Shauna Smith

Bellevue University – DSC630

Professor: Andrew Hua

Milestone 5 – “Final Project Paper Presentation”

# Objective

**Introduction**

* Problem statement
* Explain why the problem is important/interesting
* Who would be interested in solving this problem, i.e., who would you be trying to sell this project to?
* Where did you get your data?
* Why is this data useful to solve the problem?

**Methods/Results**

* What did you find out by exploring the data?
* Are there any visualizations that help tell a story with your data?
* What steps did you perform to prepare the data?
* What type of types of modeling are you using on your data?
* What metric(s) are you using to measure your results?
* Why did you choose the metric(s) you chose?

**Conclusion**

* What did you learn?
* What recommendations would you make based off your analysis?
* Is your model ready for deployment?
* What work still needs to be done?
* What do you need to consider ethically regarding the data, your model and the presentation of results?
* What ethical implications exist, or could exist if this project were live in production?
* What could be done to mitigate the ethical concerns, if anything?

**References**

* Include at least three properly cited references at the end of your paper
* Also include in-text citations

# 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 for dealing with the considerable declines in reading and math scores caused by learning loss from the pandemic were discussed. 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. The goal of educators and trusted institutions of authority such as the Department of Education, is 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 went on to exclaim, "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. Today, I propose 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 such a model, deriving predictable outcomes will grant us a workable aid in finding the best solution for discernable differences regarding both the acting & actionable elements worthy of consideration.

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 are the NEAP and the DOE (Department of Education) referenced above. These considerations are explicitly 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 address the needs to the audience of guardianship for the students themselves.

Data Source & Usefulness for our proposal

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, the given and actionable elements of student success will provide us the best mitigating feature for positively affecting overall student exam scores. Therefore, 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 implementation strategy will be gained. Ultimately, setting into motion an effective and efficient way to use provisional methods for interventions or supporting advancements given as mitigation efforts to students across the country.

## Methods & Results

Data exploring

The process I used to prep my data consisted of a few steps. These steps were based on both logical assumptions and model requirements for both analysis and model performances. As previously touched on, the feature attributes in the original dataset pertaining to qualities that were immutable to either the student or the familial relations, 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.



Visualizations that tell a story & steps in preparing the data

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

A picture containing text, screenshot

Description automatically generated

Ultimately, the original datapoints of prep courses, lunch standards, and exam scores remained somewhat 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” simply renamed to just “scores”:

A picture containing screenshot, font, text, number

Description automatically generated

Models selected for discovery

During the investigative process, 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 a similar criterion as ‘gini impurity’. This was served to account for the likelihood of class via the dataset’s distribution. However, each model was adjusted per the model’s required application and delivery for 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 a metrics for output. 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. (please see below)

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 with “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 proposed Random Forest model too, by way of aforementioned “SVC to RFC” AUC graph, functioning as a complimentary comparator (previously shown above), and with 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 as a ‘Working Data Frame” (wdf) split for the Random Forest (RF) model:

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 randomly split for the BRF model creation. Ultimately, showing the lowest accuracy score out of all the created model’s:

A picture containing text, font, line, screenshot

Description automatically generated

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

A screenshot of a computer program

Description automatically generated with low confidence

Metrics used to measure results & why they were used

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 which helped show for an additional unbiased true prediction estimate. Finally, the F1 score was approx. 99.5 which is extremely close to 1.0 correlation, and the confusion matrix demonstrated 198 of the 200 as true positive predictions, all of which are listed below.

-The F1\_score(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 top 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. These both 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 overall.

-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

# Conclusions

What we learned & Recommendations

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 demonstrations for repeatable feature importance. Conclusively, the chosen RF model was found without need for further data augmentation through testing the model 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 feature availability. 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. Since evidence based implementation is required for submissions into the federal ARP funds program, the RF model’s demonstrations easily support both features in a hierarchical manner.

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. This shows provable yields as a singular procedural mitigation effort adequate for program proposal submission.

Model readiness

When addressing model readiness, again we are essentially inquiring if the model would perform at or above baseline accuracy. We had a limited feature selection from which to train our model which places some impacts to the level of noticeable error thresholds or fixed nodes along the model’s decision process. However, there is no demonstratable evidence for a need of data augmentation, and our target accuracy has been sufficiently demonstrated for our proposal purposes as a usage base. Should follow up data be fed in the future, some considerations to the model’s efficiency and criteria are easily established and stand to support the model’s use to achieve these determined benchmarks. If additional features are worthy of calculations, then further maintenance to the model and classifications can be considered.

Potential risks & ethical considerations with final points

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.

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 as concrete evidence. They are instead, elicited to portray a fictional case dataset for problem and solution exercises intended for training model development. In this use case, project assignment directives and congregated model training will not represent the proposed problem statement derived from a real-world concept. Therefore, future feeds of collected data are suggested in the proposal’s diagnostic supplementary evaluations, and model maintenance for future confirmation. Baseline or usage accuracy are once again recommended as revising or review for a strong method in periodic practice.

## 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>

\*Smith, Shauna, (5/2023), “Smith Power Point proposal slide show link”, working( .ppsx) file for audio and proposal presentation slides, [SmithSDSC630Milestone5Presentation.ppsx](https://1drv.ms/p/s!AsRy_XxFji_CgP4xpJ4of5V7FFoSOQ?e=zqtvSG)