Wine Review Prediction

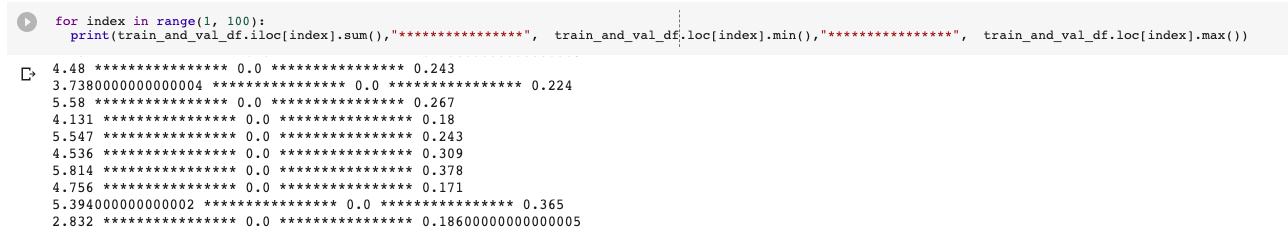
After providing the material for the case-review, a quick review was made on the data.

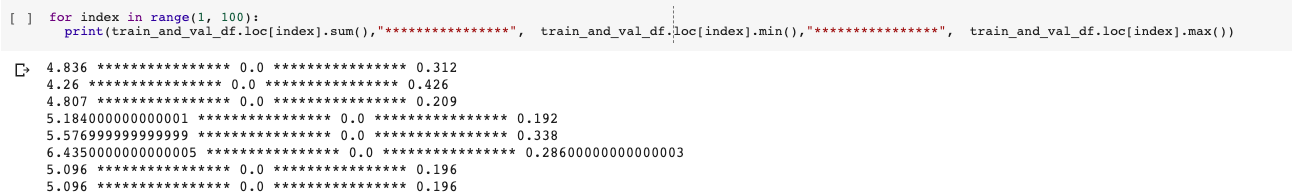
The preprocess process was started in the dataset where wine quality scores were estimated through various features.

First of all, the feature text file was placed as a column in the train data. Then the column of label scores was appended to the train data.

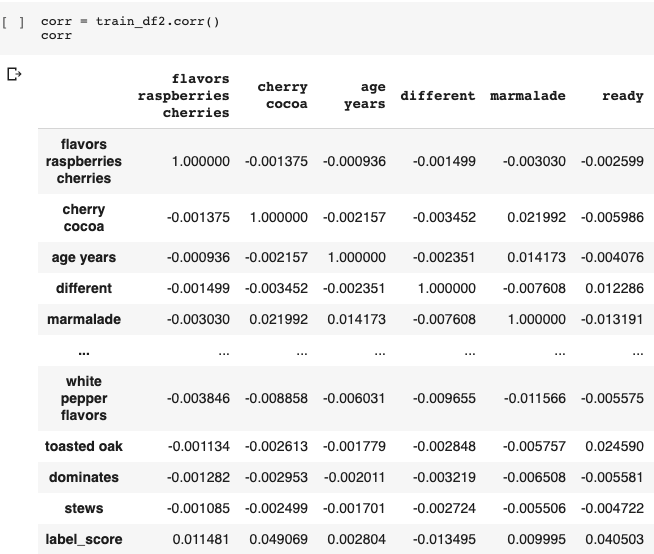
This notebook was executed on google.colabs so if you also use this way, you have to execute “from google.colab import files” and “uploaded = files.upload()” command step by step.(In this step, please be careful because colabs has a bug, sometimes you have to execute “upload()” method twice.

While preprocess step, at columns and rows, any constant summation wasn’t found. We can understand by this information, every columns and rows are independent from each other.



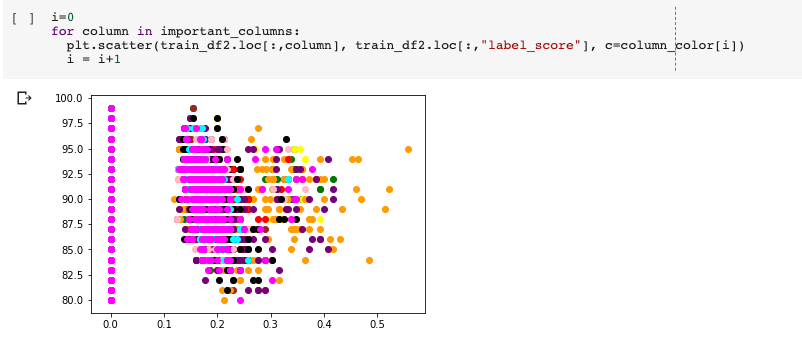


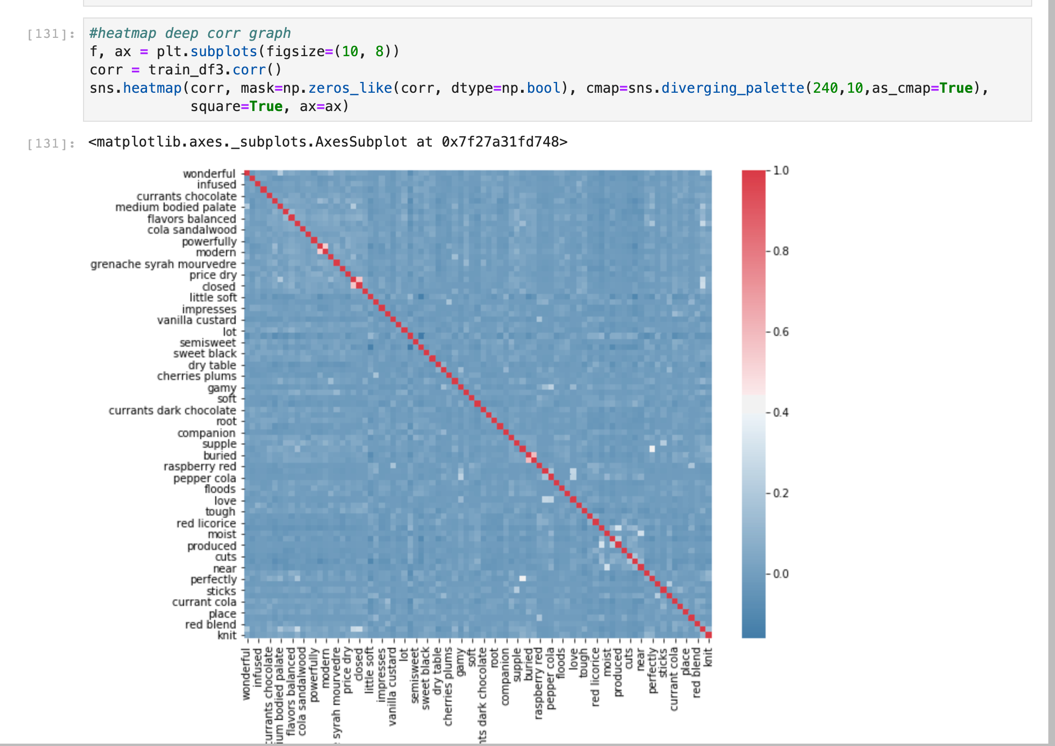
Secondly, in train data, I tried to understand correlation between columns. At the final of this step, I collect high ratio correlated columns. After this step, I worked on this columns to predict label score.



Selected columns

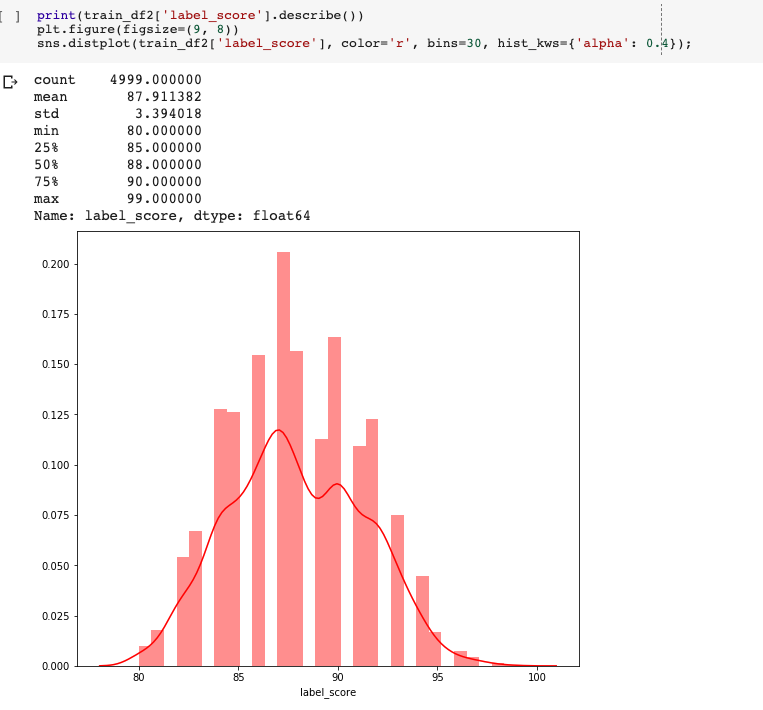






Heatmap Corr

Then in the step, I tried to make sense of the validation data. I wanted to understand what kind of character he was.



After preprocess, I started to try different regression models. I think it is regression problems because;

1)There is no time dimension, so it isn’t basic time-series

2)At the beginning, I contradicted between regression and classification but the facts below are directed to me regression solution.

***\*a.That predictive modeling is about the problem of learning a mapping function from inputs to outputs called function approximation.***

***b.That classification is the problem of predicting a discrete class label output for an example.***

***c.That regression is the problem of predicting a continuous quantity output for an example.***

\*Ref: <https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/>

Test 1) #Linear Regression Model



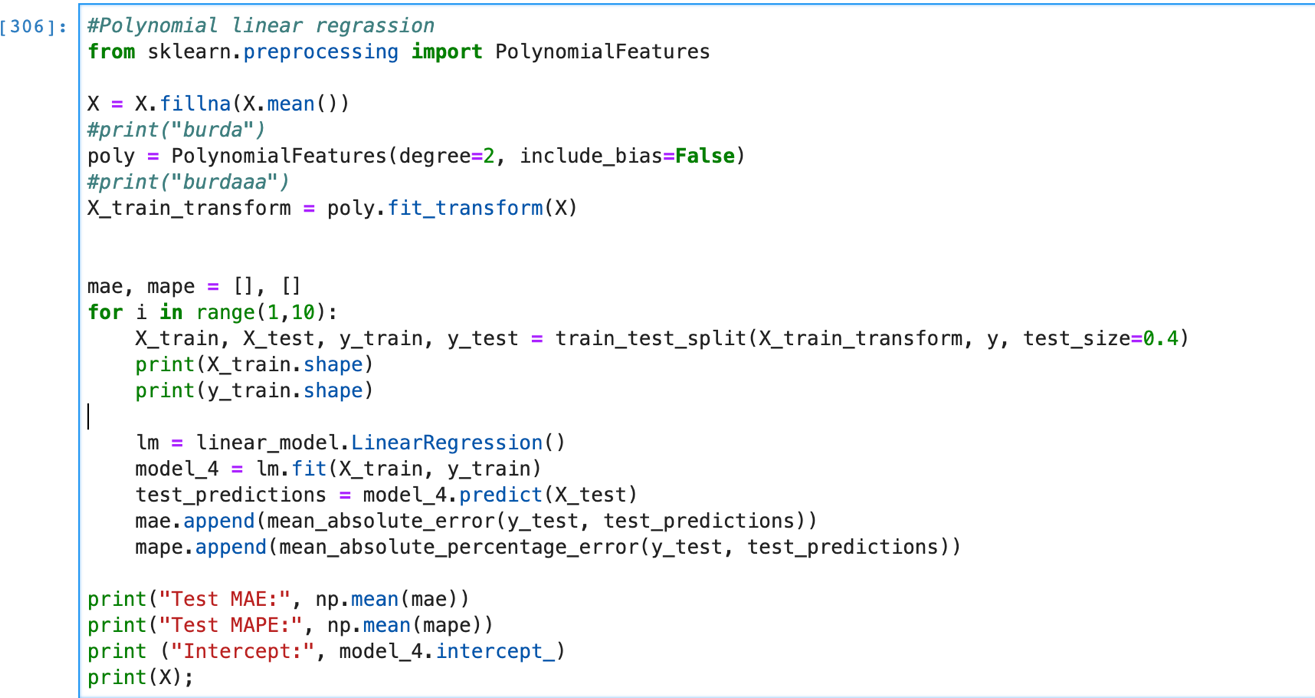
Results :

Test MAE: 1.8722902468710647

Test MAPE: 2.125544086691307

Intercept: 87.15742460460439

Test 2) #Polynomial linear regression



Results :

Test MAE: 10905336498.012377

Test MAPE: 12263723371.917559

Intercept: 86.63691958133074

Test 3) #XGBRegressor Model

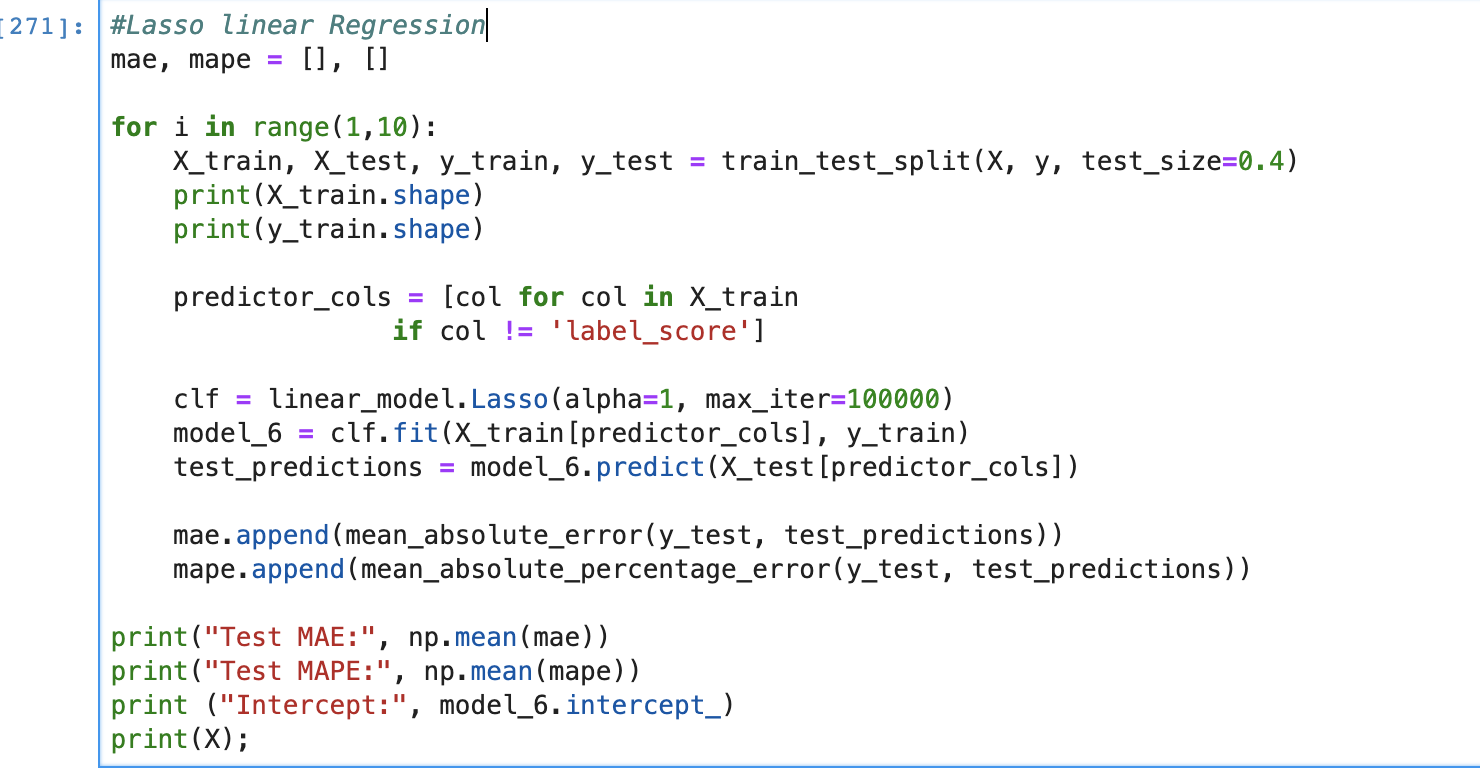


Results :

Test MAE: 2.015515528784858

Test MAPE: 2.2885070079860017

Test 4) #Lasso linear Regression



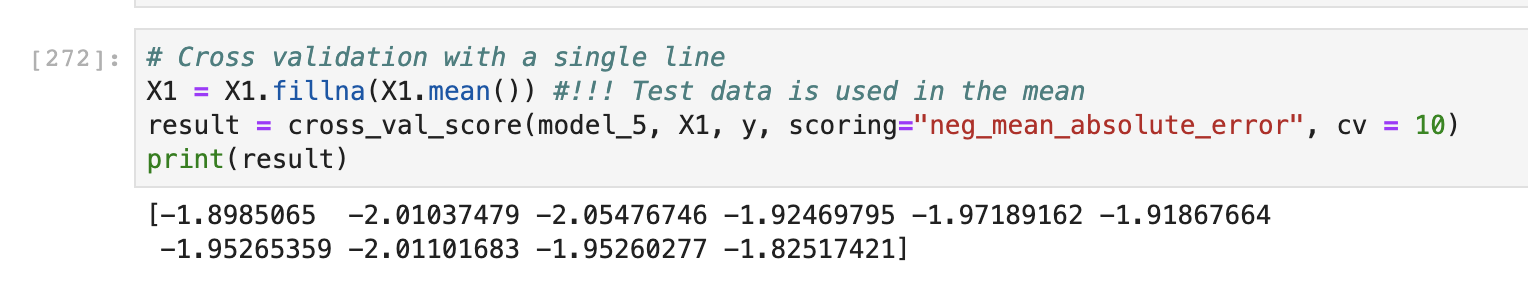
Results :

Test MAE: 2.78211685376607

Test MAPE: 3.166863134113734

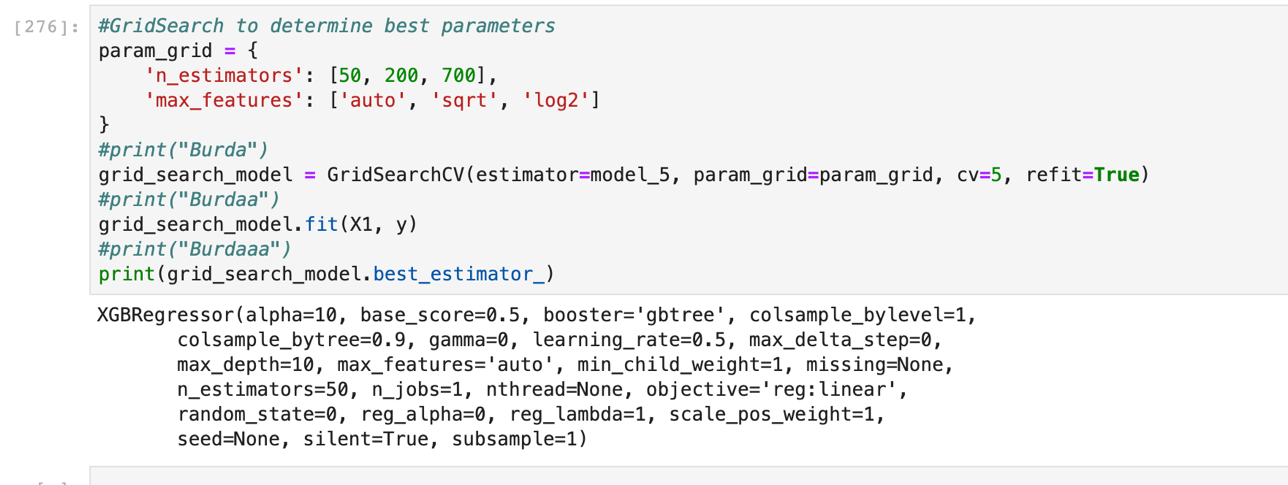
Intercept: 87.91030343447817

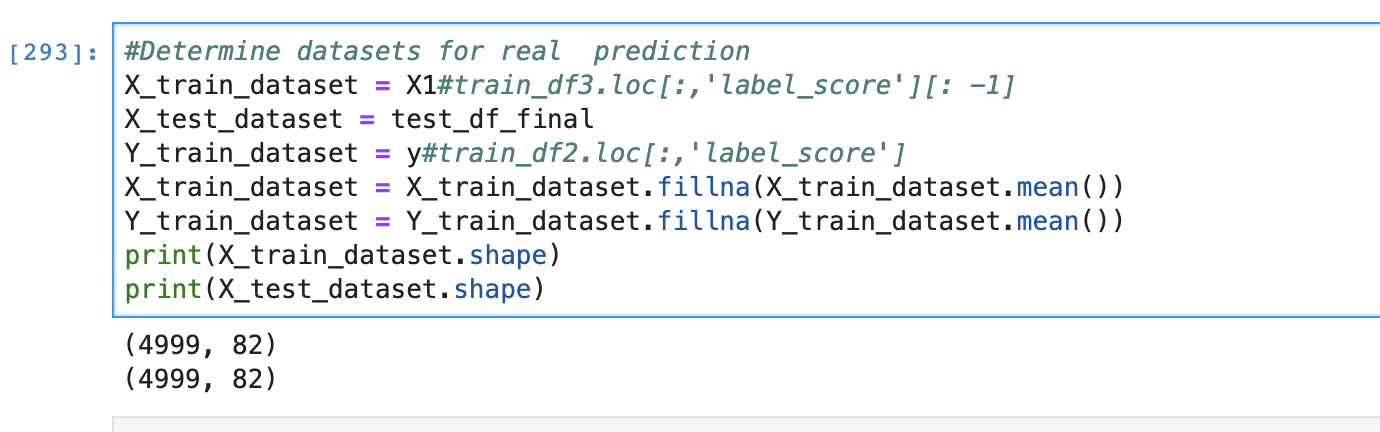
In addition, I checked the dataset in point of view of cross validation.



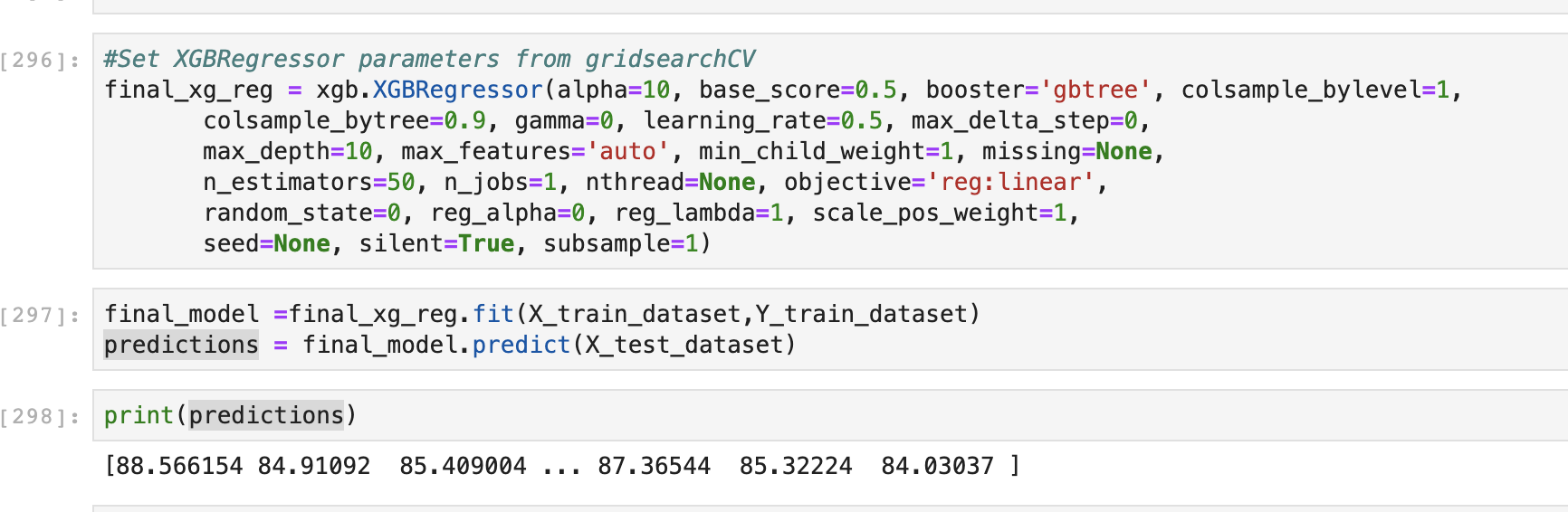
Conclusion:

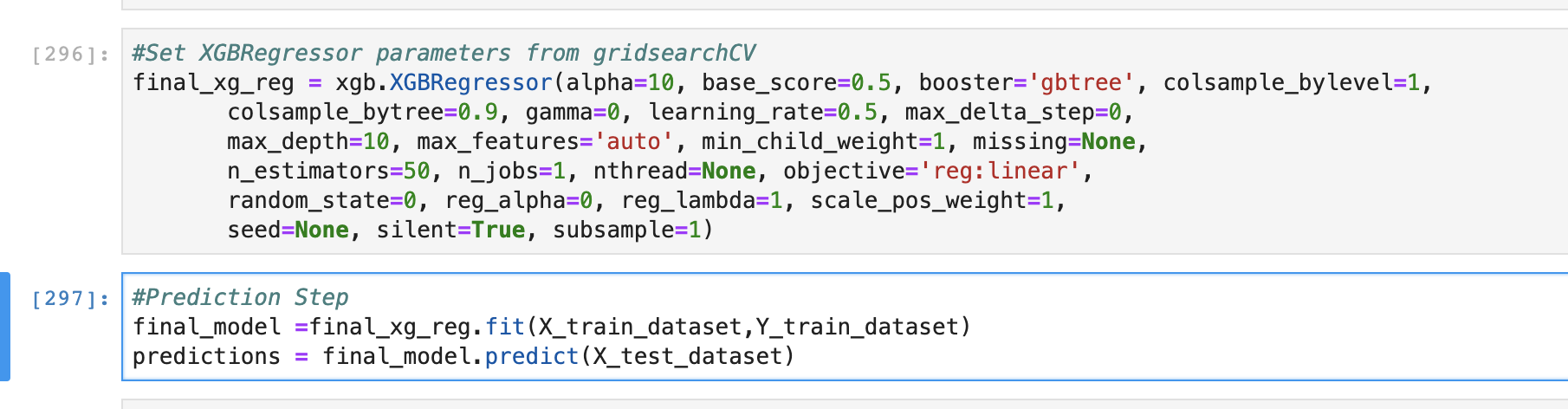
According to the experiment results, I choose XGB Regressor method. After this choise, I found best parameters by using grid search method.





I used gridsearch CV method’s results as a parameter of regenerate my XGBRegressor method.





At the end of my experiment, result plots can be seen. When we compare the prediction plot and test data plot, It is not difficult to notice the similarity.

