.081	6.043	0.000	0.328	0.645
ALEKS	0.5714	0.062	9.262	0.000	0.450	0.693

Harmony STAAR Question Bank:

Coefficient: 0.5893

P-value is very low.

Confidence interval does not include the zero.

We can conclude that this feature has a statistically significant positive effect on the dependent variable.

Harmony Reteach Sheets:

Coefficient: 0.4239

The P-value is also very low.

Confidence interval does not include the zero.

Has a significant positive effect on the dependent variable.

Texas EOC Practice:

Coefficient: 0.3409

Low p-value.

Confidence interval does not include the zero.

Has a significant positive effect on the dependent variable.

Self-prepared Resource, STAAR Review Package, ALEKS:

These features also have high positive coefficients and significantly low p-values. Their confidence intervals do not include the zero value also. We can say that these resources also have a significant positive effect on student performance.

Exit Ticket:

Coefficient: -0.0257

P-value: 0.788

While the coefficient is negative it is not statistically significant because the p-value is high. We cannot conclude that this feature has a significant effect on the “Improvement” variable.

Think Up!:

Coefficient: -0.2513

P-value is low.

Different from “Exit Ticket”, this feature has a negative coefficient with a significantly low p-value meaning that this resource can be associated with a decrease in the student performance.

Problem Solving A/B:

Coefficient: -0.2513

P-value: ~0.0

Similar to “Think Up!” feature, results suggest that this feature has a significantly negative effect on the dependent variable.

Conclusion

Think Up! and Problem-Solving A/B negatively affects the reteaching process.

Harmony STAAR Question Bank, Harmony Reteach Sheets, Texas EOC Practice, Self - prepared resources, STAAR Review Package, and ALEKS positively affects the student performance.