|  |  |
| --- | --- |
|  | PREDICTING students' average performance  April Meyer  DSC680-T301 Applied Data Science (2215-1) |

# **Executive Summary**

Standardized testing has been included in the educational system throughout history and has caused controversy. This type of testing measures a student’s progress at any given time in their education. There are socioeconomic aspects that could impact a student test scores. The purpose of this project is to show how using data could prove or disprove the impact of these socioeconomic influences and thus proving standardized testing is not biased or unbiased. The impact could be negative, positive, or nothing.

The methods used was Exploratory Data Analysis in Python. Linear regression and random forest regressor models were used from Supervised machine learning models in Sklearn.

# **Preliminary Analysis**

## **Background of the problem**

Standardized testing has been included in the educational system throughout history and has caused controversy. This type of testing measures a student’s progress at any given time in their education. It became even more of a controversy after the No Child Left Behind Act was passed in 2002. One side of the debate is that this type of testing causes teaching for the test and does not allow for creativity or critical thinkers; while the other side states it is a reasonable and unbiased way to quantify student’s performance. However, there is more to this debate. There are socioeconomic aspects that could impact a student test scores. The purpose of this project is to show how using data could prove the impact of these socioeconomic influences and thus proving standardized testing is not unbiased.

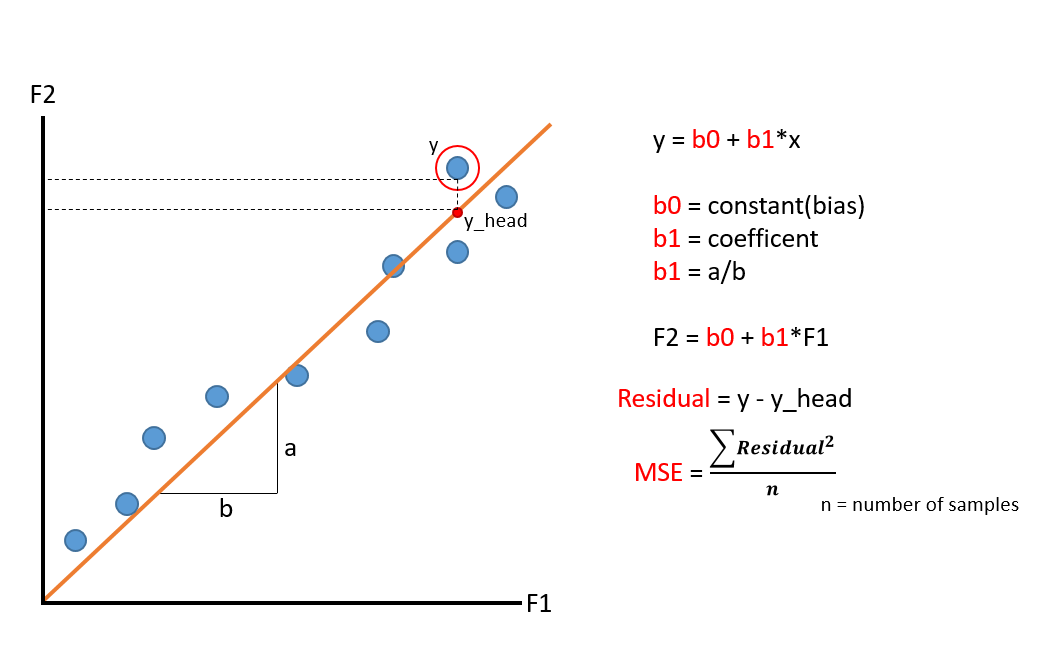
## **Problem Statement**

There are numerous influences that can affect how a student will perform on exams, including physical, social, and environmental. The impact could be negative, positive, or nothing. This analysis seeks to show how data could be used to understand these influences.

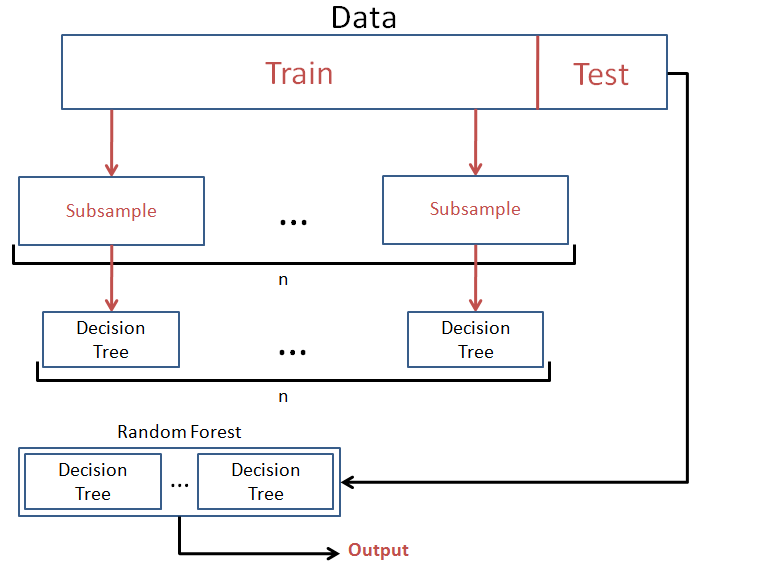
## **Methods**

Supervised machine learning models were used for this project, linear regression, and random forest regressor.

Linear regression is a linear model that looks at the relationship between a dependent variable and explanatory or independent variables (3).



Random Forest Regressor is an ensemble learning method that runs by assembling a variety of decision trees at training and outputs the mean of predication of individual trees (1).



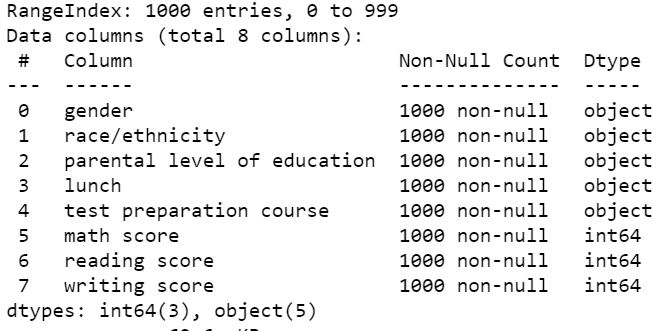
## **Data Sources**

The dataset contains math, reading and writing scores of students along with some student information such as gender, race, parents’ education, lunch type, and if the student took test preparation courses (2). It is a fictional dataset used to show how machine learning could be applied in education. Link to data, http://roycekimmons.com/tools/generated\_data/exams.

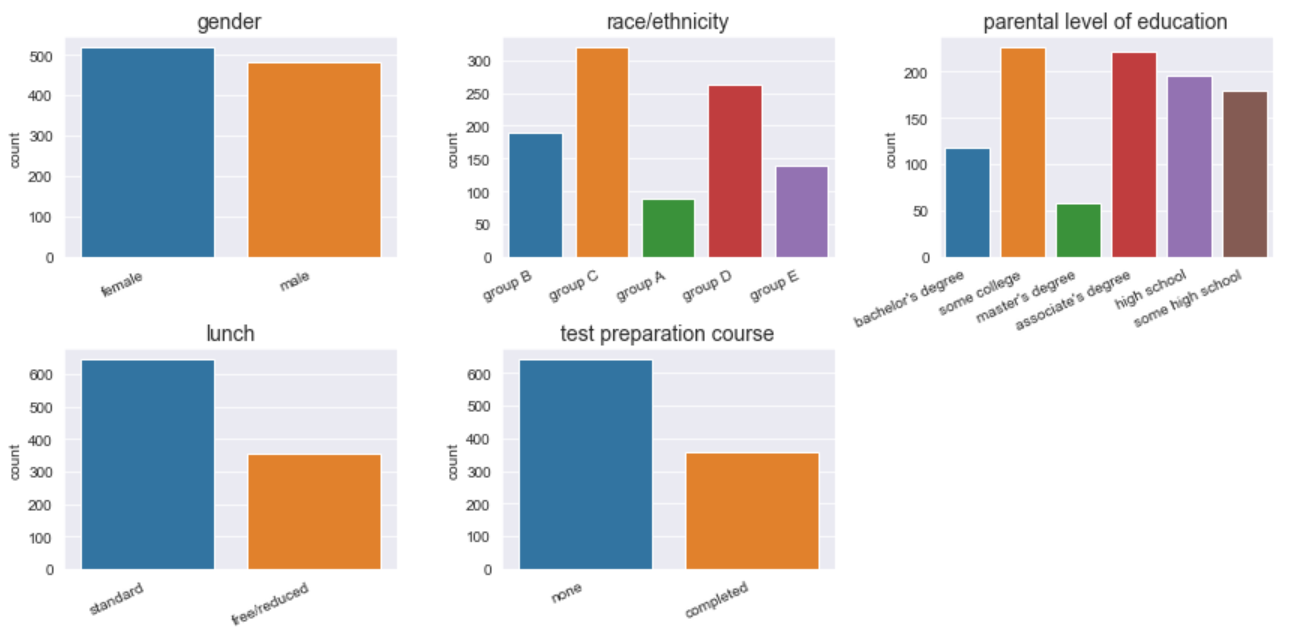
## **Data Exploration**

The initial analysis of the attributes for the dataset are below:

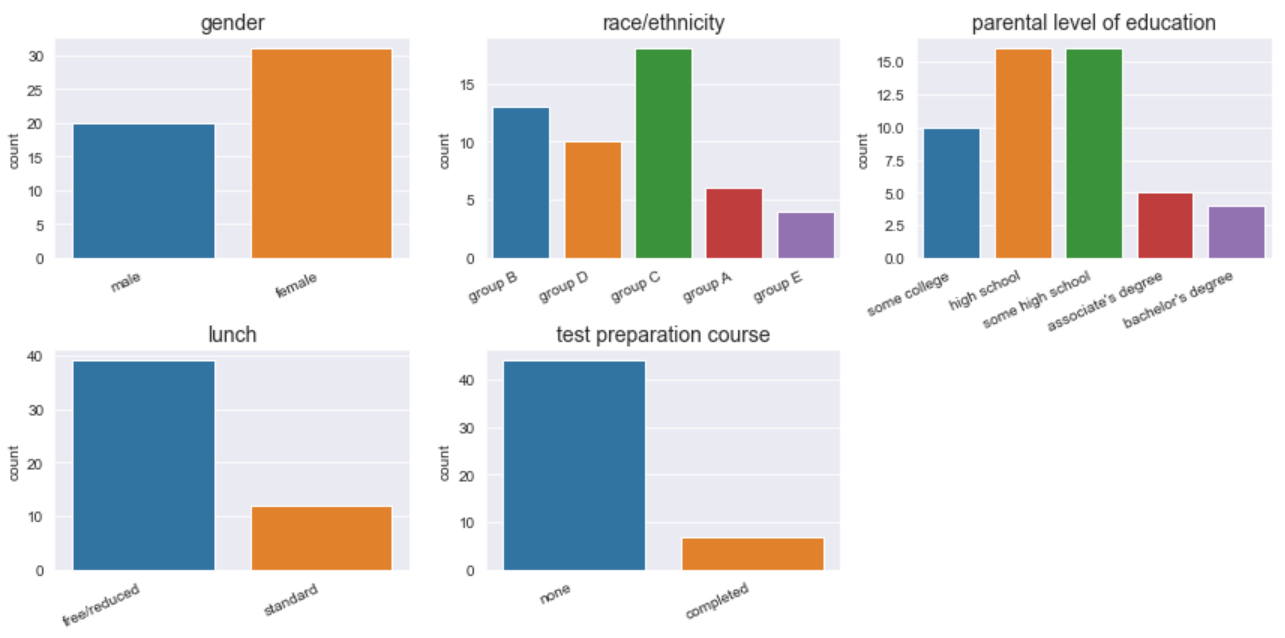
* gender: male or female
* race/ethnicity: represents five (A-E) groups
* parental level of education: six levels of education
* lunch: standard or free/reduced.
* test preparation course: none or completed.
* math score: score for math
* reading score: score for reading
* writing score: score for writing



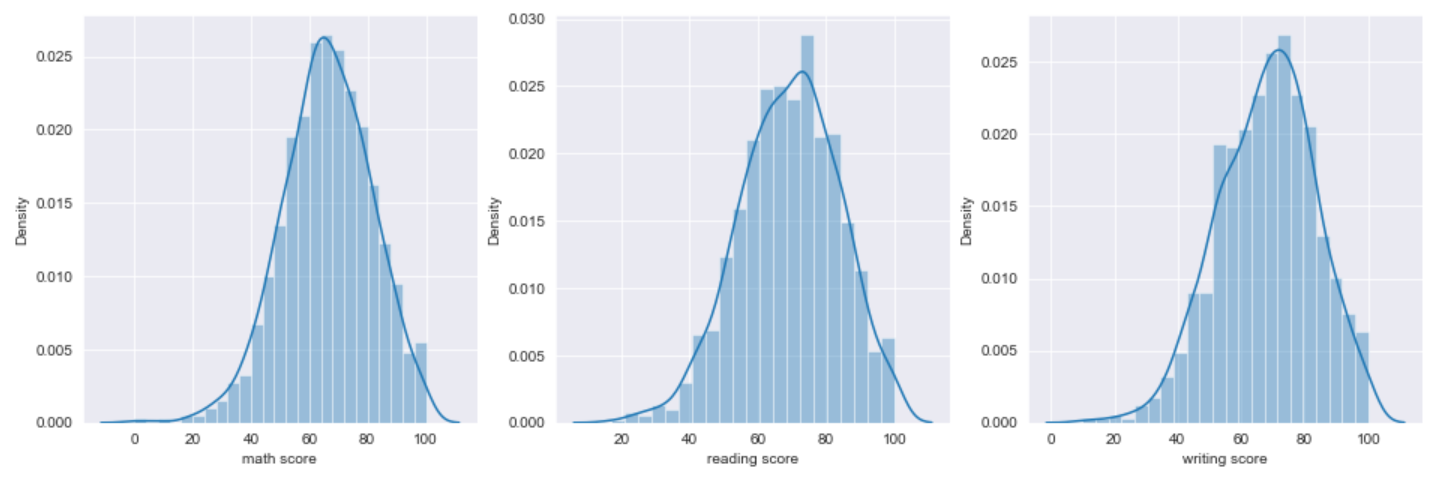
**Categorical Count of All Variables:**



**Categorical Count of Variable where Scores are below 40:**

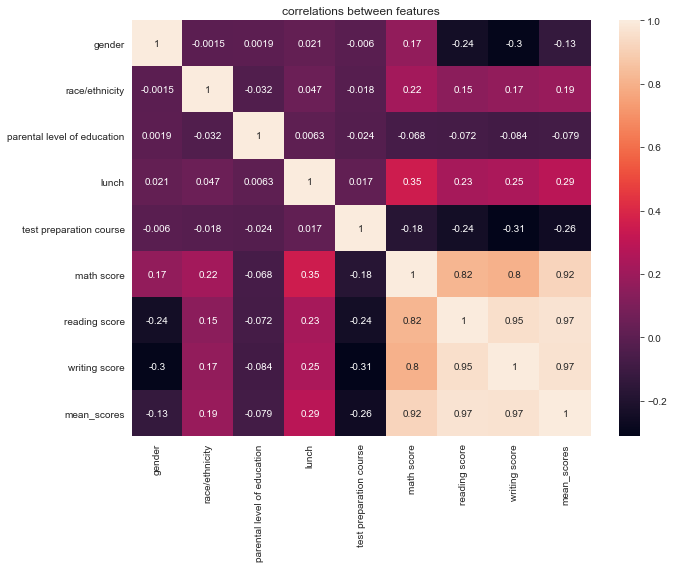


**No skewness in the Target Columns:**



## **Data Preparation**

For data preparation categorical features were transformed to use in model and correlations were evaluated.



After transforming categorical variables, a new feature was created called mean\_score. It is the average of math, reading, and writing scores. It will be used as the target variable for the machine learning models.

## **Modeling**

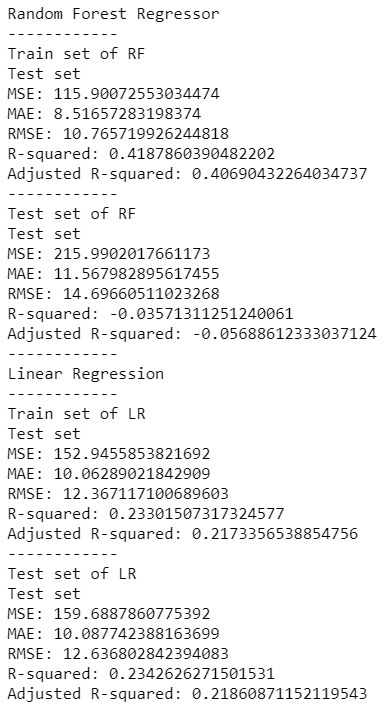
After the data preparation the data was split using train\_test\_split. It is a function from Sklearn for splitting data arrays into two subsets: for training data and for testing data (3). test\_size was set to 0.4 which specifies the size of the testing dataset.

## **Results**

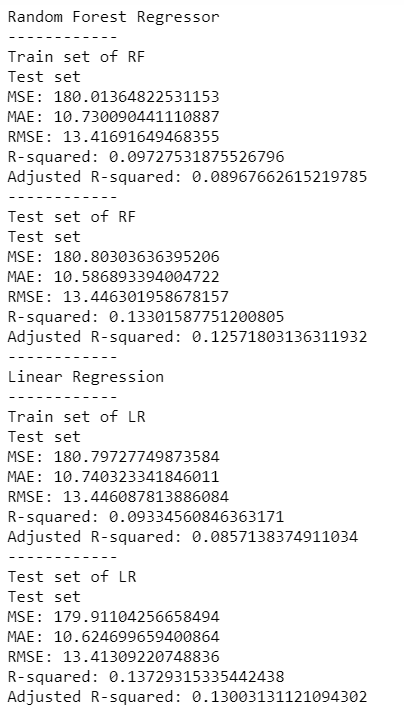
The model was validated using the three different metrics for regression:

1. R Square/Adjusted R Square: R Square is the square of correlation coefficient (R) and represents how much of variability in the dependent variable can be explained by the model (7). If there are too many independent variables there could be an overfitting problem. Adjusted R Square takes into consideration the additional independent variables (7). The R2 for Random Forest Regressor was negative meaning I was doing worse than the mean value (horizontal line). The Linear Regression performed a bit better, but not much. About 22% of dependent variability can be explain by the model when using all independent variables.
2. Mean Square Error (MSE)/Root Mean Square Error (RMSE): Mean Square Error is an absolute measure of the goodness for the fit (7). Root Mean Square Error is the square root of MSE (7).
3. Mean Absolute Error (MAE): MAE is a more closely related to the sum of error. MSE gives larger penalization to big prediction error by squaring it (7). A lower value of MAE, MSE, or RMSE suggests higher accuracy of a regression model (7).

A function was built called evaluateRegressionModel(real, predict). It returns the above metrics. Below are the results using all the available independent variables.



Below are the results using only ‘lunch’ and ‘race/ethnicity’ as independent variables:



## **Discussion/Conclusion**

Overall, the two models did not perform well. It does suggest that finding variables that would impact student’s scores could be difficult. However, using a different model or real-life data could give different results. The data correlation does suggest that lunch could impact scores. Real-life data could give more variables to work with and would be worth exploring.

## **Acknowledgements**

I would like to acknowledge the Kaggle Repository for providing this dataset I used for the project.

# **References:**

# **K, D. (2021, April 08). Random forest regression in python using scikit-learn. Retrieved April 17, 2021, from** [**https://heartbeat.fritz.ai/random-forest-regression-in-python-using-scikit-learn-9e9b147e2153**](https://heartbeat.fritz.ai/random-forest-regression-in-python-using-scikit-learn-9e9b147e2153)

Blog explaining random forest regression. It will be used as an example to perform this model from scikit-learn on the dataset.

# **Kimmons, R. (n.d.). Exam scores. Retrieved April 17, 2021, from** [**http://roycekimmons.com/tools/generated\_data/exams**](http://roycekimmons.com/tools/generated_data/exams)

Website with different datasets that are generated for learning purposes. This is where the original dataset is from.

# **Robinson, S. (n.d.). Linear regression in Python with scikit-learn. Retrieved April 17, 2021, from** [**https://stackabuse.com/linear-regression-in-python-with-scikit-learn/**](https://stackabuse.com/linear-regression-in-python-with-scikit-learn/)

Blog explaining linear regression. It will be used as an example to perform this model from scikit-learn on the dataset.

# **Seshapanpu, J. (2018, November 09). Students performance in exams. Retrieved April 17, 2021, from** [**https://www.kaggle.com/spscientist/students-performance-in-exams**](https://www.kaggle.com/spscientist/students-performance-in-exams)

Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. This reference is for the data set used.

# **Sklearn.ensemble.randomforestregressor¶. (n.d.). Retrieved April 17, 2021, from** [**https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html**](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html)

Scikit-learn implementation guide website. It explains the model parameters and gives an example.

# **Sklearn.linear\_model.LinearRegression¶. (n.d.). Retrieved April 17, 2021, from** [**https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LinearRegression.html**](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html)

Scikit-learn implementation guide website. It explains the model parameters and gives an example.

# **3 Best metrics to evaluate Regression Model (n.d.). Retrieved April 17, 2021, from https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b**

Scikit-learn implementation guide website. It explains the model parameters and gives an example.