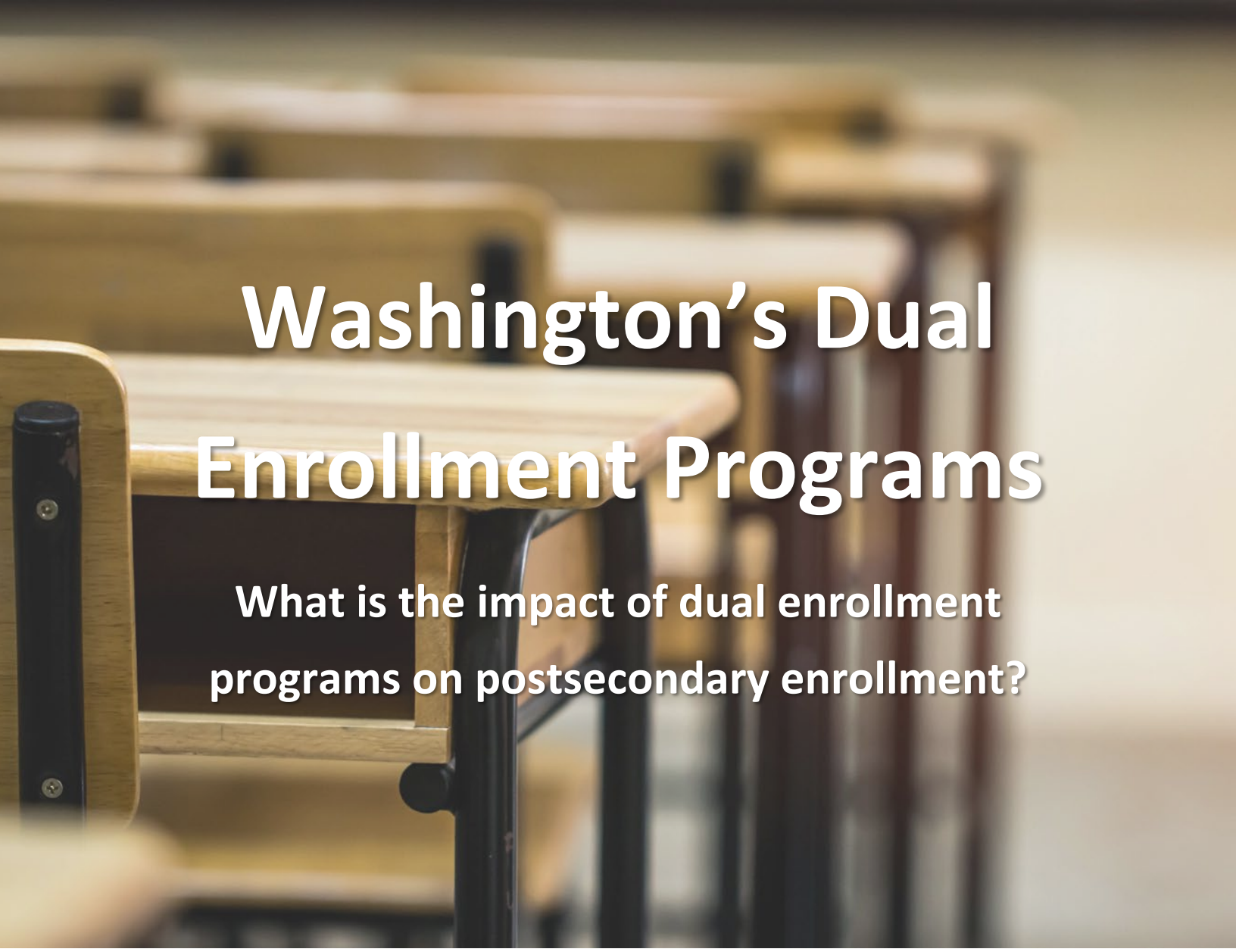

Washington's Dual Enrollment Programs

What is the impact of dual enrollment programs on postsecondary enrollment?

September 2021

ERDC
Education Research & Data Center

Washington Student Achievement Council

Foreword

This study was a collaborative effort led by the Washington State Education Research and Data Center and the Washington Student Achievement Council, carried out by data scientist Dr. Mike Preiner. Part of the rationale for the project was to demonstrate that aggregate data could be effective in its use to develop rigorous analyses. Generally, individualized record-level data allows for a richer analysis and ability to examine intersectionality and more specific outcomes. But in many cases, aggregate data is sufficient to perform a rigorous analysis that can estimate a program's effect. This technical discussion paper details how aggregate data was used to estimate the impact of dual enrollment on postsecondary enrollment for Washington students.

The study was completed as part of a larger program funded primarily by federal grant CFD #84.372A NCES 15-01 awarded by the Institute for Education Sciences in the United States Department of Education to the State of Washington's Office of the Superintendent of Public Instruction and carried out by the Office of Financial Management's Education Research and Data Center. The total program cost is $8,492,963.38. Eighty-four point eight percent (84.8 percent) ($7,203,021) of the total cost of the program is financed with this federal grant money, and 15.2 percent ($1,289,942.38) by the State of Washington.

Table of Contents

Executive Summary

Washington invests heavily in Dual Enrollment (DE) programs, with over $150 million spent in 2017 alone. The goal of this investment is to improve systemic inequities and improve postsecondary enrollment rates across the state. While there has been some preliminary evidence suggesting that DE may be successful in those goals (*Dual Credit Task Force Agenda*, 2020), there has not yet been a systematic study examining the effectiveness of these programs after controlling for other factors that also impact postsecondary enrollment, such as demographics, family income, and standardized test scores.

In this report we analyze DE data from the 2017 class of Washington graduating seniors to estimate the impacts of DE on postsecondary enrollment after controlling for a wide range of confounding factors. Part of the rationale for the project was to demonstrate that aggregate data could be effective in its use to develop rigorous analyses. Generally, individualized record-level data allows for a richer analysis and ability to examine intersectionality and more specific outcomes. But in many cases, aggregate data is sufficient to perform a rigorous analysis that can estimate a program's effect. This technical discussion paper details how aggregate data was used to estimate the impact of dual enrollment on postsecondary enrollment for Washington students.

Our models are built using anonymized, aggregate data, and with the exception of DE status, all of the data used is publicly available (*Data.WA.gov - the General Purpose Open Data Portal for the State of Washington.*, n.d.). The models include data for 155,931 high school graduates over 3 years (2016, 2017, and 2018), including 24,245 DE participants across 171 high schools in 2017. We incorporated data from two DE programs: Running Start and College in the High School, and the analysis revealed several clear results.

Both Running Start and College in the High School had a positive, statistically significant impact on postsecondary enrollment. It is worth noting that the two programs have some significant differences (*College in the High School Frequently Asked Questions*, 2019):

- Running Start is taught on a college campus by college faculty, while College in the High School is taught by high school teachers on a high school campus.
- Running Start is only eligible to students in 11th and 12th grade, while College in the High School is open to students in 10th, 11th, and 12th grade.
- Running Start tuition is fully subsidized for students, while students in College in the High School are responsible for paying application tuition fees in order to earn college credit.

Given these differences, it isn't surprising that the two programs also differed in their impacts on postsecondary enrollment: **participation in Running Start had approximately 4 times the impact on postsecondary enrollment than participation in College in the High School.**

We also find that most of DE's impact is via increased enrollment into 2-year programs. **Both Dual Enrollment programs were roughly twice as effective in increasing enrollment into 2-year programs than into 4-year programs in Washington.**

Finally, we compare the impact of DE to other approaches to increasing postsecondary enrollment. While DE (especially Running Start) has a positive impact on postsecondary enrollment, we find that other factors have larger impacts, especially for 4-year enrollment. **For example, increasing high school math proficiency by 1% would increase 4-year college enrollment almost 8 times more than a 1% increase in Running Start participation.** This may have important policy implications in terms of where to most effectively allocate resources towards increasing postsecondary attainment.

While our analysis covers the vast majority of DE participants attending Washington colleges in 2017, we would expect the relationship between DE and postsecondary enrollment to change over time. We recommended extending our analysis as new data becomes available, both in terms of additional years of DE data and in terms of additional variables to the model.

Introduction

Part of the rationale for the project was to demonstrate that aggregate data could be effective in its use to develop rigorous analyses. Generally, individualized record-level data allows for a richer analysis and ability to examine intersectionality and more specific outcomes. But in many cases, aggregate data is sufficient to perform a rigorous analysis that can estimate a program's effect. This technical discussion paper details how aggregate data was used to estimate the impact of dual enrollment on postsecondary enrollment for Washington students.

Dual Enrollment (DE) programs have a long history in Washington state. For example, the Running Start program has been offered since 1990 (Smith, 2014) with a goal of improving systemic inequities and improving postsecondary enrollment rates across the state (*Dual Credit Task Force Agenda*, 2020). While there have been numerous studies on access to and participation in these DE programs (Miller et al., 2019, Zinth & Barnett, 2018), there has been less focus on estimating the *causal* impacts of dual enrollment programs on student outcomes, such as postsecondary enrollment.

Many previous studies on the effectiveness of DE programs have performed comparisons of outcomes between students who participate in DE and those who do not (*Running Start 2005-06 Annual Progress Report*, 2006). However, this approach suffers from sample bias: the students who participate in DE are fundamentally different from the students who don't participate. For example, students who participate in DE are more likely to pass the state standardized tests than their non-participating peers. When it comes to determining the impact of DE on college enrollment, we thus need to separate the impact of DE from that of higher test scores, along with any other confounding factors.

In this report we attempt to isolate the impact of DE programs on postsecondary enrollment using a decision-tree based regression modeling approach based on aggregated student data. We control for a large number of confounding factors, including school-based measures and student-group measures, which are described below in more detail.

Analysis

Our postsecondary enrollment model is fundamentally built at the school-year-student group level. This means that each row in our processed data has the following form, where RS %

denotes the group's participation rate in Running Start and CHS % denotes the group's participation rate in College in the High School:

School	Grad Year	Student Group	District Code	...	Graduation Rate	RS %	CHS %	Postsecondary Enrollment %
School A	2017	Asian	55010	...	0.91	10.2	4.5	32.1
School A	2017	Low-income	55010	...	0.70	7.5	3.4	18.2

This level of aggregation allows us to use publicly available, anonymized data, while still providing enough resolution to disaggregate results by demographic groups. Previous research has suggested that data aggregated at this level is often appropriate to answer the type of research questions posed in this report (Jacob et al., 2014, 44-66).

Measures

There are a large number of factors that we could reasonably expect to impact postsecondary enrollment. To attempt to isolate the effect of dual enrollment, our model controls for many of these. They include the following:

Year: we include the year of the graduating cohort in our model to estimate any variability by time in our model. Our model includes 2016, 2017, and 2018.

School Effects: these are factors that reflect properties of the school environment. They include

- *District Code:* the 5 digit district code is included as a variable, allowing schools to be grouped by district.
- *School Type:* The Washington Office of the Superintendent of Public Instruction (OSPI) defines several different types of schools in Washington. Our data includes both "standard public schools" and "alternative schools".
- *Number of Students:* the number of students in the graduating class.
- *Demographic Data:* we calculate the demographic percentages of a variety of student groups at each school, and include these metrics in our model. For example, we include the percentage of students that are classified as from low-income families. We include data from the following groups as school-based percentages in our model:
    - American Indian/Alaska Native students
    - Asian students
    - Black/African American students
    - English-language learners (ELL)

- Hispanic/Latino students
- Homeless students
- Low-income students (defined by participation in the free-and-reduced price lunch program)
- Migrant students
- Students with disabilities
- Students in foster care
- Students with military parents
- White students

Student group characteristics: we also track a number of factors that measure the characteristics of the specific student groups within each school. These include:

- *Group category:* this is a categorical variable with a unique value for each student group (spanning all of the options listed in the *Demographic Data* section of School Effects above).
- *Assessment Scores:* we track the percentage of students passing both the math and English Language Arts (ELA) portions of the Smarter Balanced Assessment (SBA).
- *Running Start Participation Rate*: the fraction of students enrolled in Running Start
- *College in the High School Participation Rate:* the fraction of students enrolled in College in the High School.
- *Graduation Rates:* the fraction of students in the 12th grade class that successfully graduated.

Outcomes Measures: our outcomes of interest are postsecondary enrollment. We separate outcomes into three categories:

- *Enrollment in 4-year institutions:* the fraction of students that enrolled in a 4-year college.
- *Enrollment in 2-year institutions:* the fraction of students that enrolled in a 2-year college or technical program.
- *Enrollment in any postsecondary:* the fraction of students that enrolled in *any* postsecondary education.

Sample

The data used in our analysis comes from two different sources:

1. Publicly available aggregated and anonymized data from both OSPI and the Washington State Education and Research Data Center (ERDC) (*Data.WA.gov - the General Purpose Open Data Portal for the State of Washington.*, n.d.).

2. A sample of detailed postsecondary outcome data aggregated by DE participation (provided by ERDC).

The two data sets span two different (but overlapping) student groups. The publicly available data spans 2016, 2017, and 2018 and contains data for 155,931 students at 258 schools.

Our base DE data covers both 2015 and 2017. However, we are unable to join the 2015 data to our publicly available data, as the publicly available datasets don't include the relevant assessment data for the 2015 cohort. The joined DE data (for 2017) spans 171 schools and includes 24,245 students. It includes 13,084 Running Start participants and 13,143 College in the High School participants. There are a small number of students who participated in both DE programs.

Analytical Approach

To address this fundamental issue of a large number of correlated covariates in our model, many of which exhibit non-linear relationships, we employ a histogram-based gradient boosted regression model (*Light Gradient Boosting Machine*, 2021). It is a decision-tree based machine learning approach that natively deals with missing data and can be used for analyses with large numbers of correlated variables.

Results

Unless otherwise noted, we will focus on the results for the overall postsecondary enrollment model. To illustrate the relationships between some important model parameters and postsecondary enrollment (PSE), a series of scatterplots is shown below. Each dot represents a unique school-year-student group (for example: Low-income students from Roosevelt High School in 2017), with the dot size proportional to the number of students in each group. The plots show that while PSE generally increases with all of the four variables (graduation rate, Math SBA pass rate, ELA SBA pass rate, Running Start fraction, and College in the High School fraction), the strongest and clearest relationship is with Math SBA pass rate.

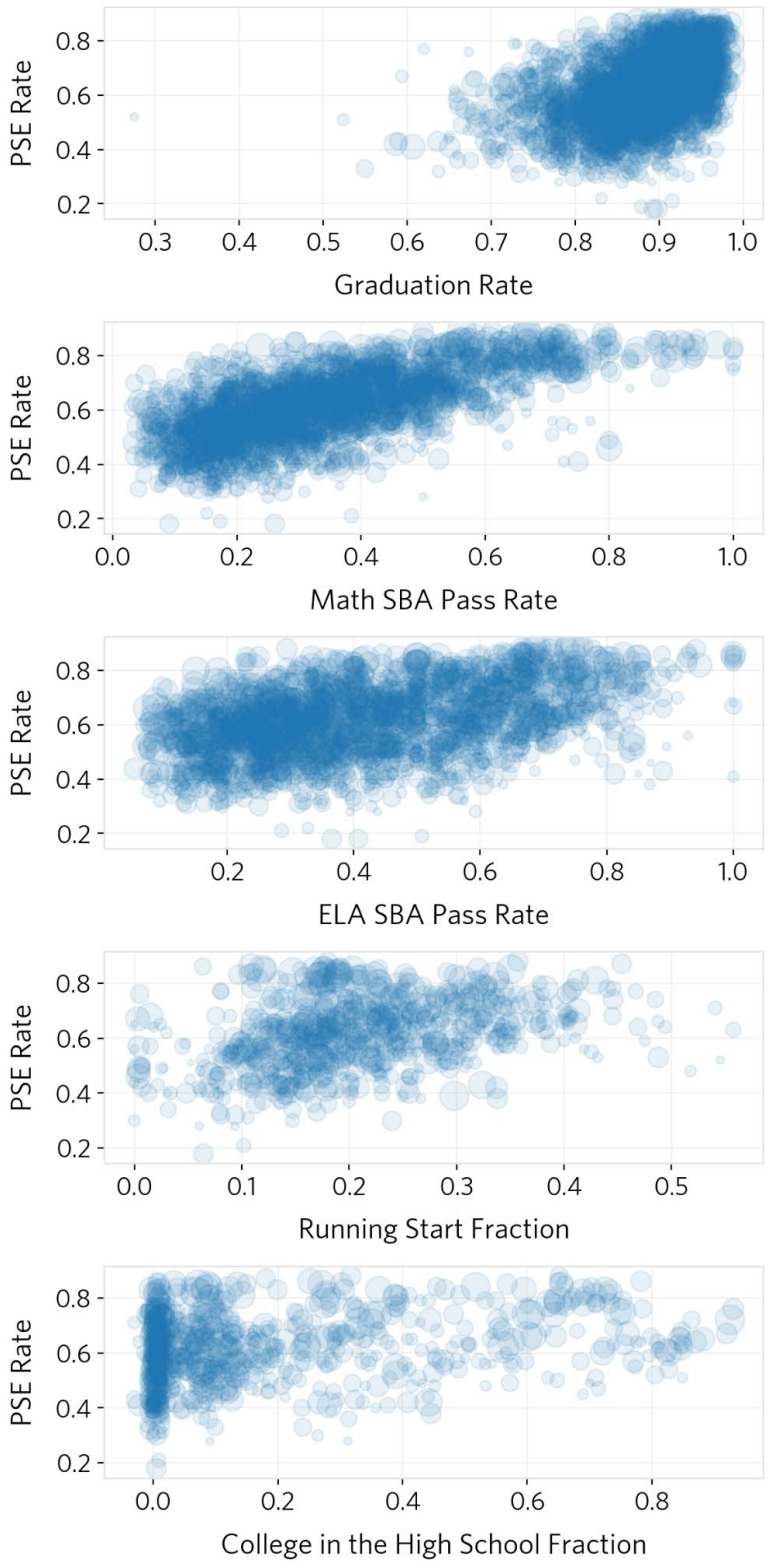

Figure 1. Plot of postsecondary enrollment (PSE) rate versus various student-group characteristics.

A full correlation matrix heatmap of all numerical model variables is shown below. We can see that the variables with the strongest correlations with postsecondary enrollment (in descending order) are: SBA Math Pass Rate (0.62), Low-income Fraction (-0.46), Asian Fraction (0.42), Graduation Rate (0.39), SBA ELA Pass Rate (0.39), and Running Start Fraction (0.35).

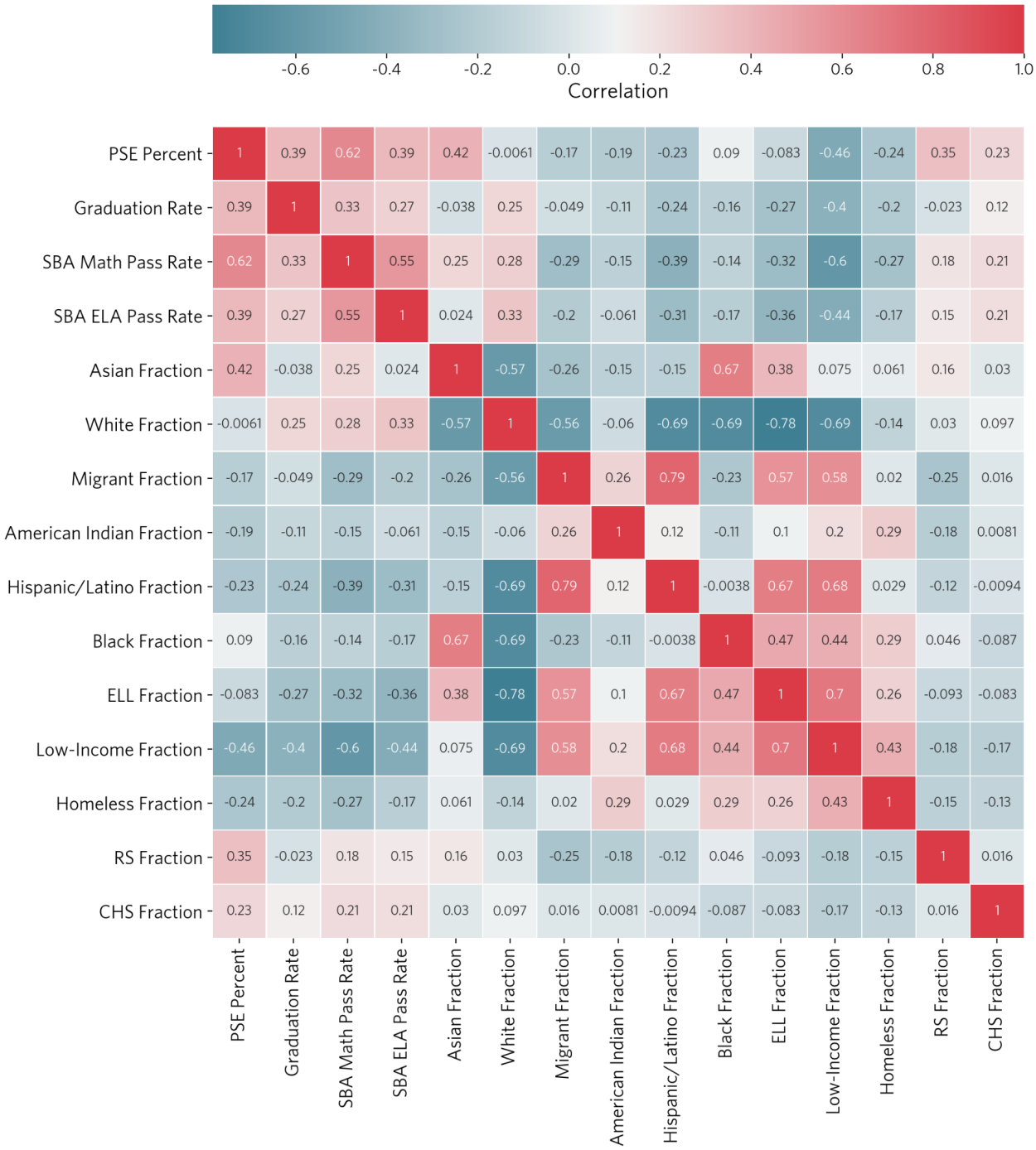

Figure 2. Heatmap showing correlation coefficients between the different numerical variables in the postsecondary enrollment model.

Model Accuracy

Before using a model to predict the impact of changing inputs on an output variable, it is important to establish that the model can accurately predict the variables of interest. To assess the predictive power of our model, we use 5-fold cross validation. This technique randomly selects 20% of the data to withhold for model assessment, and then uses the remaining 80% of the data to train the model. After training, model accuracy is assessed by comparing predictions for the remaining 20% against the actual values. This process is repeated 4 more times so that all of the data is eventually withheld for assessment. From these results we calculate the mean absolute model error, median model error, and R^2. We then repeat the process 20 times to remove any artifacts from the random sampling, and then average the results.

The final accuracy metrics are shown in the table below. We can see that our average prediction error for the postsecondary enrollment fraction is 0.040 (or 4%), and median absolute error is 0.030. This means that on average, our model can accurately predict the postsecondary enrollment of school-year student groups to within 4%.

Accuracy Metrics for "Enrollment in Any Postsecondary" Model

Mean Absolute Error	**Median Absolute Error**	**R^2**
0.040	0.030	0.823

To have a more complete view of the model error for all school-year-student groups, we plot a histogram of all of the results of the cross validation below. Over 90% of the school-year student groups have an error less than +/- 9%.

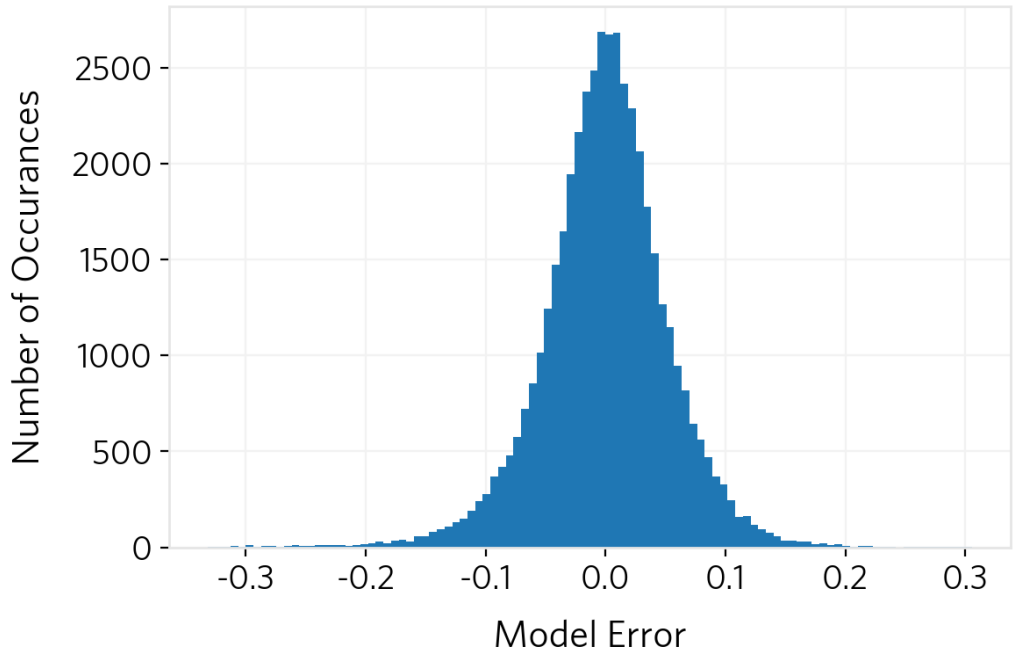

Figure 3. Histogram of model error. The error was calculated with 5-fold cross-validation and averaged over 20 repetitions.

Partial Dependence Analysis

After our model is trained, we can analyze the relationships between our output variable (PSE Fraction) and individual input variables after controlling for all other model factors. We do this using a partial dependence plot (*Partial Dependence and Individual Conditional Expectation Plots*, 2021). Categorical variables (such as school type and student group) were encoded as numerical variables. The partial dependence plots for all of the variables in our model are shown below.

Figure 4 confirms many of the relationships seen in both our scatterplots and correlation heatmap. For example, we can see a strong positive relationship between SBA Math Pass Rate and PSE Fraction, and a strong negative relationship between Low-income Fraction and PSE Fraction. This would match our naive expectations: all else being equal, we'd expect college enrollment to increase as test scores increase. Similarly, all else being equal, we expect college enrollment to decrease as a school's average family income decreases. In the case of the DE programs, the partial dependence plots show a greater impact on PSE Fraction for Running Start than College in the High School, which again matches our previous analyses.

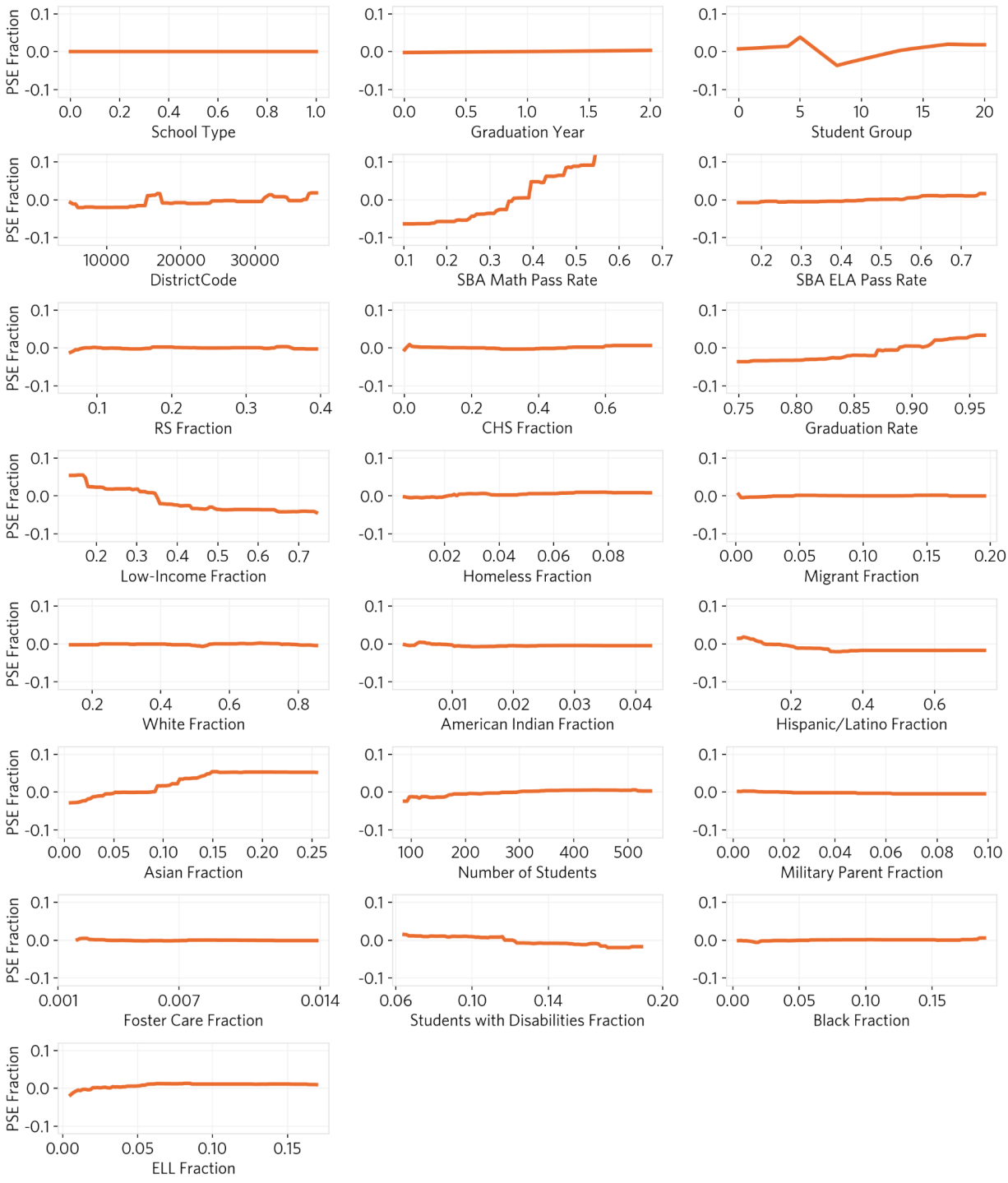

Figure 4. Partial dependence plots illustrating the relationship between each numerical model feature and the postsecondary fraction (while marginalizing over all other features).

Sensitivity Analysis

To answer our fundamental research question regarding the impact of DE participation on postsecondary enrollment, we employed the following bootstrapping strategy:

1. Build/train an enrollment model using 80% of the available data (data chosen randomly with replacement)
2. For every school-year student group in the model with DE data (this step was repeated for both Running Start and College in the High School):
    a. Predict postsecondary enrollment under the current DE participation rate
    b. Predict postsecondary enrollment with an DE participation rate increase of 10%
    c. Take the difference of b) and a) and divide by 0.1 to calculate our *impact coefficient*: the effect of a unit change in DE participation on postsecondary enrollment.
    d. Determine the average impact (step c) across all school-groups.
3. Repeat steps 1 and 2 one thousand times.

We can use the resulting data to gauge how sensitive our impact coefficients are to the specific data used to build our model. Histograms of the estimated impact coefficients are shown below. The mean impact coefficient for Running Start is 0.074, and that for College in the High School is 0.019. The practical interpretation is that a 10% increase in Running Start participation would lead to a 0.74% increase in postsecondary enrollment, while a 10% increase in College in the High School participation would lead to an 0.19% increase in college enrollment.

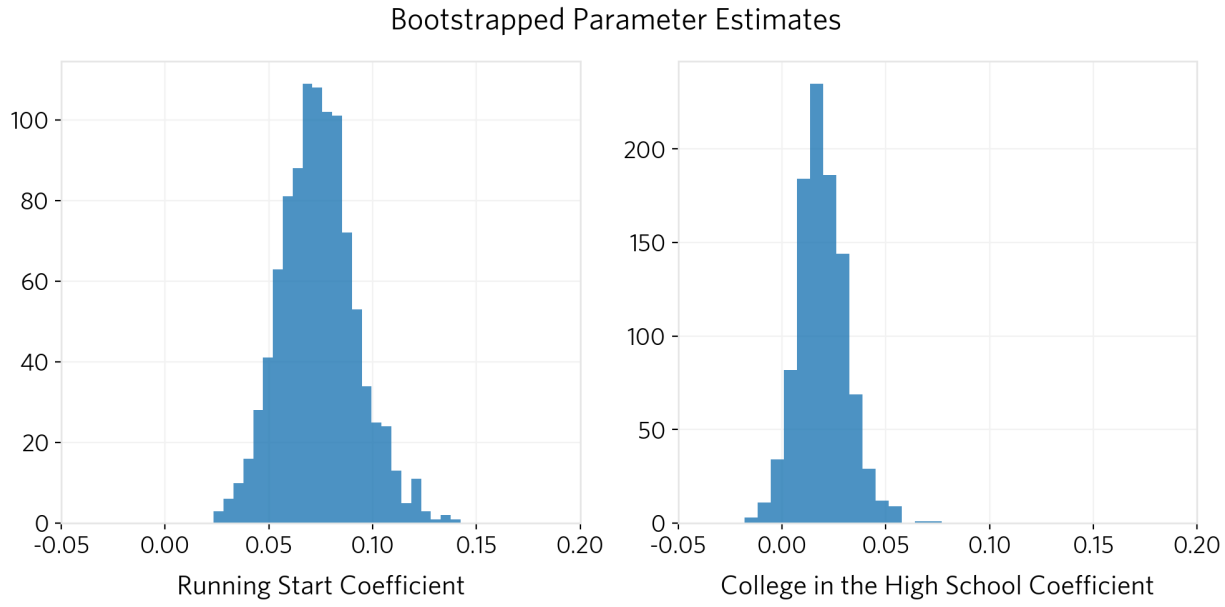

Figure 5. Histograms of the impact coefficients for Running Start and College in the High School, created by bootstrapping our model with an 80% hold-out ratio and 100 repetitions.

Comparisons to Other Impact Factors

While the average impact coefficients for Running Start and College in the High School are unambiguously positive, it is natural to compare them to impact coefficients of other model variables. For example, we can also estimate the impact that increasing math skills (via the SBA Math Percent Pass variable) would have on postsecondary enrollment. The impact coefficient for SBA Math Percent Pass is 0.186, more than twice that of Running Start, which is consistent with our observations in the raw scatter plots (Figure 1), correlation matrix (Figure 2), and partial dependence plots (Figure 4).

4-year vs. 2-year Enrollment

In addition to modeling *any* enrollment in postsecondary as an outcome, we can create separate 4-year enrollment and 2-year enrollment models.

The table below summarizes the impact coefficients across all of the postsecondary outcomes analyzed. Each of the coefficients are calculated independently from the model data, and we do not introduce any constraints to force the "any postsecondary" coefficient to equal the sum of 2-year and 4-year coefficients.

	Running Start %	College in the High School %	SBA Math Pass Rate
Any postsecondary	0.074	0.019	0.186
4-year	0.027	0.008	0.211
2-year	0.057	0.020	-0.030

Figure 6. Impact coefficients for three model variables (Running Start Participation, College in the High School participation, and SBA Math Pass Rate) for 3 outcomes. The outcomes are enrollment fractions in any postsecondary institution, in 4-year institutions, and in 2-year institutions.

A few details stand out from the table. The first is that the impact coefficients for each model variable and "any postsecondary" are roughly the sum of the 4-year and 2-year impact factors. This demonstrates that our model results are consistent, even though they were built and trained with 3 separate outcome variables. The table also shows that the majority of postsecondary enrollment gains for DE programs come from increasing 2-year enrollment: the 2-year enrollment impact factors are more than twice that of the 4-year enrollment impact factors. This is in contrast to the impact of increasing student math proficiency, which shows very large gains in 4-year enrollment, and a slight *decrease* in 2-year enrollment (due to some shifting of student enrollment from 2-year to 4-year institutions).

The size of the impact factors has important implications for policy decisions involving DE programs. Generally speaking, Running Start has a much larger impact on postsecondary enrollment than College in the High School. Furthermore, both DE programs are significantly more effective at increasing 2-year enrollment than at increasing 4-year enrollment. Finally, there are other factors (such as math proficiency on state standardized tests) that can have much larger overall impacts on postsecondary enrollment than the DE programs studied here. All of these factors should play a role in deciding where to most effectively allocate resources aimed at increasing postsecondary attainment for students in Washington.

Further Work

There are several natural extensions to the work discussed in this report. The first is to extend the analysis to include more years of DE data: we strongly recommend that our analysis be repeated with additional years of DE data. This should greatly increase the generalizability of the findings.

In this report, we've focused on *average* impact factors, which involves averaging results over all of our available student groups. However, our approach can be used to disaggregate impact factors: in other words, it can be used to see how the impact of dual enrollment may vary by different student groups. It seems likely that DE programs will have meaningful differences in impact for different student groups, and further research here could facilitate the design of more targeted programs to increase postsecondary attainment.

Finally, replicating the analysis in this report using student-level data could provide useful insight into the extent to which conclusions derived from aggregate group level data may differ from those derived from student-level data.

References

College in the High School Frequently Asked Questions. (2019). College in the High School Frequently Asked Questions. Retrieved September 01, 2021, from https://www.k12.wa.us/sites/default/files/public/ossi/k12supports/pubdocs/CHS%20FAQs%208.20.19.pdf

Data.WA.gov - the general purpose open data portal for the State of Washington. (n.d.). data.wa.gov. Retrieved September 17th, 2021, from https://data.wa.gov/

Dual Credit Task Force Agenda. (2020). Dual Credit Task Force Agenda. Retrieved August 31, 2021, from https://wsac.wa.gov/sites/default/files/2020-11-18-03-Dual.pdf

Jacob, R., Goddard, R., & Kim, E. S. (2014). Assessing the Use of Aggregate Data in the Evaluation of School-Based Interventions: Implications for Evaluation Research and State Policy Regarding Public-Use Data. *Educational Evaluation and Policy Analysis*, *36*(1), 44-66. https://www.jstor.org/stable/43773451

Light Gradient Boosting Machine. (2021, April 12th). Github: LightGBM. Retrieved September 6th, 2021, from https://github.com/Microsoft/LightGBM

Miller, M., Boatwright, J., & Mahoney, K. (2019). *Covering the Costs of Dual Credit for Students and Families*. Covering the Costs of Dual Credit for Students and Families. Retrieved September 7th, 2021, from https://www.k12.wa.us/sites/default/files/public/communications/2019-11-Covering-the-Costs-of-Dual-Credit.pdf

Partial Dependence and Individual Conditional Expectation plots. (2021). SKLearn Documentation. Retrieved September 6th, 2021, from https://scikit-learn.org/stable/modules/partial_dependence.html

Running Start 2005-06 Annual Progress Report. (2006, December). Running Start 2005-06 Annual Progress Report. Retrieved September 7th, 2021, from https://files.eric.ed.gov/fulltext/ED496209.pdf

Smith, K. (2014, May). *Access and Diversity in the Running Start Program: A Comparison of Washington's Running Start Program to Other State Level Dual Enrollment Programs Hosted on a College Campus*. Access and Diversity in the Running Start Program: A Comparison of Washington's Running Start Program to Other State Level Dual Enrollment Programs Hosted on a College Campus. Retrieved September 7th, 2021, from https://wsac.wa.gov/sites/default/files/2014.04.24.052c.Report.Running%20Start.pdf

Zinth, J., & Barnett, E. (2018, May). *Rethinking Dual Enrollment to Reach More Students*. Rethinking Dual Enrollment to Reach More Students. Retrieved September 7th, 2021, from https://www.ecs.org/wp-content/uploads/Rethinking_Dual_Enrollment_to_Reach_More_Students.pdf

Appendix

Model Accuracy Metrics for 4-year and 2-year enrollment models

To assess the predictive power of 4-year and 2-year college enrollment models, we employ the same process described for the overall postsecondary enrollment model. The results are shown in the two tables below.

We can see that while the 2-year enrollment model has smaller absolute error than the 4-year model, the 4-year model has a significantly higher R^2 value. This is because the enrollment rates for 2-year colleges have a much smaller range (typically 10-40%) than those for 4-year colleges (which can range from 5%-80%).

Accuracy Metrics for "Enrollment in 4-year Institution" Model

Mean Absolute Error	**Median Absolute Error**	**R^2**
0.039	0.030	0.857

Accuracy Metrics for "Enrollment in 2-year Institution" Model

Mean Absolute Error	**Median Absolute Error**	**R^2**
0.032	0.023	0.653