Matt Nelson

Final Report:
Teaching Analysis

Problem Statement

As we are approaching the second half of the third year of public education after a global pandemic, it would be appropriate to take a look at where we've come. I am in a special situation for my current high school, Mt. Scott Learning Center (MSLC), because my first year, we and the rest of the world closed down before the completion of the third quarter of the 2019-2020 school year. Then, we returned to teaching through distance learning in the 2020-2021 school. Finally, for the 2021-2022 school year, we came back to in-person instruction. We have had two years to get back on track, to find our educational footing.

Yet, since lockdown, there has been a reckoning of learning loss, inequitable teaching, chronic absenteeism, and grade inflation just to mention a few. But much of the news is sensationalized and misrepresented. MSLC is an alternative high school in Portland, OR, that takes in students who don't fit the mold. Where I teach, the students have already been pushed out of mainstream schools for skipping or fighting or smoking, or have self-selected to our smaller environment in order to succeed in a school with a higher teacher to student ratio. It is my goal to swim within my local data pool to find a way to predict student outcomes in order to better serve my students, those who have already been mistreated or under served. I am also interested in analyzing student outcomes as a means of professional reflection. I also aim to find some trends in my students' progress that can help me be a better educator.

Data Wrangling

My first data set was actually a collection of datasets acquired from the office manager and registrar of the office. Much of the analysis could not have been completed without her help. There were 15 csv files consisting of each quarter of each year, minus Q4 from 2019-2020 year due to it being missed during lockdown. After I merged all the datasets, the final product was 1601 rows of students with 9 columns

(Course Title, TermCode (Quarter), Period (1-8), Teacher, Abs1 (days Absent), SISNumber (instead of student name), Grade, Comment1, and SectionID. I immediately created a new column for the school year, as well as one for my teaching year at MSLC, and dropped the Comment1 and SectionID columns. This data is a closed loop. Because it is not gathered by individuals or an outside source, every row should have been complete. But after doing a search for null values, I found that one person was missing a grade and deleted that row. This brings up a fact of my dataset that is compounded by the nature of MSLC. Since we are an alternative pathway to graduation, many of our transfer students come after the first quarter because they haven't yet been pushed out. We also lose a fairly large percent of our school due to it not being the right fit for students. So even though the dataset contains 1600 student recorders, there have been many more students that did not make it or entered the quarter too late to be included in the gradebook.

I then added a race and zip code dataset. This dataset was again furnished by our office manager, but was a hardcopy. I had to create the digital version. It is 332 rows with 3 columns, SIS Number, Race, and Zip. There are 6 racial categories provided by the school: White, Native Hawaiian/Other Pacific Islander, Asian, Black/African American, Two or More, Hispanic, and White. I created a preliminary graph to see the count of students based on race (Figure 1), but I also wanted to see how our school's demographics stacked up against the rest of the school district (Portland Public Schools (PPS)) and the two nearest feeder schools (Cleveland High School (CHS) and Franklin High School (FHS). MSLC has a 5.5% increase in the percentage of students who are White than the whole District, as well as a 2.9% in the percentage of students who are Hispanic (Figure 2); compared to CHS, MSLC has a lower percentage of students who are White and students who are two or more races, and MLSC has a higher percentage of students who are Hispanic and students who are Black, whereas, compared to FHS, MSLC has a higher percentage of students who are White, students who are Hispanic, students who are two or more races, and students who are Black; both CHS and FHS have a much higher percentage of students who are Asian (Figure 3).

Later I created another dataset that was created from my gradebooks for just math classes that had 504 rows, but included the Grade Percent column. A feature

which I would need later for regression models. One final dataset that was useful for creating maps was a collection of 2021 inflation-adjusted median income from the Census for each of zip codes present in the student databases.

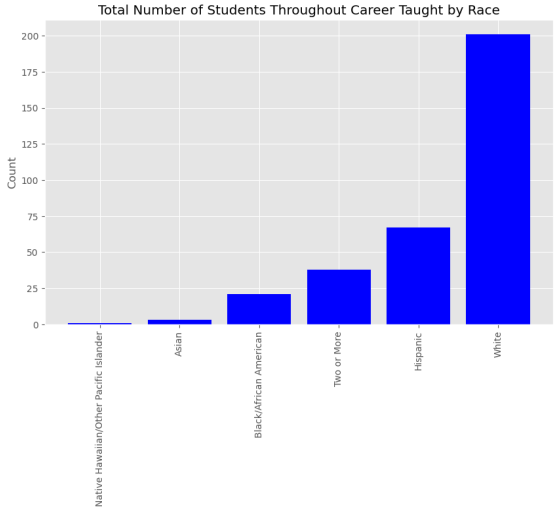

Figure 1: Count of Students by Race

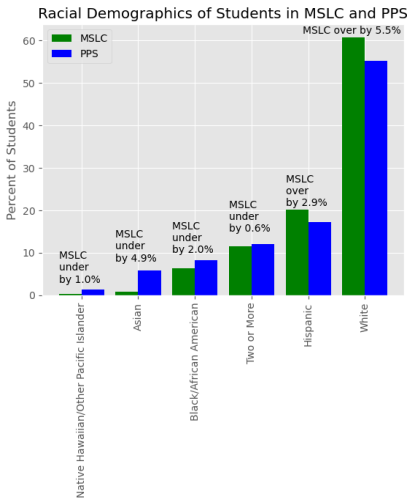

Figure 2: Racial Demographics of Students at MSLC vs PPS

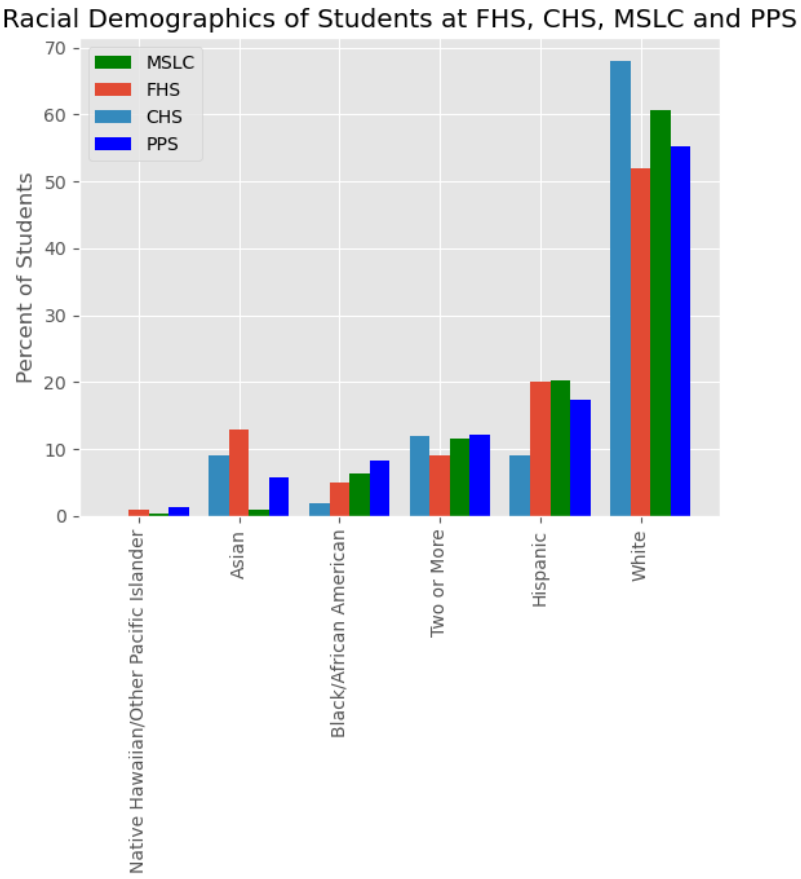

Figure 3: Racial Demographics of Students at FHS, CHS, MSLC, and PPS

Exploratory Data Analysis

To begin the EDA, I need a few more columns for analysis. The first was a Credit/No Credit (C/NC) binary column that would be useful for my classification models later. Any grade that was a P (for pass), A, B, C, D received a 1, and any other grade received a 0. I also converted the normal A, B, C, D, F scale to a numeric scale of 4 - 0 in another column titled GPA. After that, I started exploring the distribution of grades over my career as a teacher. I found that I gave more As than Bs, more Bs than Cs, and more Ds than Cs (Figure 4). I also found that my students were 8 times more likely to receive credit than not (Figure 5).

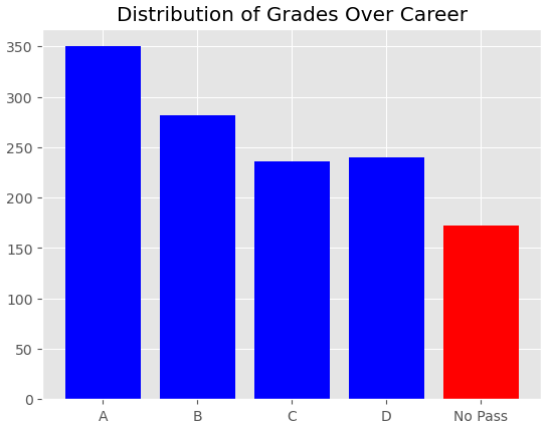

Figure 4: Count of Grades over career

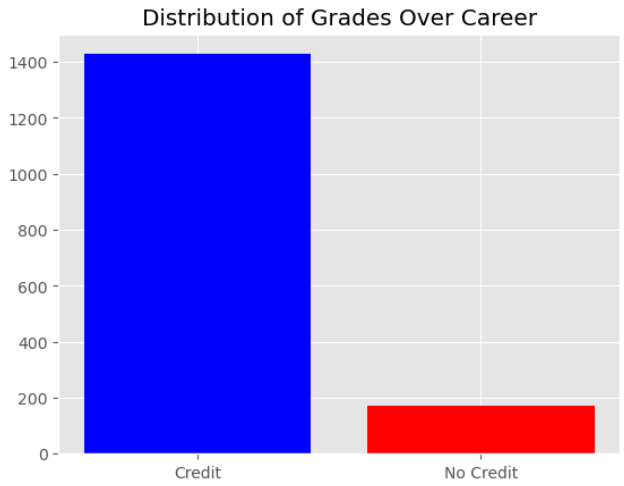

Figure 5: Count of Credit vs No Credit Awarded

I continued to do some rudimentary EDA, checking to see what the grade breakdowns were per race, class, periods, quarters, and years, and also creating a new column Math/Not Math (M/NM) to delineate between math classes and electives. What I found was interesting.

Because the number of students who are Asian or Native Hawaiian/Pacific Islander is so low, I want to focus my analysis on students who are White, Black/African American, Hispanic, or two or more races. I found that I gave the highest percentage of As to students who are Black/African American (38.6%), the highest percent of Bs to students who are Hispanic (26.2%), the highest percent of Cs to students who are two

or more races, and the highest rate of Ds to students who are Hispanic (24.3%) (Figure 6).

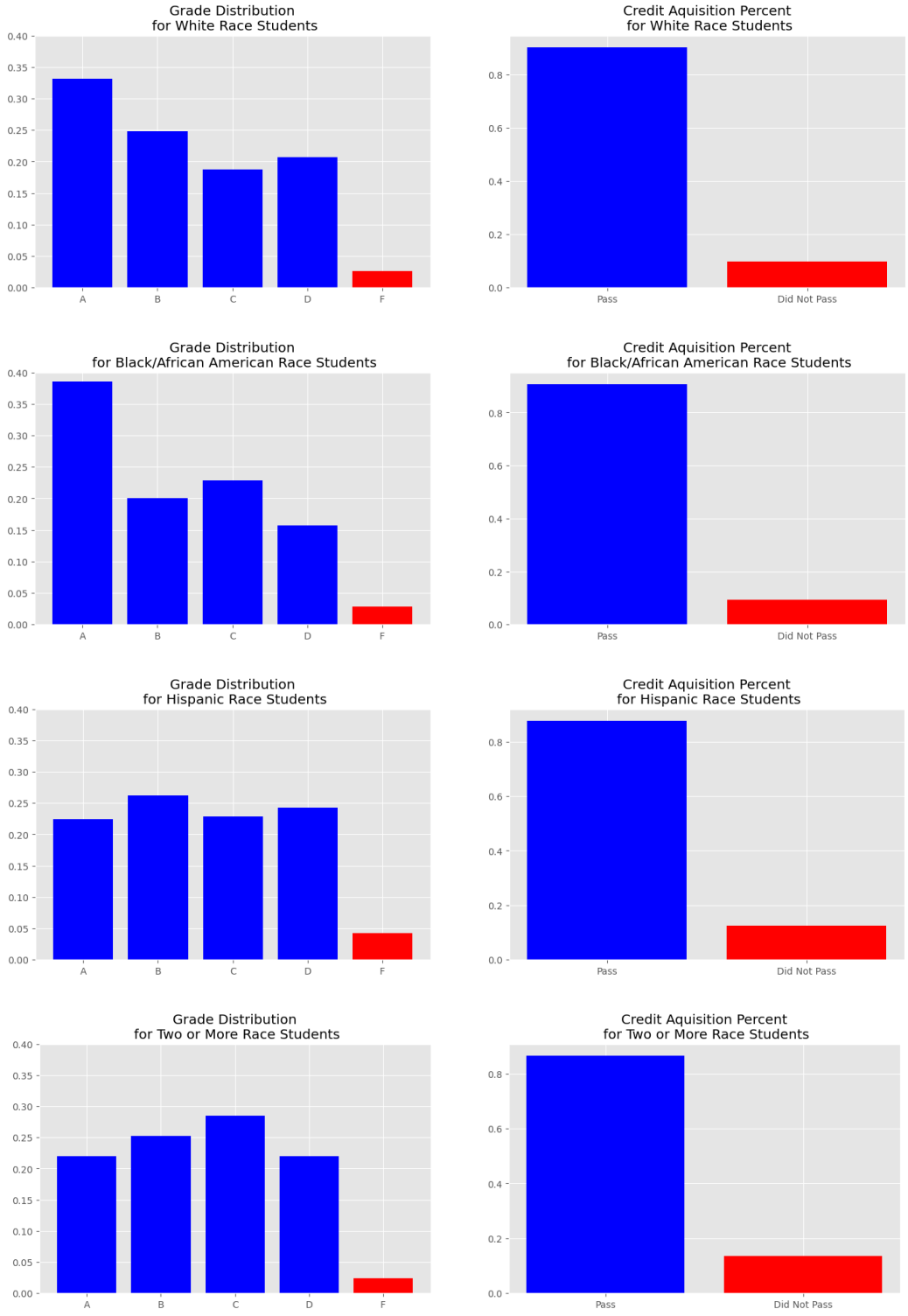

Figure 6: Count of Grades and Credit vs No Credit Awarded for White, Black, Hispanic, and Two or More Races

I also found that students have steadily been getting more As (Figure 7), third quarter is maybe the hardest quarter (Figure 8), periods 5 and 6, those that bookend lunch, have some of the lowest grades (Figure 9), and students perform worse in Math classes (Figure 10).

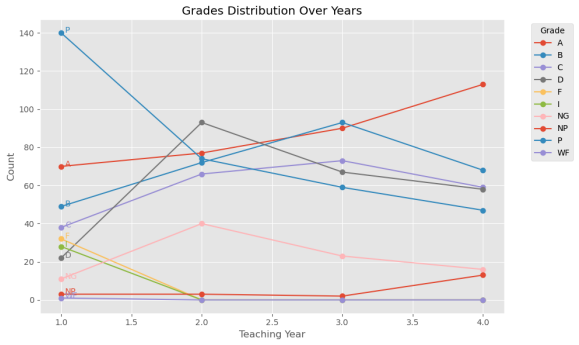

Figure 7: Count of Grades by Year

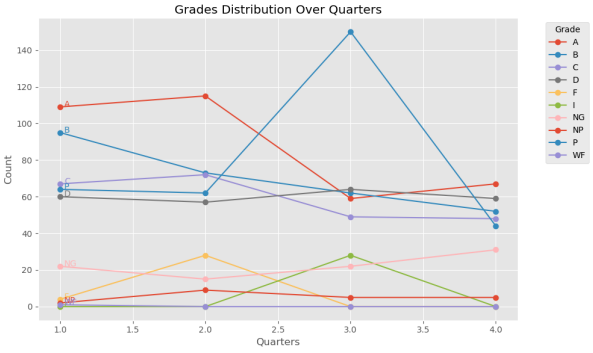

Figure 8: Count of Grades by Quarter

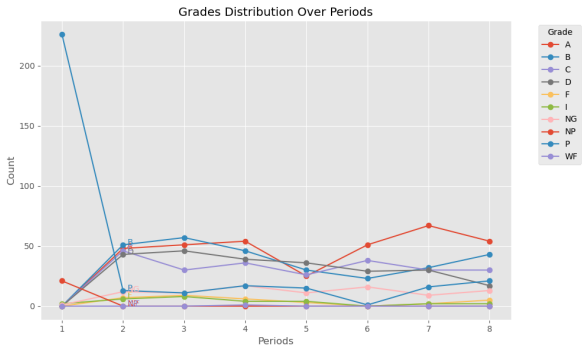

Figure 9: Count of Grades by Period

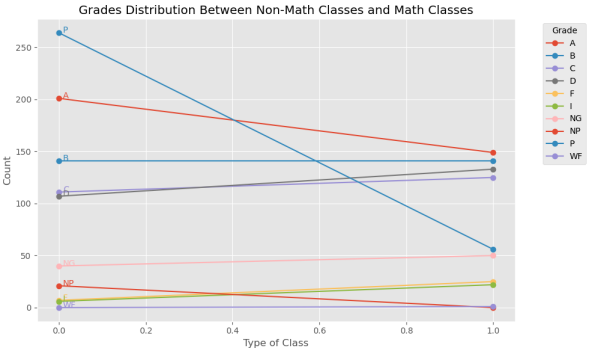

Figure 10: Count of Grades by Non-Math vs Math Class

Before I did a correlation matrix, I had to fix the Zip Code column to exclude the extra four digits. I also created another column, thanks to data gathered by, you guessed it, the office manager, of the number of days in specific quarters so that I could create a days absent ratio column. Some quarters don't just feel longer, they are longer! Here are some of the interesting correlation matrices. From the initial correlation matrix, the most striking numbers are for teaching year and credit (0.14) and GPA (0.11), and for days absent and credit (-0.30) and GPA (-0.47) (Figure 11). After this initial matrix, I created dummy variable for race and zip code so that they could be included, and didn't find anything too strong for race and credit, and there were too many zip codes to make a real conclusion. But then I made another correlation matrix between the different races and the number of days absent. I found that there was a 0.12 coefficient for

Black/African American and days absent, while there was a -0.11 coefficient for White and days absent (Figure 12).

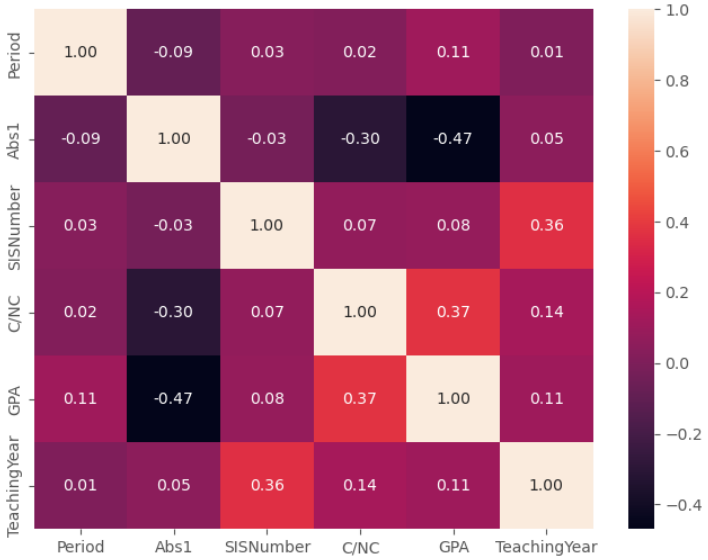

Figure 11: Correlation Matrix of Initial Numerical Values

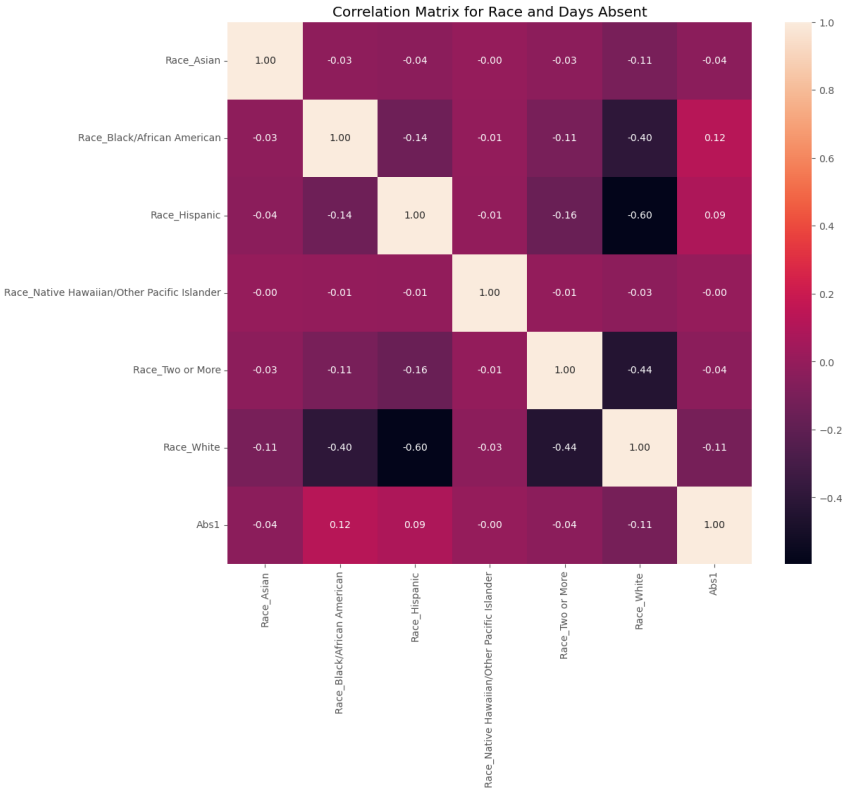

Figure 12: Correlation Matrix For Race and Days Absent

The final, and maybe most striking find was the average days absent per grade. I

found that for every drop in grade level, a student would have to miss about 4-5 days, on average (Figure 13). This is powerful information that I can definitely bring back to my students and admin.

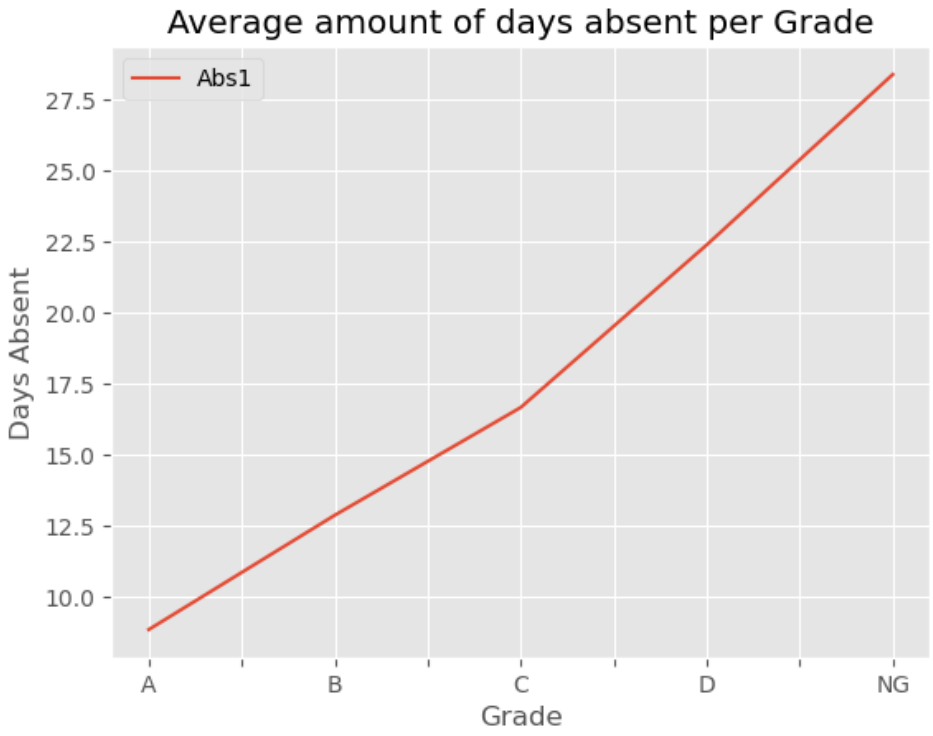

Figure 13: Days Absent and Grade

In-depth Analysis

To pre-process my data, I used one-hot encoding on the Course Title and School Year features. My target dataset was the Credit/No Credit column. I checked once more for null values but found none except in the median income zip code column I added earlier for mapping purposes. I decided to drop that whole column. I then split my features dataset into train and test sets with a 20% test size and got ready to model. I decided against standardizing the data since I was going to use classification models.

Model Selection

I decided to try out a few different supervised classification models: SciKit Learn's Logistic Regression, a Decision Tree Classifier, a Logistic Regression with an L2

penalty, a Random Forest Classifier, and an XGBoost Classifier.

My first Logistic Regression has an accuracy of 0.925 with a Pass-classification precision of 0.94 and recall of 0.96. But the fail-classification precision was 0.67 with a recall of 0.29. This is to be expected since there were so few no credits. Still, 94% of the students classified as passing were accurately picked out, and that 99% of those said to pass really were passing. I looked at the feature importance coefficients and found that the 97233 zip code had a coefficient of -2.03, the 2020-2021 school year had a coefficient of 1.21 and the zip code 97211 had a coefficient of 1.04 (Figure 14). I dug deeper into the zip codes and courses and found that 15 out of 37 zip codes had negative coefficients (Figure 15), 7 out of 11 math classes had negative correlations (Figure 16), and the distance learning year had the highest coefficient (Figure 17).

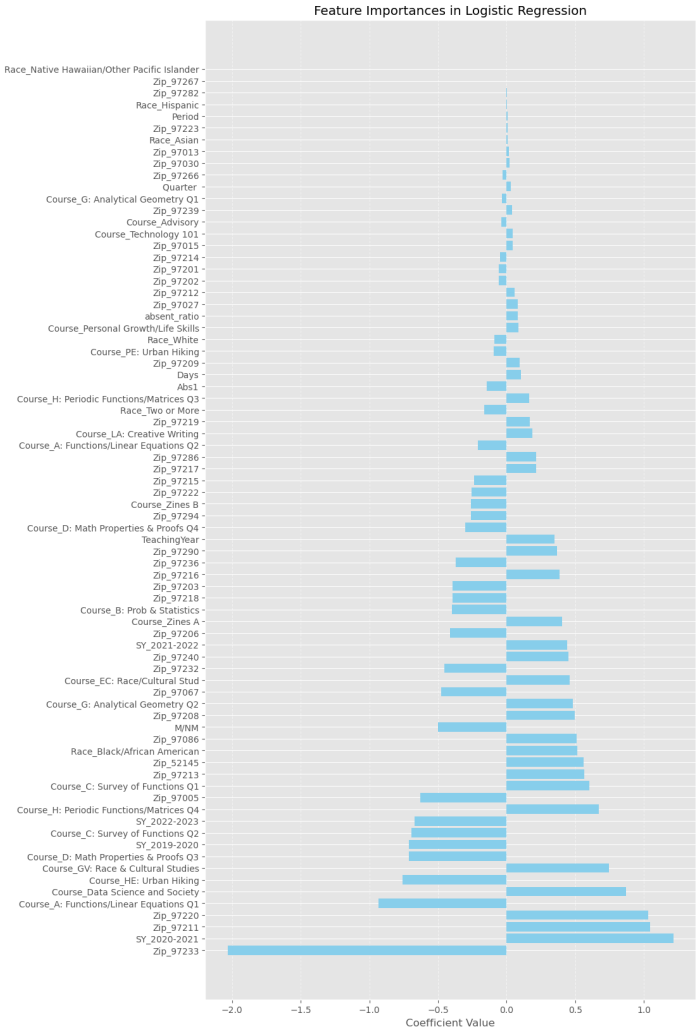

Figure 14: Feature Importances for the Logistic Regression

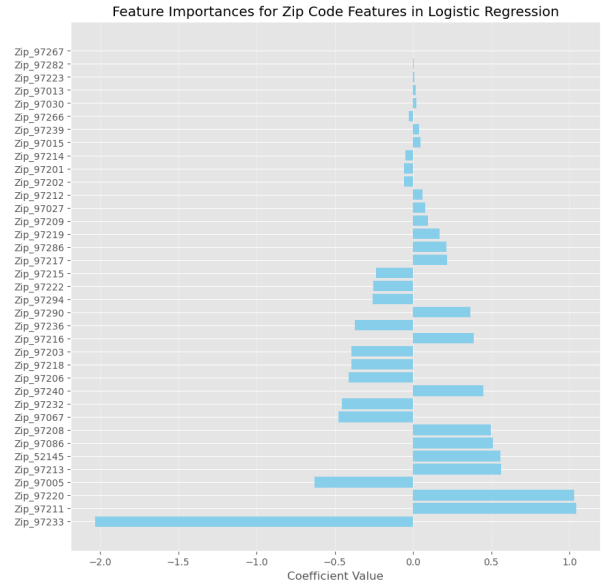

Figure 15: Feature Importances for the Logistic Regression by Zip

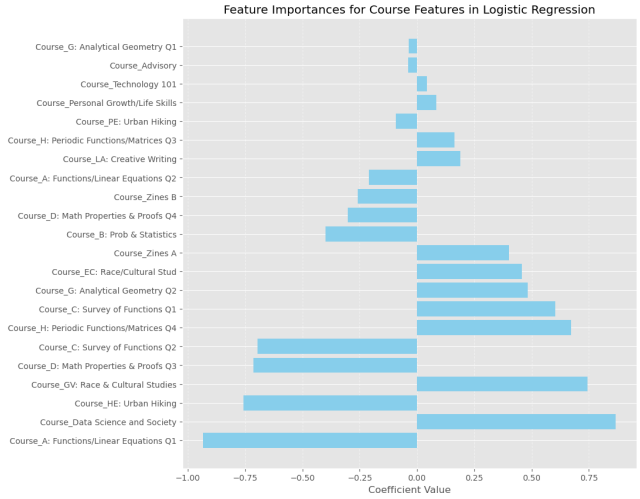

Figure 16: Feature Importances for the Logistic Regression by Course

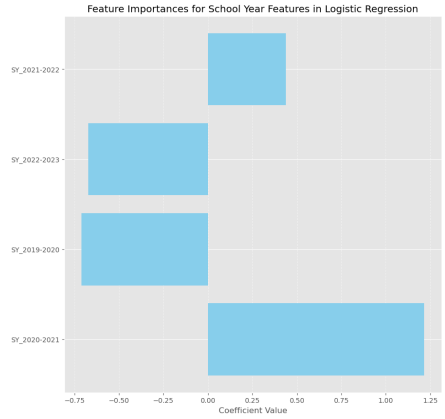

Figure 17: Feature Importances for the Logistic Regression by School Year

I decided to test the AUC ROC for the Logistic Regression and got 0.64 which is not great (Figure 18). That's when I decided to try PCA, by scaling all of my feature columns first, and then discovering the number of features that explained 100% of the variance. Turned out to be 69 out of 76 (Figure 19).

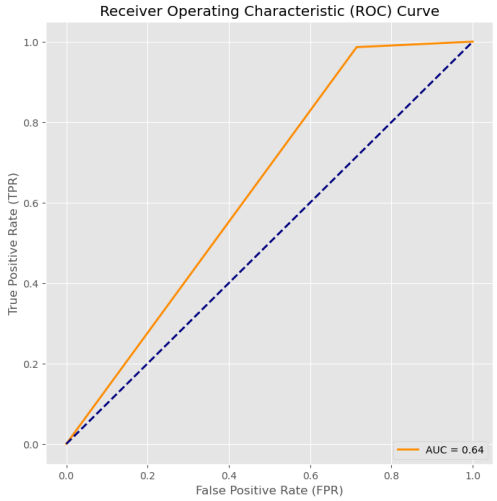

Figure 18: AUC ROC of Logistic Regression

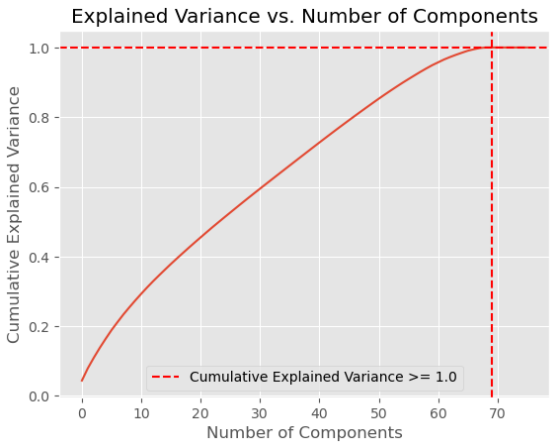

Figure 19: Cumulative Explained Variance for Features

At this point, I was ready to try some other models. I re-split the data into train and testing sets, then instantiated an out of the box Decision Tree Classifier. It had an accuracy of 0.8625 and lower precision and recall than the Logistic Regression for both classifications. I used the scaled data on the Logistic Regression model and compared the Decision tree and though the Decision Tree had less false positives, it also had much less true positives than the Logistic Regression (Figure 20).

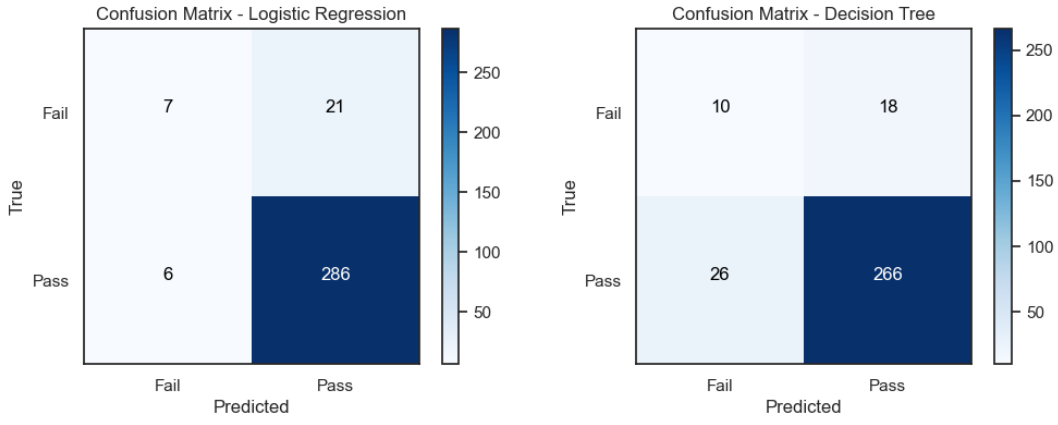

Figure 20: Comparison of Confusion Matrices for Logistic Regression and Decision Tree Models

At this point I tried an L2 penalty on my Logistic Regression because with the scaled data, the accuracy had fallen to 0.916, but the penalty had no effect.

I then tried a Random Forest Classifier, but it too had no greater accuracy (0.916) and had worse precision and recall. I decided to do a Grid Search to tune the hyperparameters, find the best parameters to be max_depth = None, min_sample leaf=1, min_samples_split =5, n_estimators=100, but there weren't substantial gains.

My final model was an XGBoost Classifier, which I started out the gate with a Grid Search, finding the best hyperparameters to be learning_rate = 0.1, max_depth=5, and n_estimators=50. The accuracy was 0.919, but the recall was slightly better than the Logistic Regression (Figure 21).

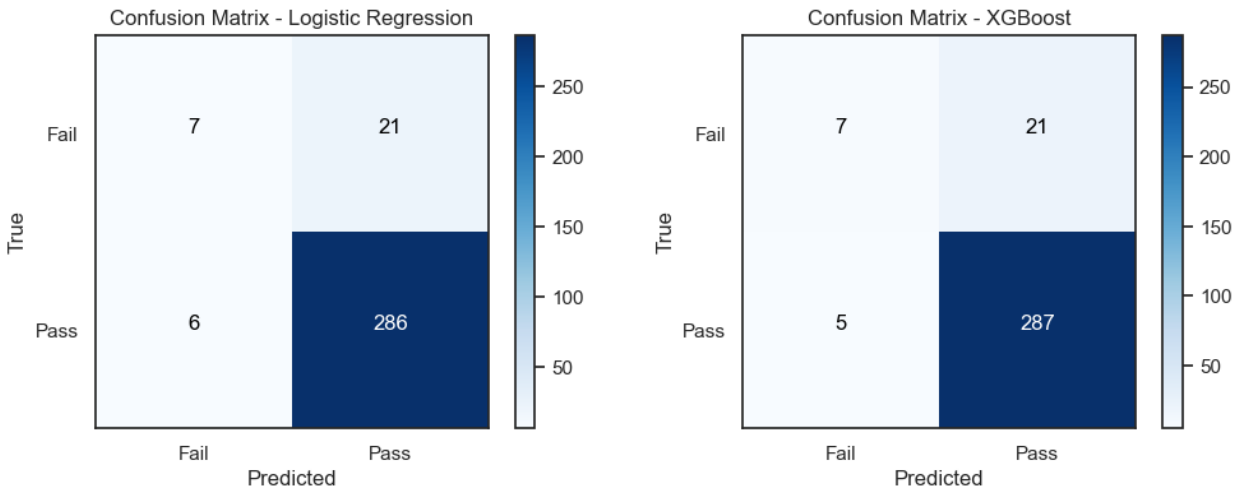

Figure 21: Comparison of Confusion Matrices for Logistic Regression and XGBoost Models

The feature importances were different for the XGBoost model than the Logistic Regression. The most important features where the zip code 97202, the teaching year, the zip code 97233, the absent ratio, and math/non-math classes (Figure 22).

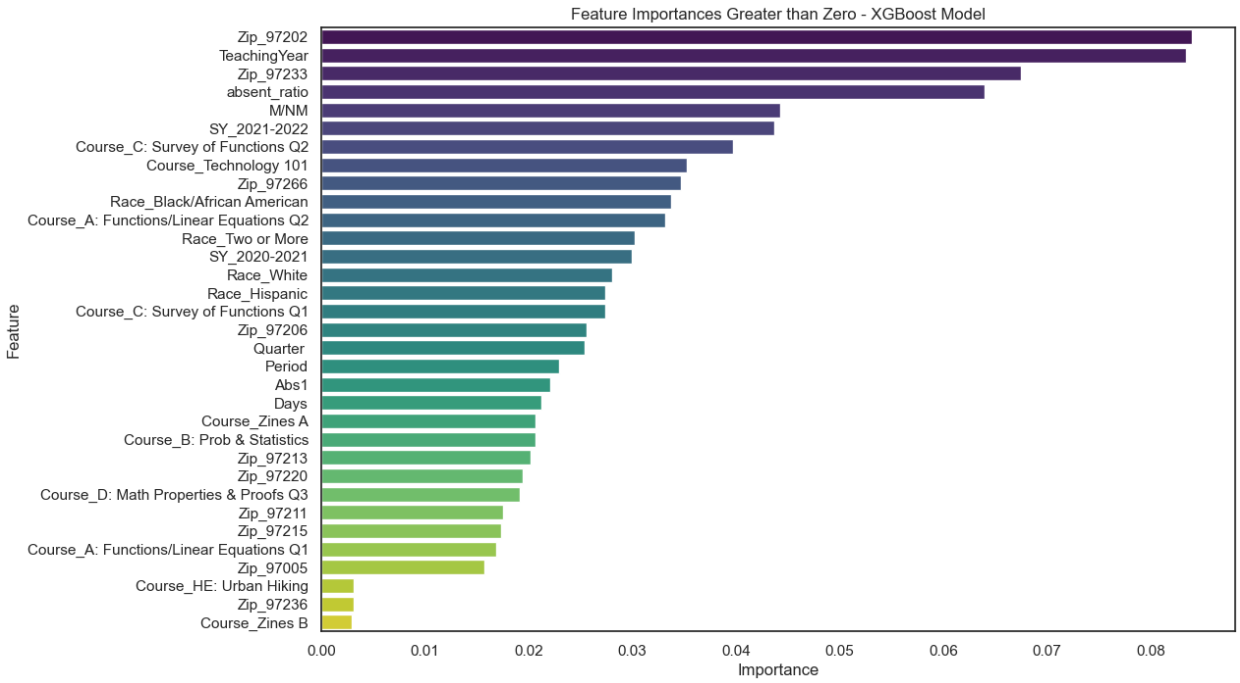

Figure 22: Feature Importances for XGBoost Classifier

Takeaways

Although the best model was the XGBoost, I found the Logistic Regression model to be impressive. That said, for future reference, the dataset was very imbalanced. I could try to resample with over-sampling the failure class (C/NC = 0). But from both of the best models, I would like to look more into the different zip codes that showed higher feature importances like 97202, 97233 for the XGBoost, and 97233, 97211, and 97220 for the Logistic Regression. I wanted days absent, or better yet, absent ratio, to be a bigger feature for Logistic Regression since that is the only feature that students have (some) control over. At least for the XGBoost it was the 4th largest feature importance.

Another takeaway was that I didn't have enough data for a great Linear Regression model. I tried to make a simple model with just the days absent ratio as the feature and the grade percent as the target and I had a Mean Squared Error of 268 and a R^2 of 0.11 (Figure 23). I feel with more data from all of my classes, or even all of the classes from the school, I could make a much better model.

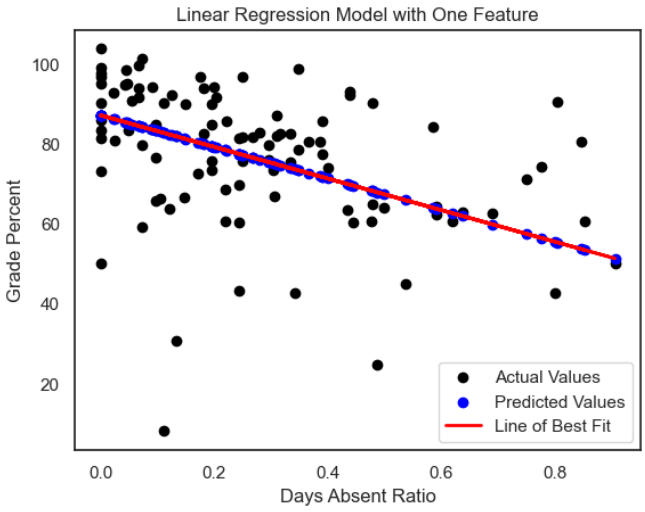

Figure 23: Linear Regression Model for Days Absent Ratio and Grade Percent

Future Research

There is so much that could be done with this type of work. I would have to tease out some more relationships between different variables like race and absent ratios, or courses or periods and absent ratios. The combinations of possible inquiries are large. Like, are there any other factors about how well a student does in classes that bookend lunch? I could make another feature that is afternoon versus morning class. But what I really need is more data. Both for department analysis and whole school investigations. Once there's more data, I think the models will become a little more balanced–especially if I could find a way to include students that drop-out or who lose their spot at our school.

In terms of modeling, I also want to try and make a neural network of the data. I feel like deep learning is the next stage to find the best predictor of student achievement.