ISyE 6740 – Summer 2024
Project Report

Team Member Names: Walter J. Simmons

Project Title: Analyzing Teacher and Student Performance On College Credit Examinations

Problem Statement:

At the conclusion of the school year in May of each year, high school students around the world take examinations to hopefully earn college credits. The performances across each of these tests are used to evaluate both student and teacher performance throughout the year. Unfortunately, these metrics are often reduced to the most basic of outcomes and used as evaluation such as: "did the average score go up?" or "did you get the score required to get credit?" This kind of analysis misses some much larger trends, and teacher and student performance can both be more specifically measured using more advanced metrics. Teachers and students both rely on specific feedback to improve their own learning and instruction, and teacher performance evaluations are often simply based on standardized testing scores as a unique average and a 30 minute observation.

By and large, teacher performance is almost never evaluated with anything other than brief observational data and the above "robust" analysis. Darling-Hammond describes the necessity for highly qualified teachers as "One of the few areas of consensus among education policymakers, practitioners, and the general public today" and that "improving teacher quality is one of the most direct and promising strategies for improving public education outcomes in the United States" (Darling-Hammond). Further, that "[there is not] a set of widely available methods to support the evaluation and ongoing development of teacher effectiveness".

Other studies have been inconclusive about the efficacy of this data for teachers. "…we cannot know the extent to which this reflects true fluctuations in performance or changes in class or school dynamics outside of a teacher's control" (Goldhaber). But, utilizing this data, specifically for the use of teachers, who could see more of those true fluctuations, educators can make much more informed decisions about their practices.

This study utilizes several more sophisticated analytical models and techniques in order to classify students test scores based on their grades in class. These models can help address the correlation between in class scores versus standardized testing. This study also analyzes year-to-year teacher performance, using a value-added model.

This project utilizes several modeling and data processing techniques to evaluate a sample teacher and classifies students into one of five groups (scores). The best performing models were a Random Forest and Neural Network, in which 73.6% of students were accurately classified into their correct score, and 93.5% of students were correctly classified as Pass/Fail.

Data Source

The college credit exam that I teach has a scoring system of 1-5. Generally, a score of 3 is considered a passing grade, a score of 4 is a high pass, and a score of 5 is mastery. Scores of 1

and 2 are functionally the same, and both typically earn no credit among any college campus. 3, 4, and 5 grades award credit depending on a multitude of factors such as university, major, and course. For example, Georgia Tech awards credits for 4s and 5s for many subjects, but this particular subject only accepts 5s (Georgia).

Included in score reports are breakdowns of performance by learning category. This assessment, combined with the appellate scoring from the previous three years was used to assess teacher performance year-to-year. Our student scoring dataset has 18 feature vectors for 155 entries. Shown below, the clusters are not obviously well defined, but determining pass/fail appears viable. The scoring breakdown dataset is 25 entries and 4 feature vectors (though they themselves are aggregate stats).

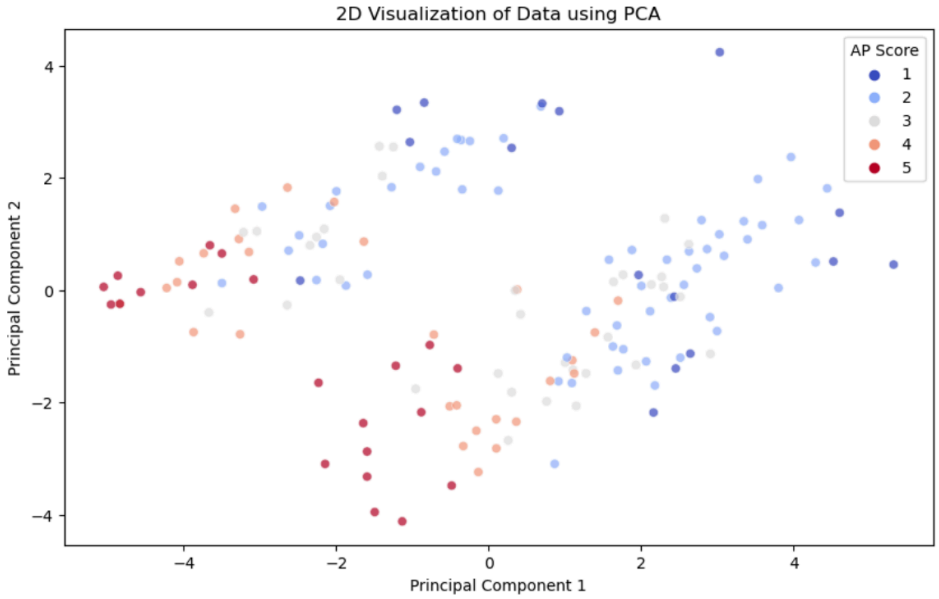

Figure: Visualization of the student dataset after performing PCA

Methodology

The methodology will first be data aggregation. This is largely done in Excel to match grades from year to year that may or may not have the same names. The data preprocessing step is quite extensive. First, a large amount of normalization is done - grades year-to-year may have had different scaling, so grades are normalized as a measure of performance relative to the class year. There are some students with missing assignments as well, so those assignments are imputed with scores based on both their average, and the class average. During this step to account for varying amounts of samples for each score, SMOTE was implemented to tune our model closer to outliers (Brownlee). This did not, however, result in any meaningful performance increase on unseen data.

The classification model was selected based on testing of the following: random forest, kmeans clustering, neural network, multinomial logistic regression, and gradient boosting. Most of these models were selected for testing for their robustness to multiple features (of which, we have a quite a few). Random forests and decision trees are given preferential treatment in analysis, as

their results are often a lot easier to interpret. Since these results would theoretically be presented to non data professionals, ease of understanding is an important factor.

The dataset was partitioned into training and testing and done so randomly so as to not overfit to a particular year. This project implements an 80/20 training/testing split and uses k-fold cross-validation to tune each of the models, and prevent overfitting to a particular set. This is largely since our dataset is not huge. GridSearchCV was implemented to explore numerous hyperparameter tunings to achieve the best performance. The models were each evaluated based on accuracy, precision, recall, and f1-score.

First, we identify student performance metrics year over year, and determine whether our relative performance has gone up, while accounting for changes in total number of students to get an accurate view of if my teaching performance has actually improved. We then evaluate these trends using some simple linear regression. We will then use the classification models to create a value-added model by looking at the residuals between predicted and actual test scores using their scores before entering a course. This is particularly useful to communicate to students, as it can help convince them to trust the process of the classroom and build confidence in their instructor. In general, students respond better when they know their teacher is capable.

This level of data analysis is largely unheard of in classrooms. In most of my team meetings, if we looked at an average it was a big data day. By providing a small framework, this project could hopefully be easily adapted to a variety of subjects and allow for teachers to obtain more robust analysis of their own teaching, and provide justification for pay increases from administration. Because of the lack of specificity, administrators could run a similar analysis and provide much more pointed feedback on teacher growth over time. Analyzing specific units, and particularly teaching strategies year over year, teachers can build upon previous years with much more specific places to focus their instruction to improve student outcomes. Teachers could try implementing different styles of teaching, evaluate them on a year over year trend, and determine the actual effectiveness of their practices.

Results

a. **Classification of Student Scores**

In our goal of predicting student test score based on their classroom grades, we can accurately predict 70.1% of student scores correctly using a neural network approach, and 61.3% accuracy using a random forest model. K-Means, Gradient Boosting, and Logistic Regression each did slightly worse, with gradient boosting achieving the lowest performance. This is reasonably significant, as this is not a simple classification, and we are far outperforming the expected results (20%). This shows that there are some clear indicators of student performance, but differentiating some scores is more difficult.

Shown below are the model accuracies for each for the dataset. We see somewhat similar performance in these models, but a noticeable improvement when using the neural network.

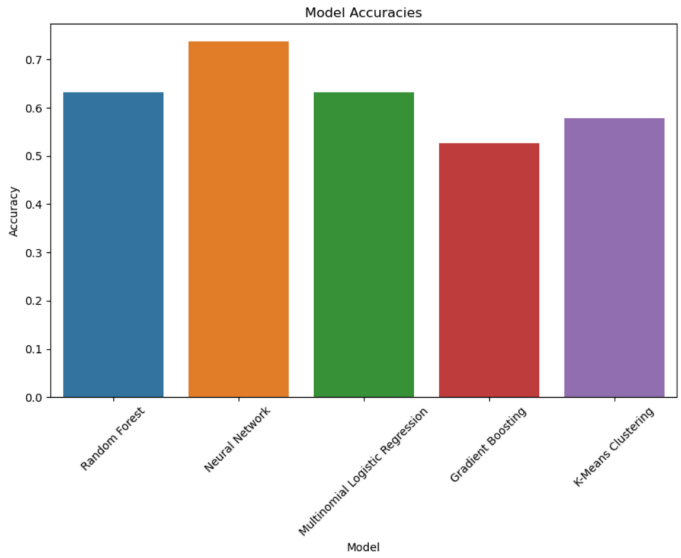

Figure: Model Accuracies for Full Prediction Model

Looking at feature importances to identify areas of improvement, we see some slightly odd results shown in the figures below.

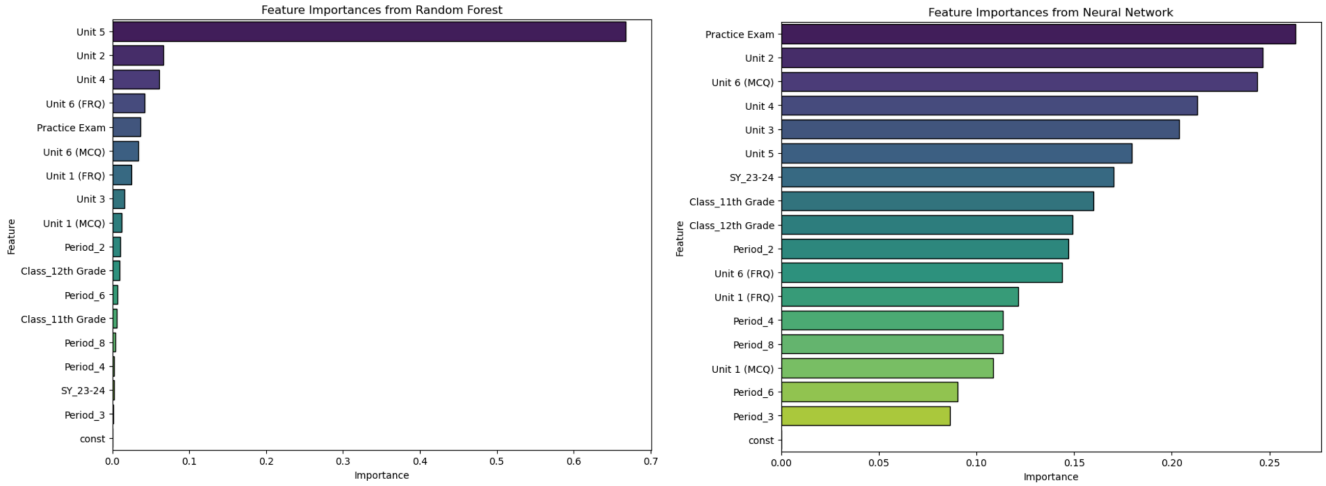

Figure: Feature importances based on each model

The neural network appears to take a more measured calculation based on many of the feature vectors. However, the neural network weights a single test far more heavily than any other feature. This shows that there may be some underlying patterns that account for Unit 5, or the grading on Unit 5 was uniquely predictive, requiring some more personal analysis on the part of the educator. Thankfully, we also show a very low importance on class period or school year. Oddly, we see that FRQ tests are less predictive than MCQ tests overall.

Examining a simple decision tree also yields similar results, as shown below where 80% of students can be classified into a pass/fail based on a single test result. This led to the question of whether or not a simple pass/fail classification might achieve better results, as the difference between a 1 and 2 is negligible, as well as a 3 and 4.

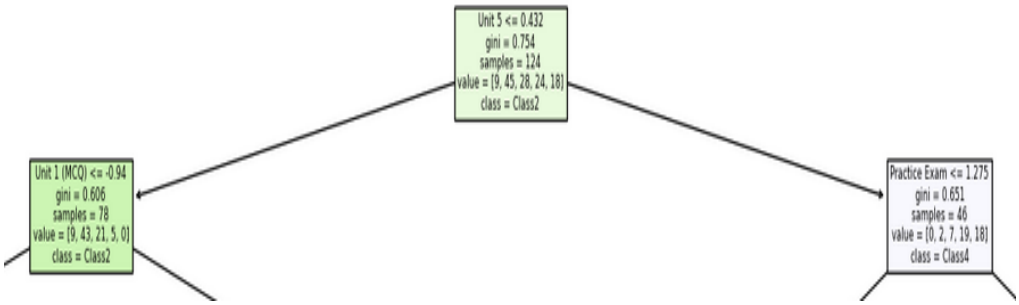

Figure: First layers of simple decision tree

When moving to a simple pass/fail classification model, we observe a marked improvement in classification accuracy, exceeding 90% for a random forest model.

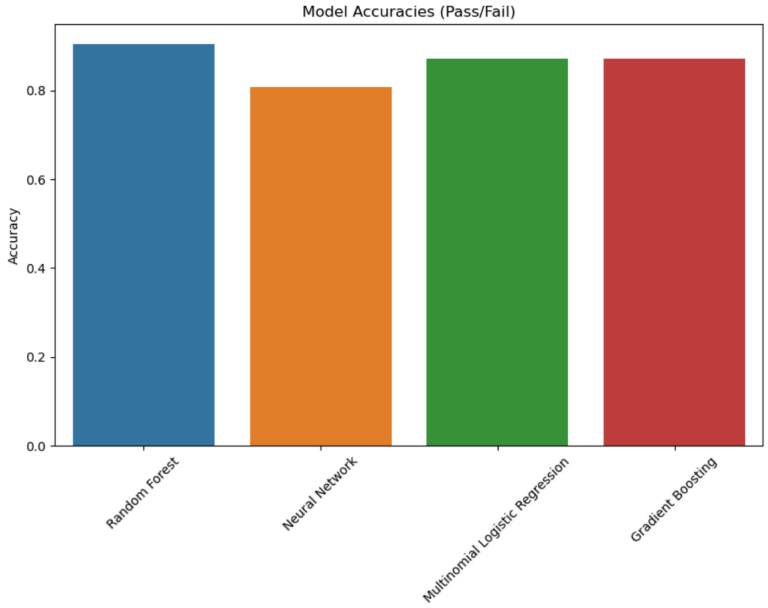

Figure: Pass/Fail Accuracies

This performance is quite good, and exceeds most standards used in educational research. We can reasonably predict the performance of a student on an AP exam on relatively few test grades with great accuracy. We can also identify that some units are significantly more important than others, contrary to published testing breakdowns.

b. Value Added From Teacher

A cursory glance at student data shows a marked improvement from when the instructor took over in 2022 to present in terms of total number of passing students.

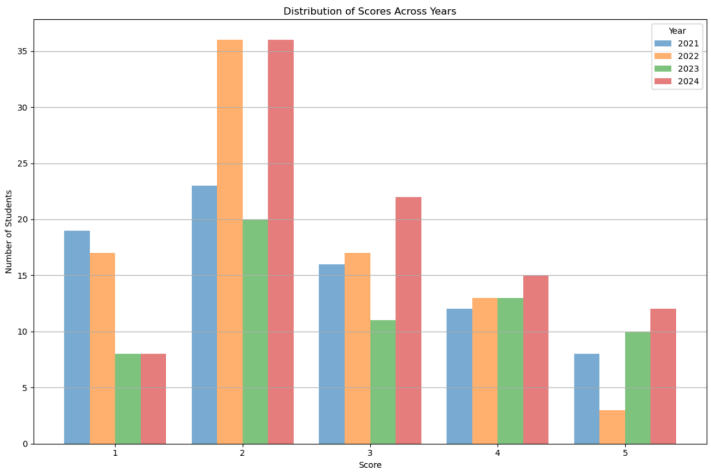

Figure: Student Score Distributions

We can also see that the most predictive factor in our random forest is also one of the instructor's lowest differential score factor, suggesting a change may be necessary in that units instruction.

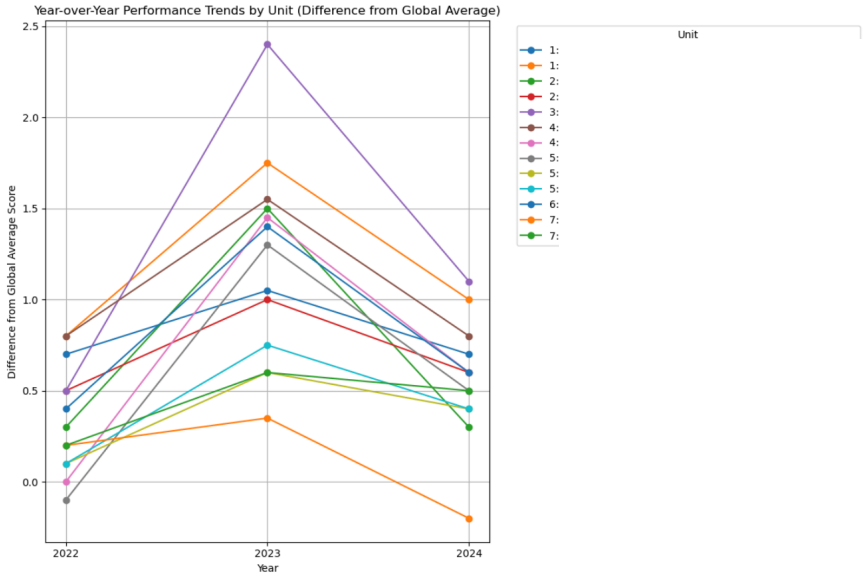

Figure: Performance trends year over year by standard

A value added model for the teacher was implemented using the following formula.

$$VA_{T,j,t} = (\overline{S_{T,j,t} - \overline{S}_t}) * n + N$$

where: $\overline{S_{T,j,t}}$: average test scores for students taught by teacher T in subject j year t

$\overline{S}_t$: global average test score for subject for year t

n: total number of students taught by teacher

N: total number of passing students

Which weights relative scores by the number of total students and passing students. This allows us to penalize doing worse than a global average, but still monitor if more students are passing each year to account for increases in student population for the course.

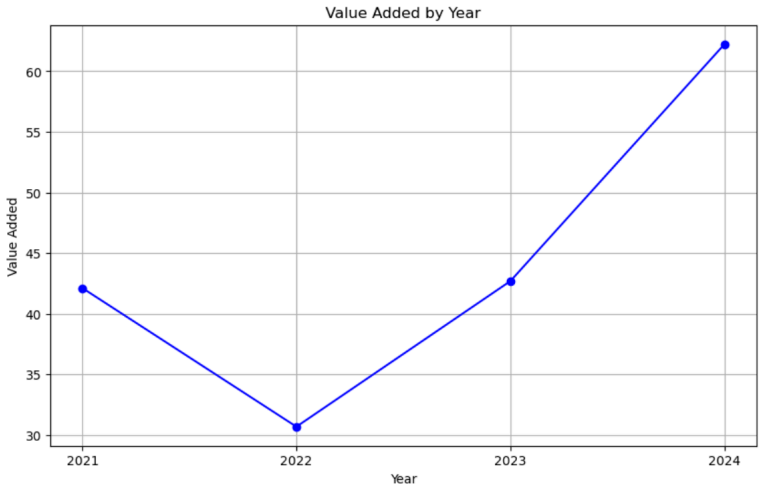

Figure: Value added by educator per year.

Conclusions:

This analysis shows a strong predictive model of student scores on standardized testing for pass fail and individual scores. It also provides a good indicator of teacher focus, allowing an educator to identify areas of improvement. It also provides a framework for a value added by teacher model, allowing an educator to advertise their own performance, as well as show their value as educators to their students. Further evaluation would naturally look at student demographics more carefully (there are none present in this dataset). It should be noted that this analysis should be applied over a longer period of time, as fluctuations year-to-year are notable.

References:

Brownlee, J. (2020, August 16). *SMOTE for Imbalanced Classification with Python*. Machine Learning Mastery. Retrieved from https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

Darling-Hammond, L. (2010). *Evaluating teacher effectiveness: How teacher performance assessments can measure and improve teaching*. Center for American Progress. Retrieved from https://eric.ed.gov/?id=ED535859

Georgia Institute of Technology. (2024). *Advanced Placement Exams*. Retrieved from https://catalog.gatech.edu/academics/undergraduate/credit-tests-scores/advanced-placement-exams/

Goldhaber, D., & Hansen, M. (2010). *Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions. Working Paper 31*. National Center for Analysis of Longitudinal Data in Education Research. Retrieved from https://eric.ed.gov/?id=ED509689