Linear Regression and Predicting Reading Score Based on One or More of the Other Features

IAF 604

Matthew Afsahi

Dr. Prashanti Manda

Problem Introduction

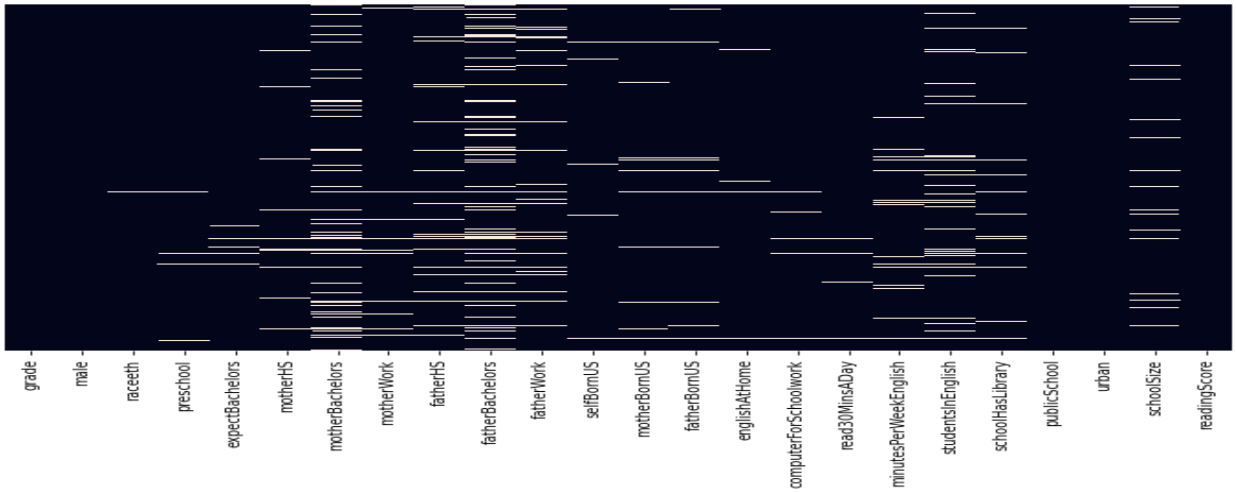
Education is one of the most important feature that let the globe grow. Hence, parents and stakeholders are researching to find what is important and how it is possible to gain the education levels for their children and themselves as well. In this analysis, I will analyze the data provided by demographics and schools for American students taking the exam, derived from 2009 PISA Public-Use Data Files distributed by the United States National Center for Education Statistics (NCES) to find out what criterions are the most importance features in education levels as well as developing machine learning models to estimate the scores of a new student based on the history of his/her index.

In particular, I will focus on Linear Regression, Random Forest Regression, Decision Tree and Artificial Neural Networks Regression's deep learning to practice and find the best fitted model for this project. Moreover, comparing the root mean squared errors, cross validation scores, and explained variance scores with the help of the residual plots conclude the best model in this project. Moreover, feature engineering selection methods such as Spearman and Pearson correlations will be implementing to get more accurate results. In addition, with analyzing the regression's coefficients of each model as well as features selection based on Random Forest of ML model give an estimate to find the most importance feature of the variables in dataset.

Dataset

The dataset had distributed by the United States National center for Education Statistics named 2009 PISA Public-Use Data Files, which provided with the two separate training and testing datasets. The data had 24 variables who I am trying to estimate the reading score as a response variable in this process. The other variables consider to be predictors' variables in this project. Data has an object variable called raceeth where it is a label variable. There are some missing data points as known as null values which should be fixing in order to gain the higher accuracy in each model.

Missing values in datasets



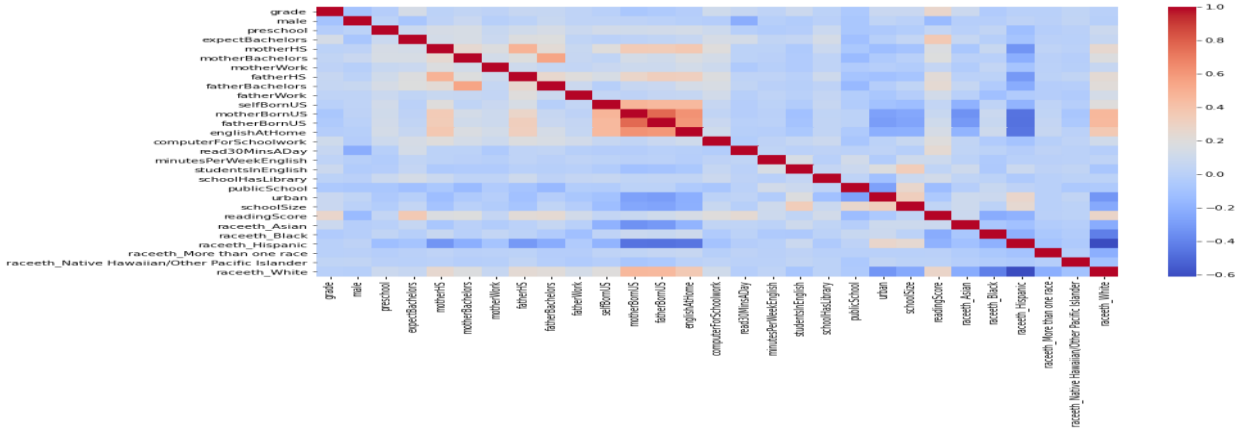
Feature and Processing

I initially started with the raw data and implemented many feature extraction methods with the help of visualizations. Also, I implemented such feature engineering selections methods to develop each model finding the extracted variables based on the top score values.

First, I used method such as most frequent variable to filling the missed categorical values and transforming them to the numeric labels with the use of Pandas' dummies function.

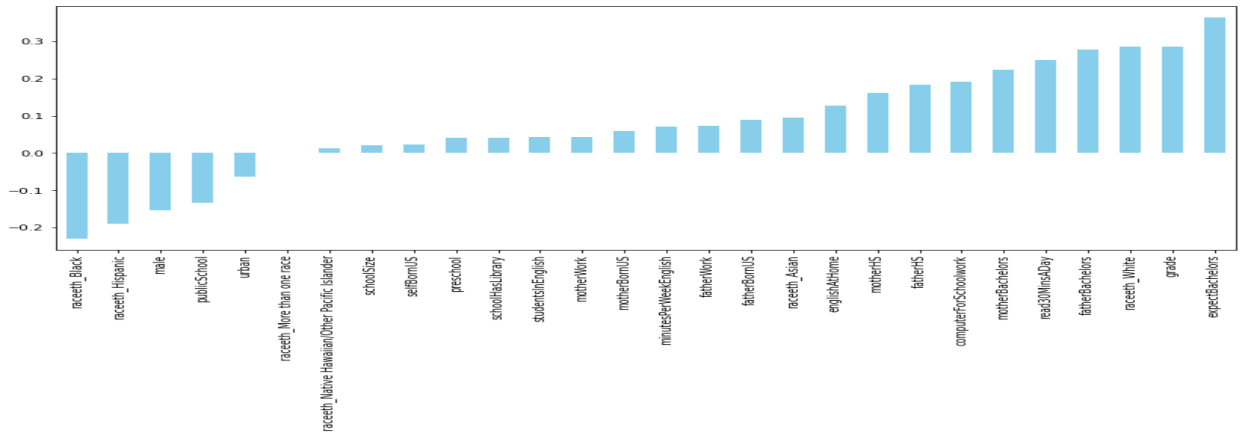
Also, implemented an Iterative Imputer module to impute the missing values based on Bayesian Ridge method which is the most accurate of the imputation of the missing values in the dataset. This function works perfectly based on the Bayesian function that can estimate and map the missing values based on the other predictors in a round robin application. Moreover, by implementing a Spearman correlation method to see what the potential variable may be the most correlated features to the reading scores response variable.

Spearman Correlations



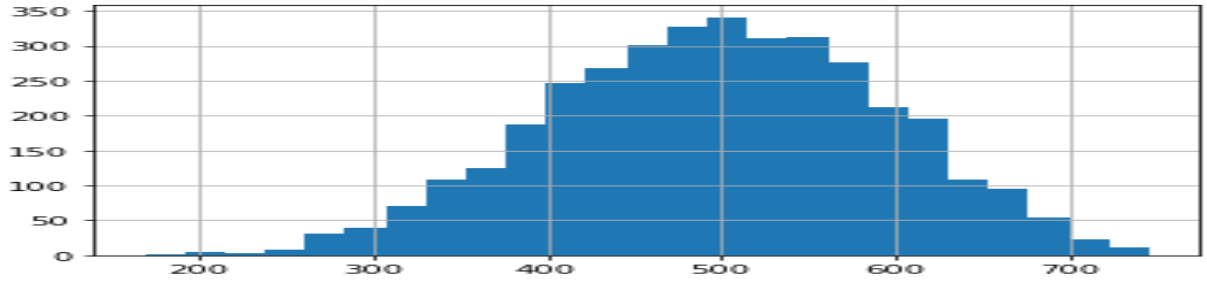
Moreover, by sorting the correlations respect to reading score in a bar chart could find more clear relations between the predictors and response variable as shown below chart.

Correlation bar between reading score and predictors



In addition, scaling the data as the data is not in a same unit is important. Therefore, normalizing the data is required which ANN models need to be normalizes as well. Finally, Inspecting the target variable with the normality assumption is required to fit the machine learning estimators. By visualizing the target variable, I ensured that the target variable does not need any transformation.

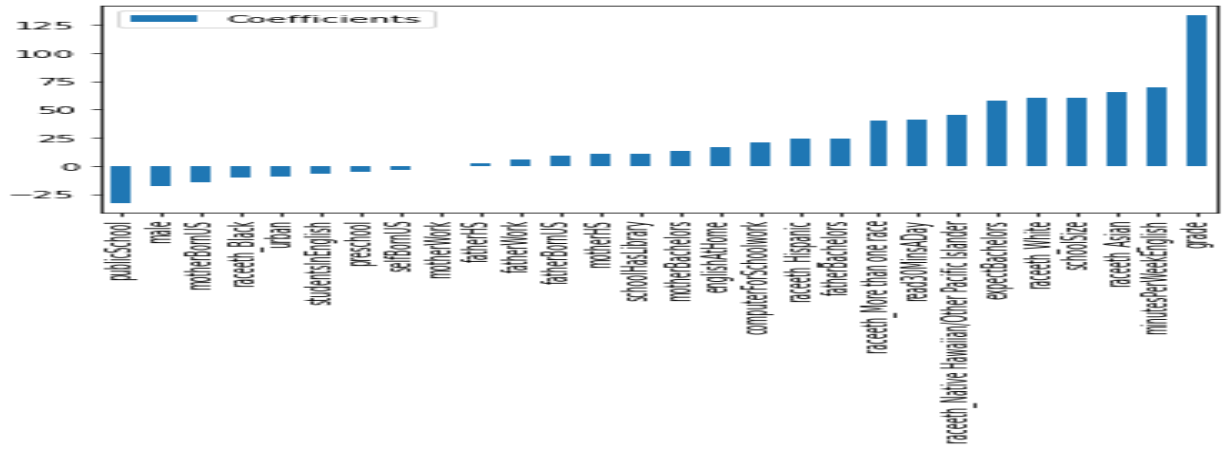
Target variable distribution



Models and Technics

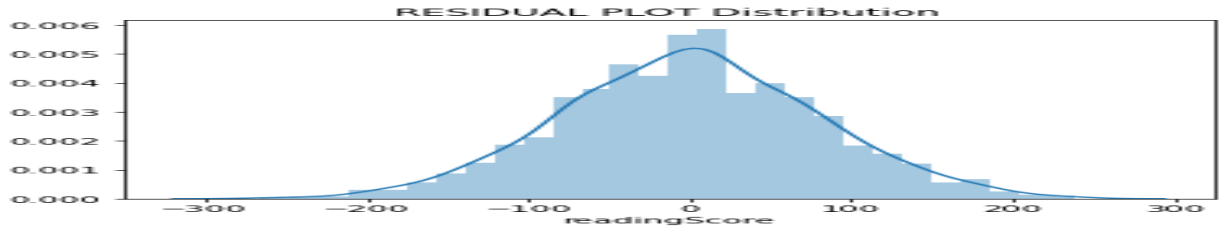
Linear Regression: As long as the target variable is a continues measure, it is the best practice to use the linear regression machine learning module. After fitting a linear regression estimator on the scaled features and target variable on the training set, I have extracted the model scores as well as coefficients where sorting them to graph a bar chart, gave me a clear visualization to find the most important predictors features respected to the Linear model.

Coefficients feature selection



Analysis of the predictions based on the test scaled dataset gives a clear estimation of finding the root mean score errors and plotting the residuals as well. The plotted residuals shows a normally distribution where it emphasizes that the variance of the model is consistence and the process probably is valid in this project.

Residual Plot of Linear Regression



Findings of the linear regression model

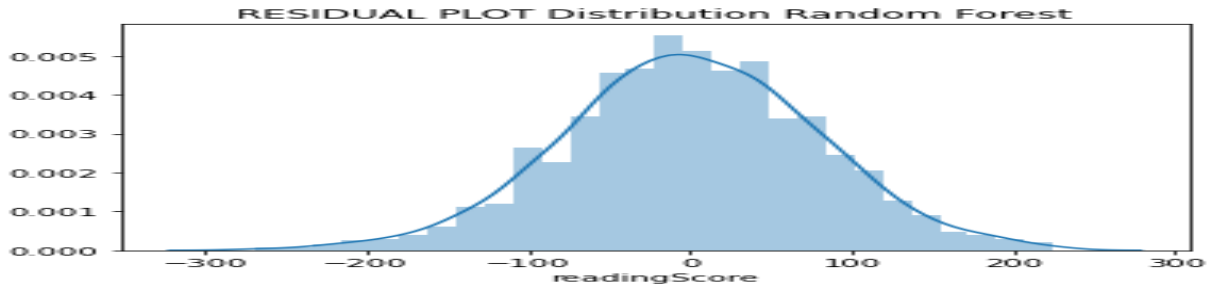
Explained Variance Score on Test Data: 31 %

Root Mean Score Error: 79

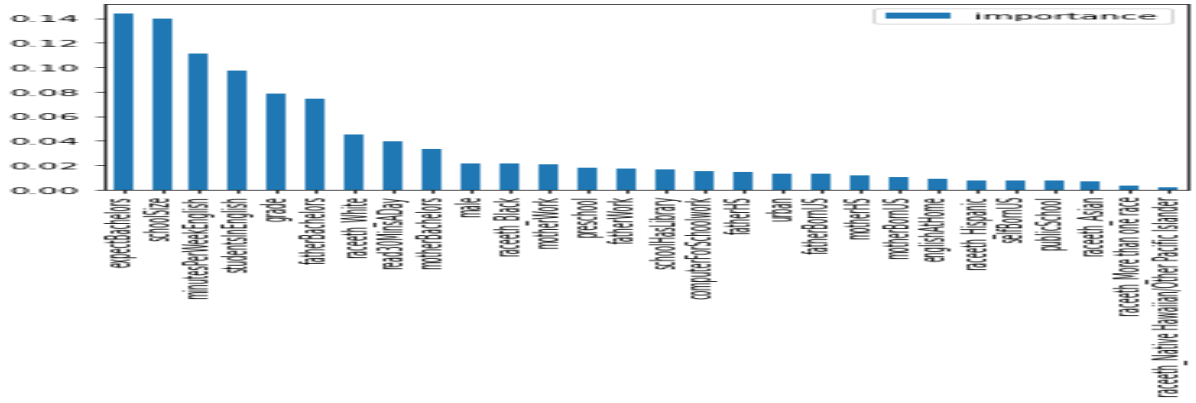
Cross Validation Scores: 0.29992933, 0.19970677, 0.20915095, 0.36079812, and 0.32814148

Random Forest: Random Forest implementation has own unique features, so by choosing the number of estimators of 100, implemented a random Forest model of machine learning to develop and extract the importance features as well as other scores to compare the models of the implantation in this project. Random Forest gave a better estimations and residual plots with the different important features shown below

Residual Plot of Random Forest



Importance Features Based on Random Forest Estimator



Findings of the Random Forest Model:

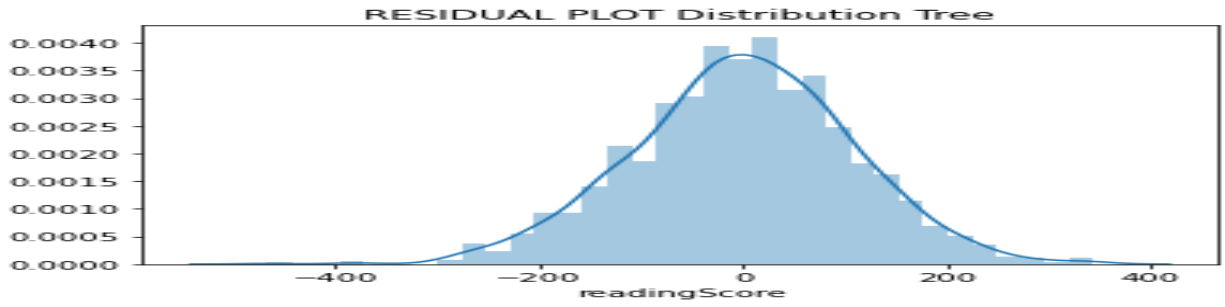
Explained Variance Score: 33 %

Root Mean Score Error: 78

Cross Validation Scores: 0.33578434, 0.22657283, 0.28154744, 0.32578176, and 0.28824247

Decision Trees: Implementing the Decision Trees machine learning model did not give a better prediction's score based on root mean squared errors and explained variance score. This needs m ore feature engineering process such as pruning the tree.

Residual Plot Decision tree



Findings of the Decision Tree Model

Explained Variance Score on Train Dataset: 99%

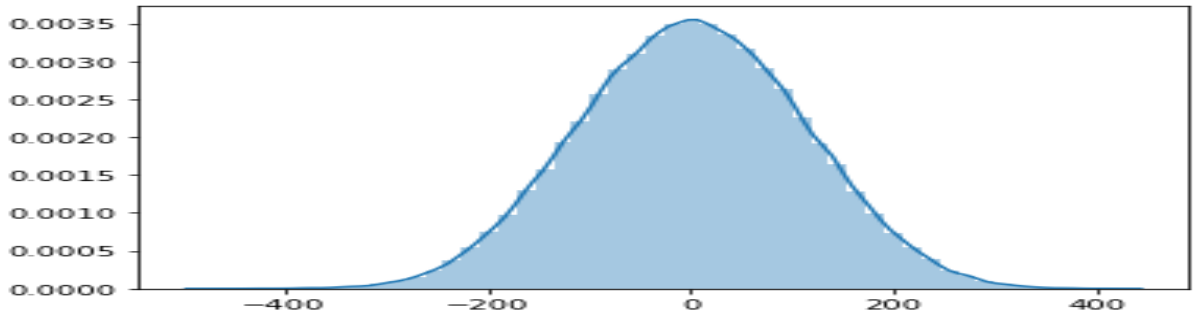
Explained Variance Score on Test Dataset: -26 %

Root Mean Score Error: 107

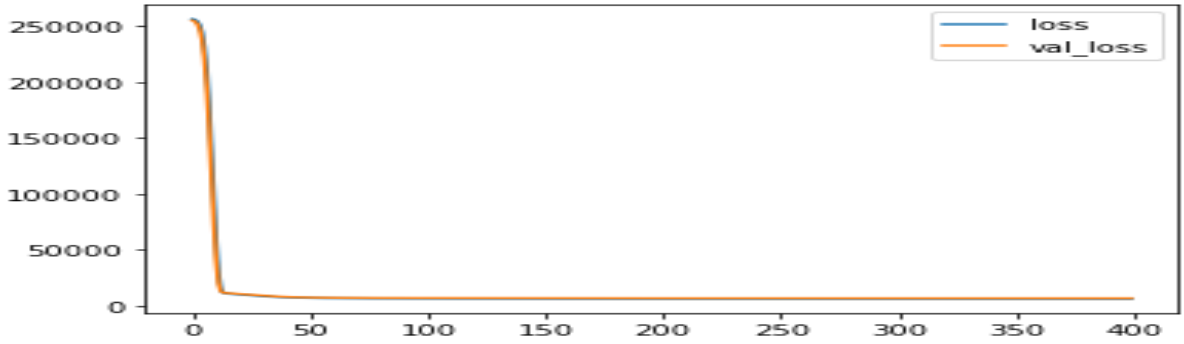
Cross Validation Scores: -0.31940013, -0.46327876, -0.45351955, -0.5872706, -0.40067615

Artificial Neural Networks Deep Learning with Tensor Flow: I implemented a ANN regression model with a input layer of 28 inputs and also finalized a designing of the three hidden layer of the 28 inputs each with of the ReLu activation function and the Adam optimizer. Also, a loss function of mean squared error with validating on test scaled feature test data and target test dataset. The batch size of 128 and epochs of 400 applied to the fitting model.

Residual Plots ANN



Loss Chart validation on test predictors dataset and target test dataset



Findings of the ANN Model

Explained Variance Score: 32 %

Root Mean Score Error: 78

Results

Finally, I have reduced the variables on this dataset from 24 predictors to 9 predictors columns based on the Random Forest Importance Features to design a better model estimating the reading scores with the higher explained variance scores and lower the root mean squared errors. The reduced predictors found that to be 'expectBachelors', 'schoolSize', 'minutesPerWeekEnglish', 'studentsInEnglish', 'grade', 'fatherBachelors', 'raceeth_White', 'read30MinsADay', 'motherBachelors'.

This report concludes a suggestion to stakeholders for education departments as well as the parents who are wondering the best choices for their children, where they can by focusing and spending more time and intention to these features somehow that the increase of one unit of these predictors could lead to increase the some unit of the reading scores depending to the coefficient’s of each predictors shown in the models and technics section above. Also, with comparing all the implemented models it has found that the best model of ML was Random Forest as well as ANN model with Tensor flow.