Shauna Smith

Bellevue University – DSC630

Professor: Andrew Hua

Milestone 4 – “Finalizing Your Results”

**Objective**

In Milestone 4, most of the technical work for the project should be done. You should include the information from Milestone 3 and address the following additional items:

Explain your process for prepping the data.

Build and evaluate at least one model.

Interpret your results.

Begin to formulate a conclusion/recommendation.

# Milestone2

**Introduction**

Problem statement

On October 24, 2022, the following article written by Kayla Jimenez for USA Today was published: <https://www.usatoday.com/story/news/2022/10/24/naep-report-card-test-scores-reading-math/10552407002/>

In the article, highlights dealing with the considerable declines in Reading and Math scores. Scores collected by NAEP (National Assessment of Educational Progress) from 2019-2022 were reviewed with the accepted belief that these declines could be attributed to effects from the pandemic. These effects consisted of both remote learning difficulties, and absentee recovery rates during in-person returns. The response to this concerning decline of scores across the United States begs an appropriate but concise response to recover our kids as expediently as possible. In the article, U.S. Education Secretary Miguel Cardona was quoted as saying, “In this moment, we must prioritize intentional collaboration and innovation. We can’t be satisfied with business as usual," Cardona said. "We must do better, and we can.” (Cardona,2022) As inspiring as this proverbial call to arms is, it leads to a very real need to understand how to best address the problem and predictive analysis could be a contributing factor to answer such a call.

Problem pertinence

To quote the infamous Bruce Lee,

"We shall find the answer when we examine the problem, the problem is never apart from the answer, the problem IS the answer, understanding the problem dissolves the problem."**(Lee, nda)**

What this asserts, is acknowledging that by taking the problem in and analyzing its many elements, we only come closer to understanding it. It will therefore be our understanding of the problem, which will yield a predictable solution. With this concept, I propose a model development for predictive analysis using the available data regarding student exams. With this model, deriving predictable outcomes will grant us a workable aid in finding the best solution for discernable differences regarding the acting elements.

Target Audience

Our target audience with resulting advice will be as a collaborative effort in supporting a path forward to Educational Institutions, and regulators. Our efforts will be in offering a concentration on best practices via the contributing variables to assist in exceling scores in reading, writing, and math for k-12 in the United States. Some of these offices, such as the NEAP referenced above and the DOE (Department of Education) with regards to access of the federal ARP funds referenced within the Biden Administration's push to "combat learning loss". However, the main impetus behind our efforts will be with regards to the audiences of students themselves.

Suggested Model Selection

Regarding model selection, efforts for a supervised learning model will be made with a Decision Tree or Random Forest Model. If multi variables converge under class category, then attempts to locate pairs of features with greatest accuracy will be selected and appropriate attentions given to overfitting will be monitored throughout the process.

Plans for Model Evaluation

An analysis through ROC values or F1 Score will be assessed for best methods of confirming feature selection of the model.

Hopeful Lessons to Learn

It is with high hopes that by uncovering the greater predictive variable to achieving the deemed useful feature or features, a garnered focus towards employing the best implementations will be gained. Setting into motion an effective and efficient way to use provisional methods for interventions or supporting advancements given to students across the country.

Risks and Ethical Implications to Consider

When it comes to ethical risks and implications there are several platitudes to consider. Contextual elements inherent to the dataset touch on aspects such as parental education, nutritional advances, and ethnicity considerations. Therefore, it is a noted contention to tread with the understanding that their impacting representation is not intended to perform as a constant, nor be assigned as a metric for diagnostic assays.

Contingency Plan for Project

The largest aspect reserved for an acting contingency plan is modification of classifiers and data augmentation to address a reduction in overfitting if needed. The aspects of future data accumulation and contributions to further analysis or data randomization as validation data versus training data will also be rigorously evaluated as needed.

Points to Consider

Potential points to consider are that all observations derived from this dataset are completely fictional according to the source dataset summary and acknowledgments. Therefore, they are not intended to convey a real-world set of feature observations from which to apply real-world solutions. They are instead intended to portray a fictional case dataset for problem and solution exercise intended for practicing model development with. In this use case, project assignment directives and congregated model training will not represent the proposed problem statement derived from a real-world concept.

# Milestone 3

Predicting Student Exam Scores

Data Source & Problem Statement

For this project's purpose, I will use the provided open-source dataset derived from Kaggle.

* Kaggle Datasets No DOI ava., with contributors:300102-เกียรติศักดิ์, Owais Bin Mushtaq, Anzar. <https://www.kaggle.com/datasets/rkiattisak/student-performance-in-mathematics>

I proposed a model development for predictive analysis using the available data regarding student exams. With this resulting model, deriving predictable outcomes will grant us a workable aid in finding discernable inferences that support improvements. Contextually regarding the given and actionable elements of student success, provides us the best mitigating feature for positively affecting overall student exam scores.

Will I be able to answer the questions I want to answer with the data I have?

“Begin at the beginning”, is wise advice from Lewis Carroll’s great contribution of *Alice’s Adventure in Wonderland*. This simple statement gives us some wonderful insight into how to proceed with our analysis. When We ask how to fix something, we can only answer by knowing what the potential cause was. In other words, where did the problem begin. In our case, origination of influences can tell us how to find the biggest correlation to the beginning decline of student performances.

On review of the provisional dataset, I am assured of plausible explanations. However, I remain cautious to an absent factor not represented in this dataset, nor the conversation. Additional considerations regarding unknown bias or underdeveloped collections are probable with the existing dataset. However, to perform a true representation these potential phantom causes are beyond our scope. Therefore, without preemptive understandings into their existence and thus a prompting for their collection, we must make do with what we have. By focusing on what we have, we can offer credible mitigation methods as a starting force. Essentially, we can at least get the discussion going and entered before the proverbial “Queen’s Court”.

*Our original data points consist of 8 features:*



Although all these features serve as illustrative representations of the students being observed, our biggest concern reflects an overall correlative factor to declining scores. Our concern then, is not incorporative of descriptive features on the students themselves. Descriptive features are irrelevant, because these feature types offer no mitigation means as applicable and actionable qualities to resolving the problem at hand. Therefore, it is prudent that we eliminate such deemed irrelevant features. These features are immutable, or at the very least, beyond the realms of our direct alteration. For this reason, aspects of gender, parental level of education, & race/ethnicity will be dropped from the model training as they are immutable qualities, and irrelevant to the intended goal.

*This will leave us with a dataset consisting of the following:*



From this dataset, I have checked for missing values to which there were none. I intend to use this refined version to train the model. My target variables will be the collective score(s), leaving the training variables of lunch and test preparation courses. These are two easily mitigated and quantifiable features. From these features we can implement a predictable gain and subsequent actionable strategy for our goal of improving exam scores and submitting a measured compensation proposal.

What visualizations are especially useful for explaining my data?

With regards to visualization, I am taking a preemptive approach in establishing a correlation by their causal differences. I will assign visuals between the scores vs. lunch & scores vs sums of test preparation courses. This can establish a positive or negative correlative value and show a comparable difference among the datapoints. Finally, I will provide a visual of the decision tree’s distinct output created by the model.

Do I need to adjust the data and/or driving questions?

When I am addressing the datapoints, I recognize that manipulations of the values will need to transpire for the purposes of properly constructing the model’s interpretations. Calculative dummies or exchanges of delimiting points of interpretive values, may be applied with some liberty. For example, by consolidating a sum value to the three classes of score values to create a singular totem is reasonable. This is done to simplify the process of an overall representation in score value, whilst also accommodating for the model’s efficiency in future applications. Once these manipulations are achieved, I am confident that the available features offer mitigation efforts from which we can reach an actionable conclusion. Overall, yielding us a predictability from which to derive an inference that assists in answering the driving question.

Do I need to adjust my model/evaluation choices?

I have proposed a predictive analysis using a supervised learning model with a decision tree classifier. This is an intentional selection calculated to process for the most predictable feature involved with achieving a greater set score standard. This calculated score standard will be established as a delimitator from which to glean out the greatest feature influence. This influence will then allow for an inference towards the best mitigation effort offered by the provided dataset.

Are my original expectations still reasonable?

My original expectations were to use the provisional features offered in the data and analyze for the most influential and actionable feature on the score outcomes. However, since we are seeking features that offer an actionable solution, only what is plausible and mutable from the dataset will be analyzed. The problem statement concerns itself with finding ways to increase student exam scores by implementing mitigations with a given and provable influential variable. The goal is then to find the highest impact from which to provide a usable mitigation that should create a predictable effect on exam scores over time. Despite this dataset having very limited actionable features from which to derive this perspective, I am still confident that the two selective features will yield insight. Additionally, the features will serve to not only offer workable methods, but also offer a provision for reasonably addressing the problem.

# Milestone 4

Finalizing Your Results

Explain your process for prepping the data.

The process I used to prep my data consisted of a few steps based on both logical assumptions and model requirements for analysis or model performance. As previously discussed, the feature attributes in the original dataset pertaining to qualities that were immutable to either the student or the familial relation, served as no actionable function to the goals set forth in the proposal. Therefore, to implement any type of useable mitigation that would yield improvements on the student exam scores, and provide substantive predictive evidence, only actionable or mutable features as independent variables were kept. Leaving features regarding gender, race, or parental education to be eliminated from the dataset at the forefront due to their deemed immutable and unusable characteristics.



Additional configurations to the dataset were performed for the purpose of visual familiarization. Here is a swarm plot visual that displays proportional attributes as co-optive correlative features to the exam scores from juxtaposed perspectives.

A picture containing text, screenshot

Description automatically generated

Ultimately, the original datapoints of prep courses, lunch standards, and exam scores remained proportionally similar, but distributed differently based on perspectives as displayed above.

All specific exam scores were combined for a single targeted variable due to the insignificance in distinguishing the curriculum matter to the proposed goal.

A black text on a white background

Description automatically generated with low confidence

Creating an overall score better shows minimal movements for classifying improvement measures and offers a means of separating the classes into lower scores vs. higher scores.

Placing the score feature into 1 or 2 rank as higher or lower scoring, was established via the median as the delineator.

A screenshot of a computer program

Description automatically generated with medium confidence

With the dependent variables, a given dummy variable was assigned to define the value category, again as 1 or 2 representing each value provided. Then, dependent on application, data was classified and used as needed for each model or visual display purposes.

-Below is a showing all features classified, with “Scored\_Sums” renamed to “scores”:

A picture containing screenshot, font, text, number

Description automatically generated

Build and evaluate at least one model.

I ended up building 4 models in total, the proposed Random Forest model, followed by a support vector classifier (svc) model as a comparator. Then a Balanced Random Forest (BRF) model, and finally a Decision Tree (dtc) model. Each tree model was fed with similar criterion as ‘gini impurity’ but was adjusted per each given model’s required application and delivery of the data input. For each successful predictive model, I gave the feature importance output, and on the tree classifiers a visual of the tree’s logic path, F1, and/or accuracy score with a confusion matrix visual when applicable.

Finally, for the svc model a created AUC ROC diagram as metrics for output was used. This was performed simply to show the model’s ability to search for both classification and regression analysis with the svc model complementary to the RFC model.

A screenshot of a computer program

Description automatically generated with medium confidence

-Below, “SVC model to RFC” model AUC ROC for predictive lower scores=2.

A picture containing text, screenshot, line, plot

Description automatically generated

However, there were some preferences for accuracy scoring as related to each model’s performance. With the BRF model, I provided an additional cross validation “Mean F1 Score”, “Mean Recall”, and the “Mean Precision”.

-BRF model:

A screen shot of a computer program

Description automatically generated with low confidence

Additional favored regards were granted for the Random Forest model too, by way of an “SVC to RFC” AUC graph as a complimentary comparator (shown above), and an added “Out of Bag (oob) vs. accuracy” score (shown below).

A screenshot of a computer program

Description automatically generated with low confidence

-Below is the data is split for the Random Forest (RF) model with the created working data frame (wdf) explained above:

A screenshot of a computer code

Description automatically generated with low confidence

-Below is the [model=(RF)] from creation to predictions:

A screenshot of a computer program

Description automatically generated with medium confidence

-Below, is the data split for the BRF model creation showing the lowest accuracy score:

A picture containing text, font, line, screenshot

Description automatically generated

-Below, is the data split and creation for (dtc) model & label encoders to the wdf:

A screenshot of a computer program

Description automatically generated with low confidence

Interpret your results.

All these metrics were established to convey the best fit for the data. Ultimately, gaining assurances that the best performing model for the given dataset, was indeed the Random Forest (RF) Model, as was predicted in the proposal. Whereby, the metrics for an accuracy score was 99%, and the ‘oob score’ was 98 when showing for an additional unbiased true prediction estimate. Finally, the F1 was approx. 99.5 which is extremely close to 1.0, and the confusion matrix demonstrated 198 of 200 true positive predictions.

-Below, shows the F1(close to 1.0) & Confusion Matrix for the RF model:

A picture containing text, font, screenshot, white

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Additional confirmation was given regarding repeated feature importance outputs from successful models. Showings for ‘prep’ at (54% vs. lunch at 45% in magnitude) in the RF model, and again for ‘prep’ at (57% vs. lunch 42% in magnitude) with the similar dtc model can be used to serve as evidence for assigning ‘prep courses’ from the dataset’s actionable features, as the best selection for focused mitigation efforts to improve exam scores.

-Below, RF model feature importance:

A picture containing text, font, line, screenshot

Description automatically generated

-Below, dtc model feature importance & F1(closest to 1.0) with following tree diagram:

A screenshot of a computer code

Description automatically generated with low confidence

A picture containing text, font, screenshot, white

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Begin to formulate a conclusion/recommendation.

It was concluded that the effectiveness of the Random Forest (RF) model was the best fit to the dataset. Further assurances were granted through the performance outputs of comparative models, or the merits of their own predictive scores with the repeatable feature importance conclusions. Conclusively, the chosen RF model was found without need for further data augmentation through testing the model’s performance. Final regards to the conclusions in conjunction with the dataset, yield a strong support in favor of the outcomes and predictive nature of the model for supplemental future uses. The RF model as it stands with the current features inherent to the dataset, conveyed a well recommended effect for change on scores via the ‘prep-courses’ as the key driver for mitigation efforts.

However, as an additional feature importance it is recommended to include the mitigation measure in favor of lunch provisions, as lunch was garnering around approx. 45% of the importance rate likely due to limited features available. Therefore, the recommendation for lunch subsidy programs is as a secondary mitigation effort from which to focus additional actionable attentions during the applied implementation process. Leaving lunch programs as a feasible supporting mitigation to be included with the prep-courses in the given proposal which is required for submissions into the federal ARP funds program.

If the federal ARP funds policy dictates allocation of funds as limited or restricted to a single given class or focus, then priority should be garnered to the subsidy of prep-courses for all given subject matter pertinent to each child. The prep-courses offered as the key mitigation effort stand to predictively garner the greatest impact on student exam score totals. Therefore, this feature should be appropriately harnessed to expedite the most efficient and impactful means for student recovery when submitted for reimbursement by the ARP program proposal as a singular procedural mitigation effort.

## References

\*Kaggle Datasets No DOI ava., with contributors:300102-เกียรติศักดิ์, Owais Bin Mushtaq, Anzar. <https://www.kaggle.com/datasets/rkiattisak/student-performance-in-mathematics>

\*Carroll, L. (1993). *Alice’s Adventures in Wonderland*. Dover Publications.

Data Pertinence

Additional reference reviews to content knowledge and data understanding were derived from the following sources:

\*Article -Poli, R. (2001). Reading and math test scores fell across US during the pandemic. How did your state fare?, USA Today. <https://www.usatoday.com/story/news/2022/10/24/naep-report-card-test-scores-reading-math/10552407002/>

\*Pdf -data.world's Admin. (2023, January). School Student Health and Wellbeing. Physical Activity, Nutrition, Lifestyle, and Emotional Health Behaviors. Kaggle datasets, <https://www.kaggle.com/datasets/thedevastator/school-student-health-and-wellbeing>

\*Smith, Shauna, (5/2023) , “SmithSDSC630 Milestone4”, working(.ipynb) file for exams dataset & Model ensembles with outputs, and visuals as link, <https://1drv.ms/u/s!AsRy_XxFji_C_Tn9Omeyx3xSXlx-?e=tEMZoN>