Kiejstut Bunikiewicz

Professor Brahma

BSAN 6070

05/05/21

Final Project Individual Portion Report

**Individual Model Attempts: Gradient Boosting Regressors and Gradient Boosting Classifier**

**Model Choices and Justification**

The three models I ran were the Gradient Boosted Regressor, the Gradient Boosted Regressor with Standard Scaler, and the Gradient Boosted Classifier. Based on our project’s initial goal of predicting the amount of CO2 emissions per person based on the various factors of country information, I had decided to implement a Gradient Boosted Regressor to attempt to predict the amount of CO2 emissions. The advantages of this technique rather than utilizing a standard Decision Tree, such as CART is that a Gradient Boosted Regression or Classifier is more likely to provide strong prediction performance through its multiple iterations of the Decision Tree. The Gradient Boosted methods aim to reduce loss, or prediction error through the improvement of the prediction through the use of multiple prediction iterations to improve its predictions with each iteration. The method of using aggregate decision making because the prediction is arrived at through a weighted average means that the Gradient Boosted models are more likely to make more accurate predictions as the model learns from its errors in prior iterations. The Gradient Boosted models are also good for prediction because they are less prone to overfitting because the Gradient Boosted techniques arrive at an average for their decision making. This means that each iteration fits the data differently, reducing the chances of overfitting. Another useful feature of Gradient Boosted Techniques is that they can provide feature importances for both regressions and classifiers, even if the model cannot explain why it has arrived at those feature importances. Another reason why the Gradient Boosted techniques are useful for analyzing the country data is that they are robust with regards to outliers because tree-based models can account for outliers in the various splits they make. These models do not require normalization or scaling of data to be effective. I also did not need to do any specific feature selection for my model other than the features agreed upon by the group.

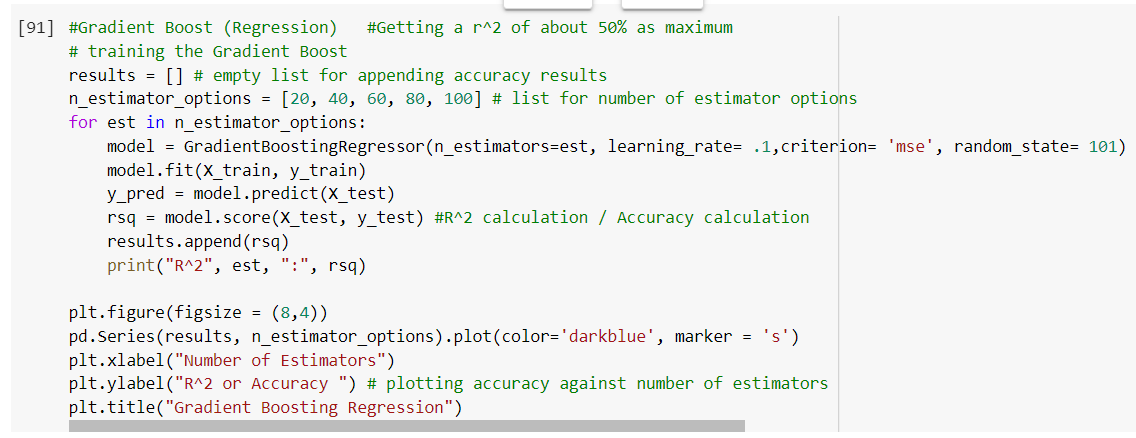
Unfortunately, the Gradient Boosted techniques also come with the limitations of being a black box model and of taking a long time to train. Our dataset was small enough that the training time was not substantial. Due to the process of arriving at a decision by aggregate, the Gradient Boosted techniques are black box models. This means that although we will be able to make predictions or classifications as well as finding feature importances, we will not be able to identify why the models arrived at such a decision. The benefits of a high ability to predict the value or the class and the reduced risk of overfitting make these models a strong model for answering our question.

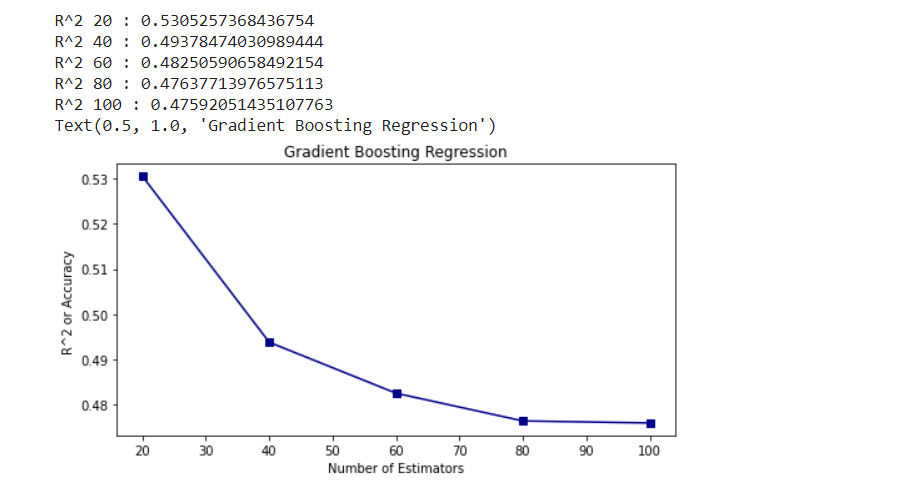
**Model Assumption Validation**

Our initial dataset was composed of numerous continuous variables relating to the country’s Population, Area, Population Density, Coastline, Net Migration, Infant Mortality, GDP, Phones, Arable land, Crop land, Other land usage, Birth Rate, Death Rate, Percent of Economy made up by Agriculture, Percent of Economy made up by Industry, and Percent of Economy made up by Service. Our only categorical independent variable was the country’s climate. For our dependent variable we used the CO2 emissions per person in Tons. These variables satisfied the assumptions that the model’s independent variables must be in numerical format and either categorical or continuous. Our initial data also satisfied the assumption for regressions that the dependent variable was continuous. The data was also transformed to include a categorical variable for whether or not a nation had above global average per capita CO2 emissions. Other than the assumptions about the target variable being either categorical for classifiers or continuous for regressions, the Gradient Boosted Techniques have no other assumptions. These models can handle null values well, even though in the group cleaning portion my group imputed any nulls. The models also do not require data standardization or normalization. That means that I did not need to perform any additional transformations for my data to be usable.

**Model Performance and Analysis**

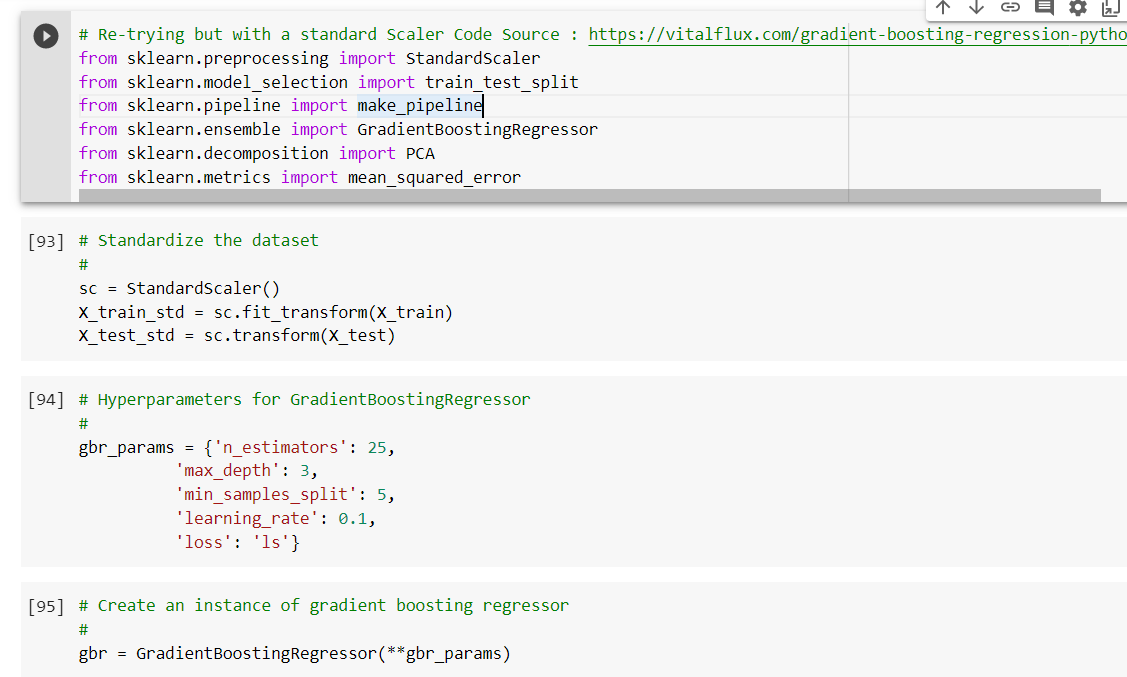
My first attempt was to implement a Gradient Boosted Regression with the default sci-kit learn hyperparameters except for preliminary tuning by iterating through a list of the number of estimators. The performance criteria used to evaluate the regressor model was the R^2 value because the R^2 indicates the amount of change explained by the model, or how effectively the model can predict quantities.

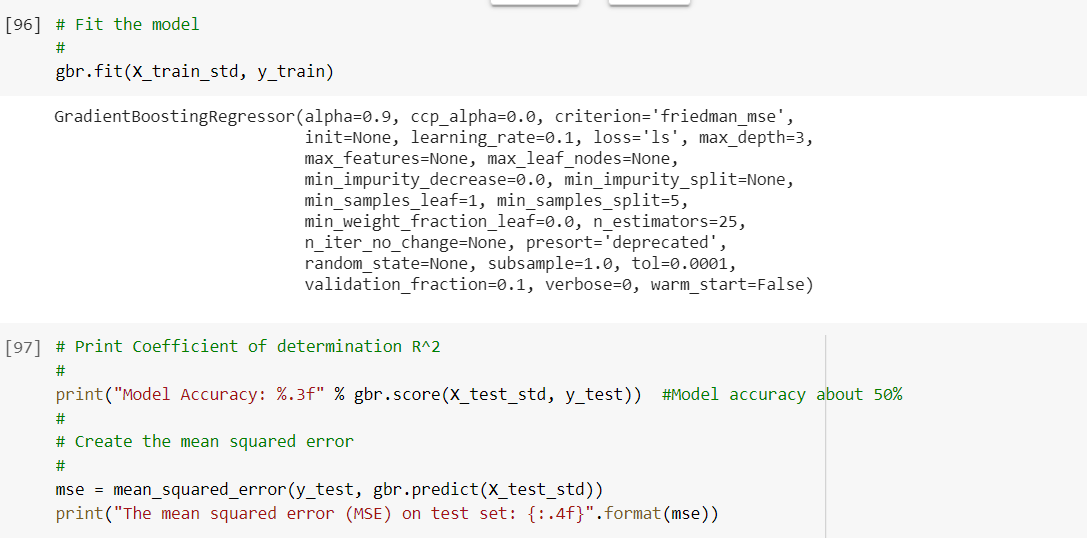
Below is the code used:

Below is the results of the regression model:

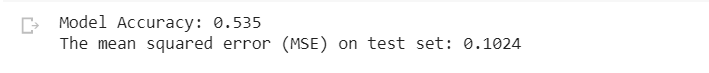
As can be seen in the above results image, the highest R^2 achieved was about 0.53. This score is unsatisfactory as it means the model can barely predict more effectively than a baseline which will have an R^2 of 0.5.

After finding unsatisfactory results, I tried to re-run the Gradient Boosted Regressor with the number of iterations with the highest results (10 estimators) as a Gradient Boosted Regressor with Standard Scaler which would normalize for any scaling issues present in the data. Using Ajitesh Kumar’s blog article “Gradient Boosting Regression Python Examples” as a source for using Standard Scaler (cited in works cited), I applied the code below to run a Gradient Boosted Regression on the scaled data.





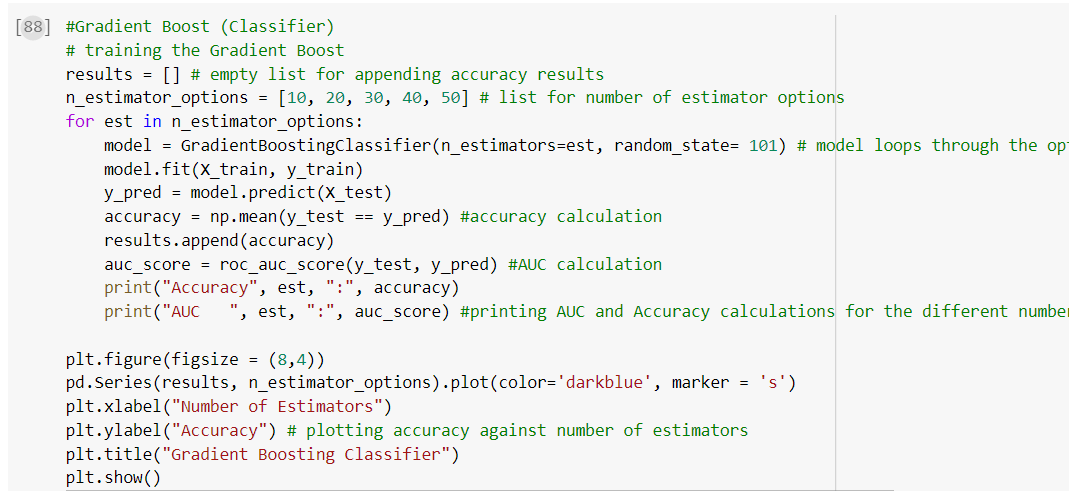
Results Below:



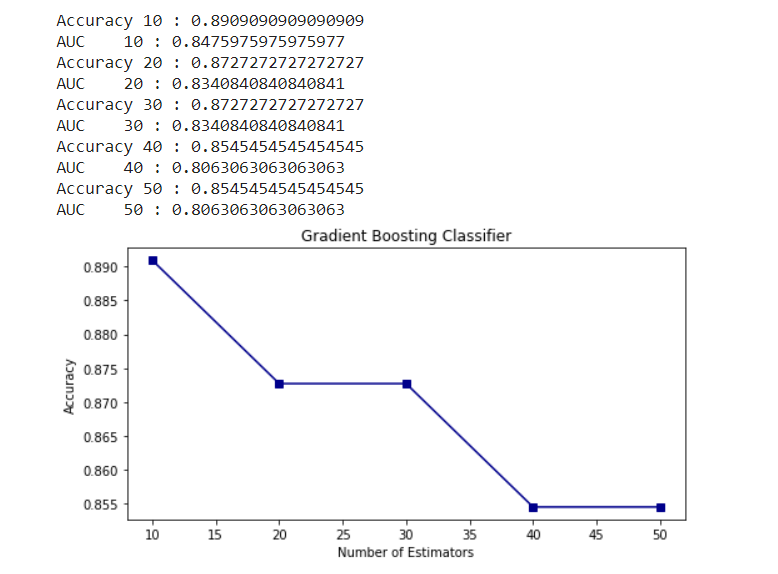
Using the model scoring results, I found that there was an improvement of only 0.005 in the model’s ability to predict. This meant that a Gradient Boosted Regression model with the Standard Scaler applied was not much better than the original Gradient Boosted Regression. Also, this model was not much more effective than the baseline prediction. Due to the weak prediction ability, I concluded that the quantity of CO2 emissions per person may not be predicted but that the class of whether or not a citizen of a country will emit more pollution than the global mean could be predicted. My group created a new classification variable that would indicate 1 if the country’s citizens would emit more CO2 than is the global mean and 0 if it is less. I did not need to do additional data cleaning or transformations on my own because my group mates were also using classifiers. This transformation would enable me to run a Gradient Boosted Classifier to classify the countries by satisfying the assumption that the classifier’s dependent variable must be categorical.

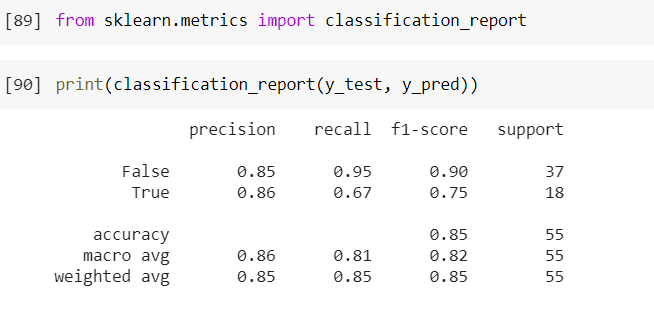
Due to this Gradient Boosted Classifier being an initial model, I used the default hyperparameters other than the number of iterators, which I varied in order to see a preliminary range of results.

The code used is below:



The results are below:





The model’s highest accuracy score was 0.89, with 10 estimators used. This led me to conclude that the data was not suited to predict quantity, but could be used to effectively predict the class of whether or not the CO2 emissions per capita would be above the global average. The classification report describes a more in-depth look at the model’s performance. The accuracy for the overall model, using the default parameters was 0.85 meaning that the Gradient Boosted Classifier could successfully identify the class of a country 85% of the time. The precision for both the True (1) and False (0) outcomes were also in line with the accuracy, which meant that the target class of above average or below average per person CO2 emissions was successfully predicted correctly 85% or 86% of the time as well. The recall for the class indicating the emissions per capita would be below average was also extremely high at 0.95, meaning that 95% of the time this model could correctly classify the below average emissions countries without incorrectly classifying them as above average. The most troubling score was the relatively low Recall for the above average CO2 emissions per capita class. The 0.67 result meant that would misidentify the above average emissions about one third of the time. However, the 0.90 F1 Score for classifying below average CO2 emissions and the 0.75 F1 score for above average meant that this model was relatively effective at correctly classifying the CO2 emissions per capita classes for each country. Overall, the classifier was my strongest model because it had consistently performed more effectively than the baseline of 0.5 and much more effectively than my regressor models which were only able to accurately predict about 53% of the quantity change caused by the different variables.

**Model Conclusion**

The Gradient Boosted Classifier was much more effective than the Gradient Boosted Regressor. The regression models I implemented could barely outperform the baseline estimate for the R-Squared, which should be above 0.5. On the other hand, the classifier was able to perform well above the baseline accuracy of 50% with minimal tuning for the number of iterations. The Classifier model was also effective because its Recall, Precision, and F1 scores were also high. This means the model was not very prone to misclassifying and was able to effectively balance that ability with its positive predictive value. The classifier also carried the benefit of showing which features were most important. In the case of this dataset, it identified GDP and Phone ownership as the two most important features. Although I cannot use feature importances to draw conclusions for the group, this additional feature could be helpful in group analysis if this model is chosen.

Works Cited

Kumar, A. (2020, December 14). *Gradient Boosting Regression Python Examples*. Data Analytics. https://vitalflux.com/gradient-boosting-regression-python-examples/