|  |  |
| --- | --- |
|  | **Introduction to Business Data Analytics** |

|  |  |
| --- | --- |
| **Homework #4 Part 2** |  |

\_\_\_Jane Doe\_\_\_\_

(put your name above (incl. any nicknames) )

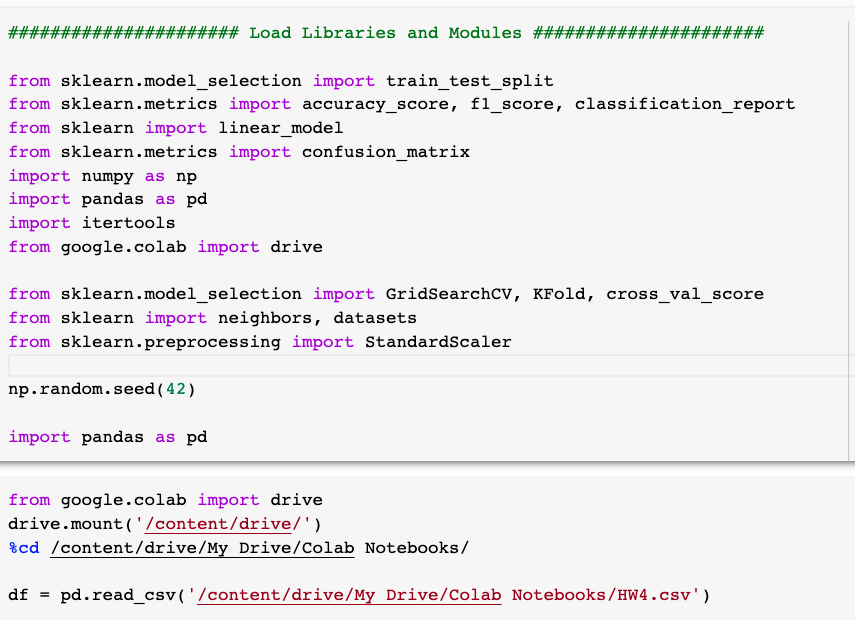
**Use numeric prediction techniques to build a predictive model for the HW4.xlsx dataset. This dataset is provided on Canvas and contains data about whether or not different consumers made a purchase in response to a test mailing of a certain catalog and, in case of a purchase, how much money each consumer spent. The data file has a brief description of all the attributes in a separate worksheet. We would like to build predictive models to predict how much will the customers spend; Spending is the target variable (numeric value: amount spent).**

**Use Python for this exercise.**

**Whenever applicable use random state 42 (10 points).**

1. **After exploring the data, build numeric prediction models that predict Spending. Use linear regression, k-NN, and regression tree techniques. Briefly discuss the models you have built. Use cross-validation with 10 folds to estimate the generalization performance. Present the results for each of the three techniques and discuss which one yields the best performance.**

Load Data and Locate Features

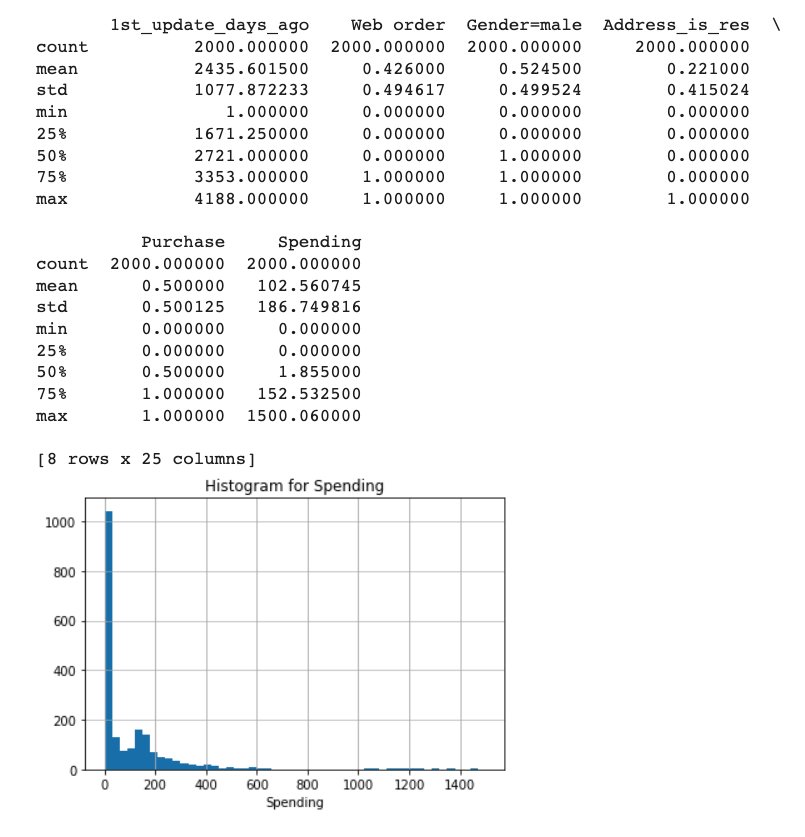


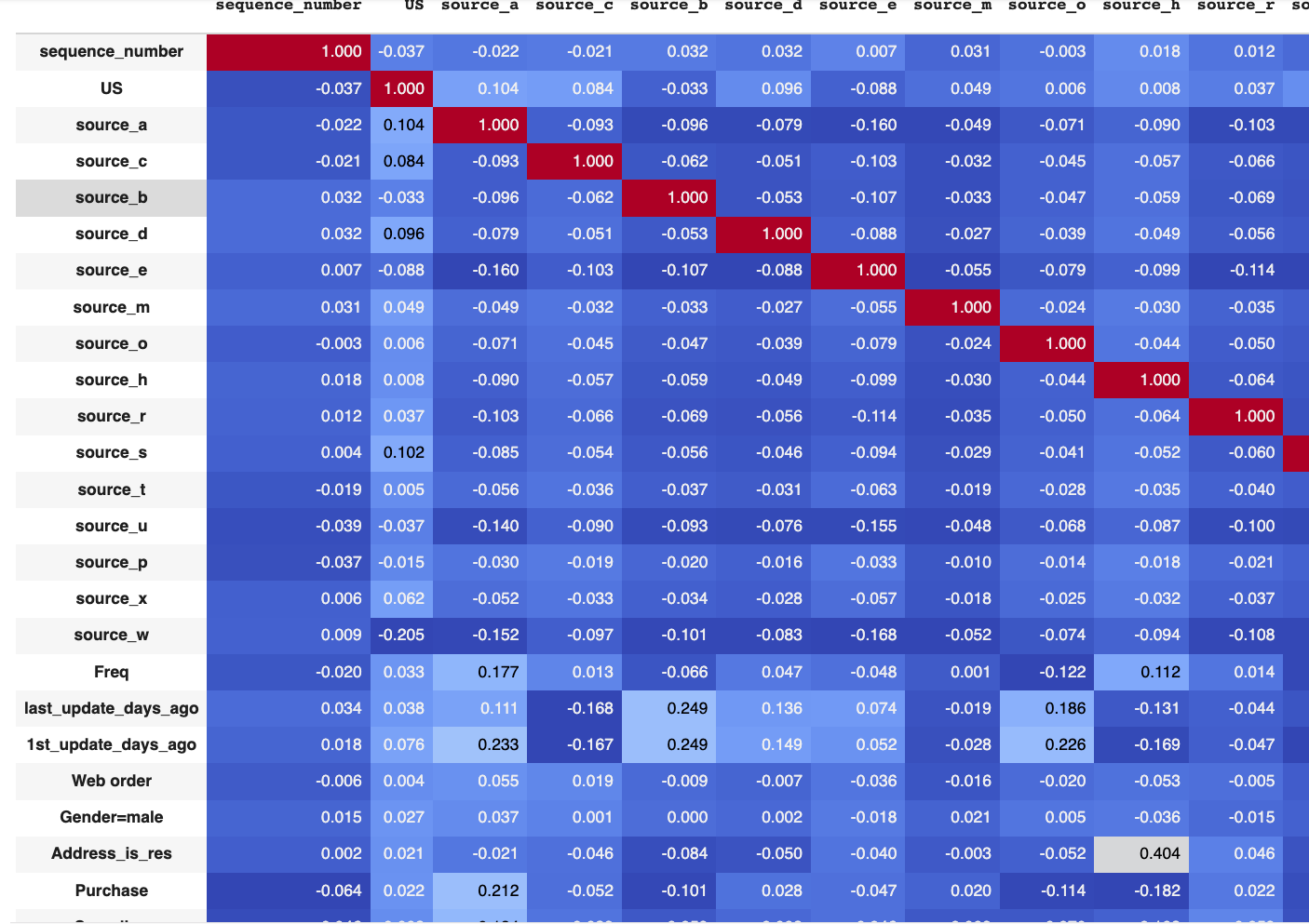
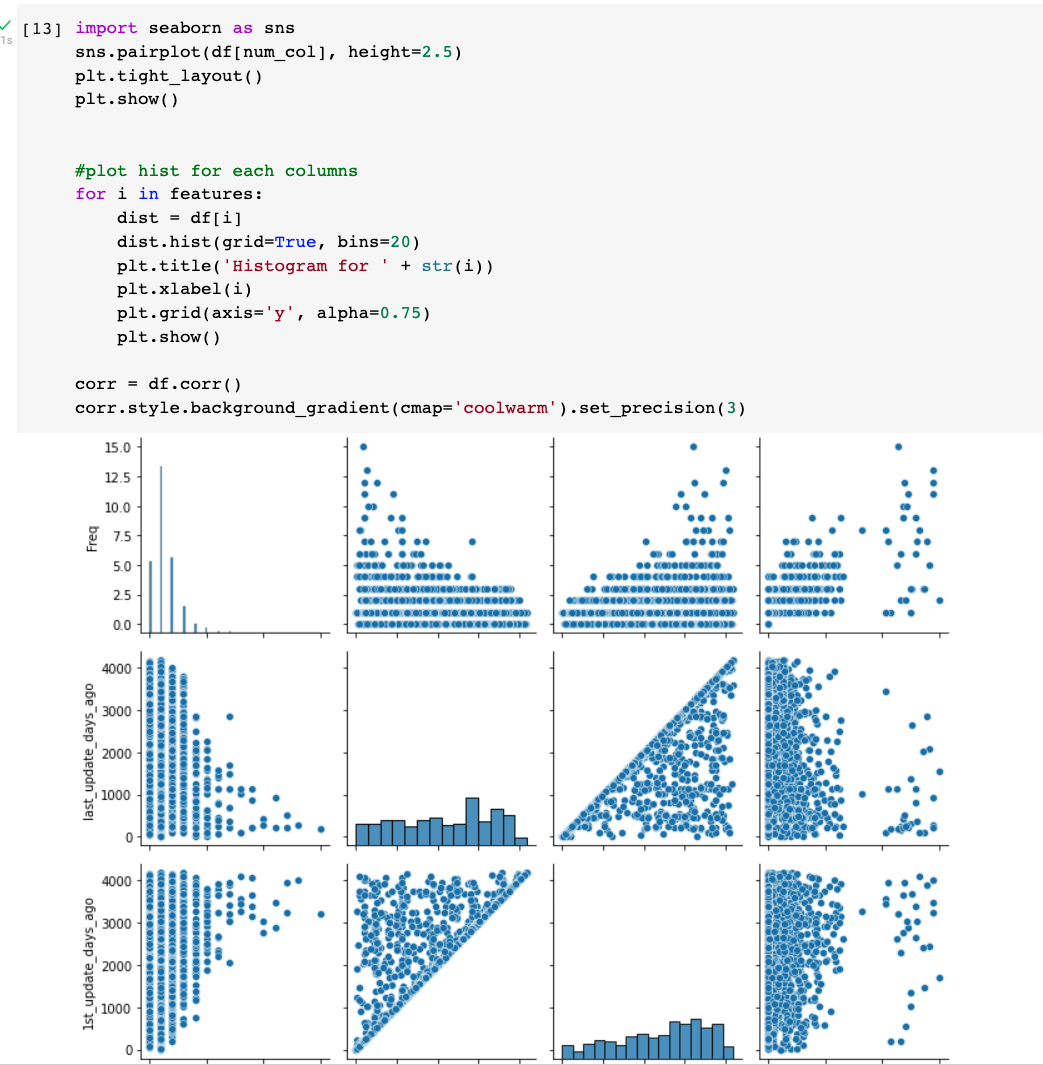
Define Target Variable and Features

Explore the Data: Descriptive Statistics and Visualizations

Table

Description automatically generated

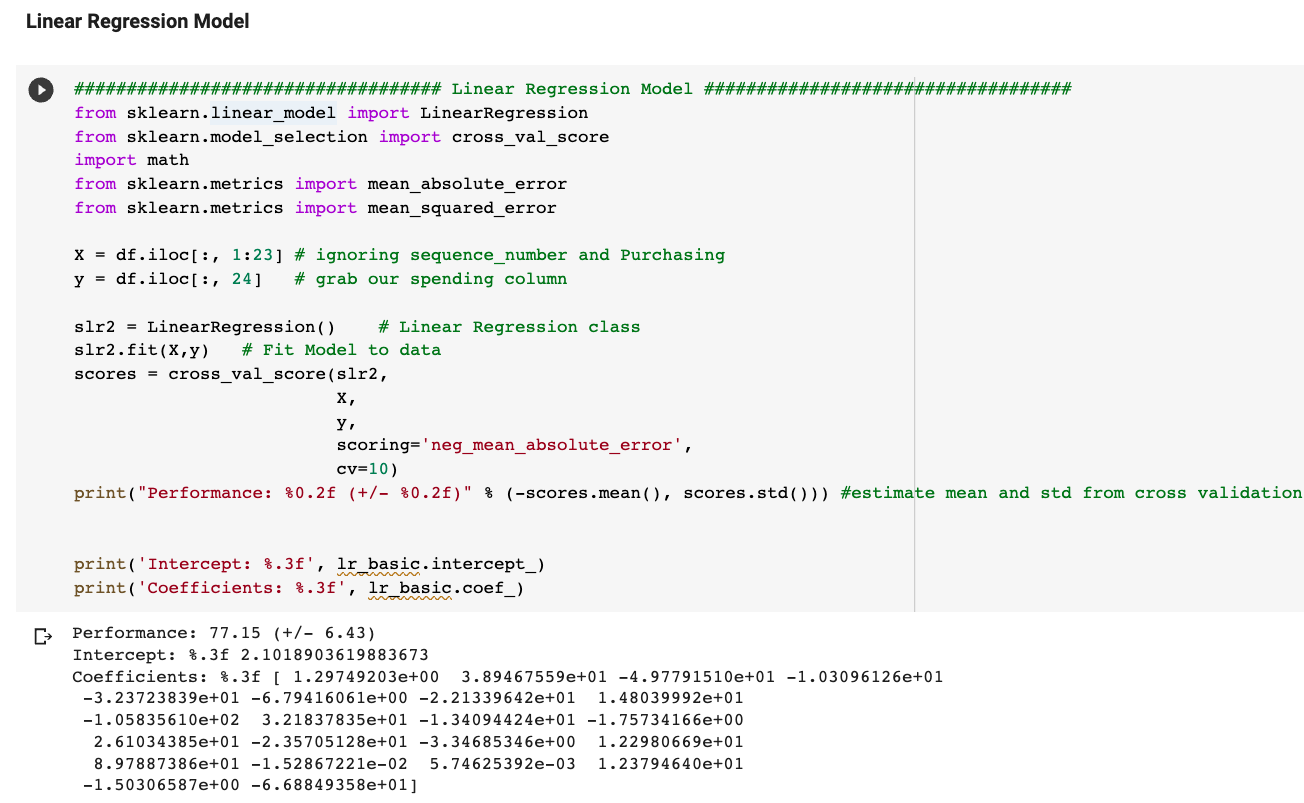


Visualizations

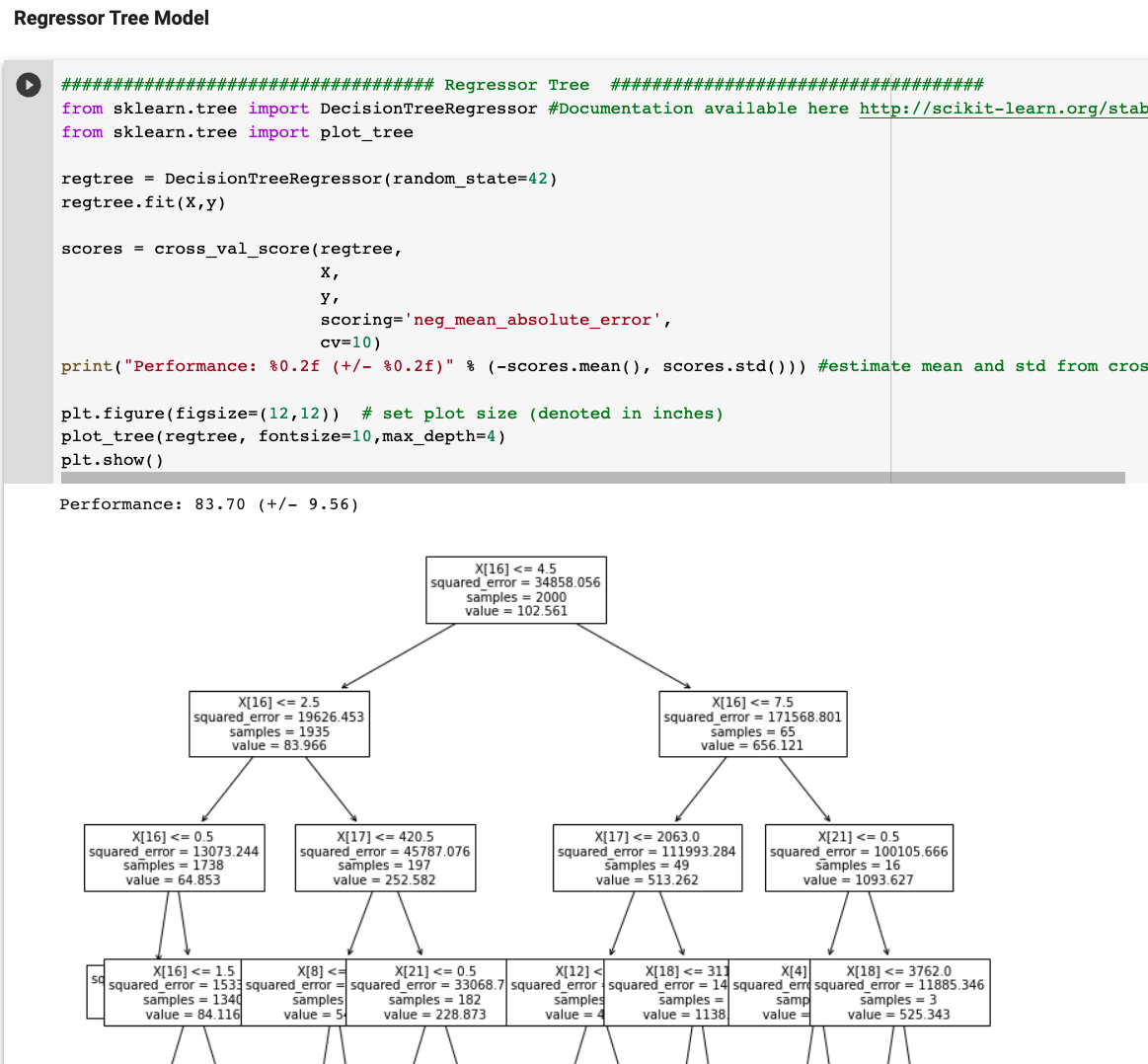
**Question a**

We will build a simple linear regression model, a kNN Regressor model, and a Regressor Tree model. After carefully inspecting the data, we have to exclude the feature sequence\_number as a predictor from our data science models as it doesn’t have any predictive power. Similarly, we have to exclude the feature Purchase since it would lead to data leakage.

For all the models, we will be using MAE as our performance metric and cross-validation with 10 folds to estimate the generalization performance of the models.



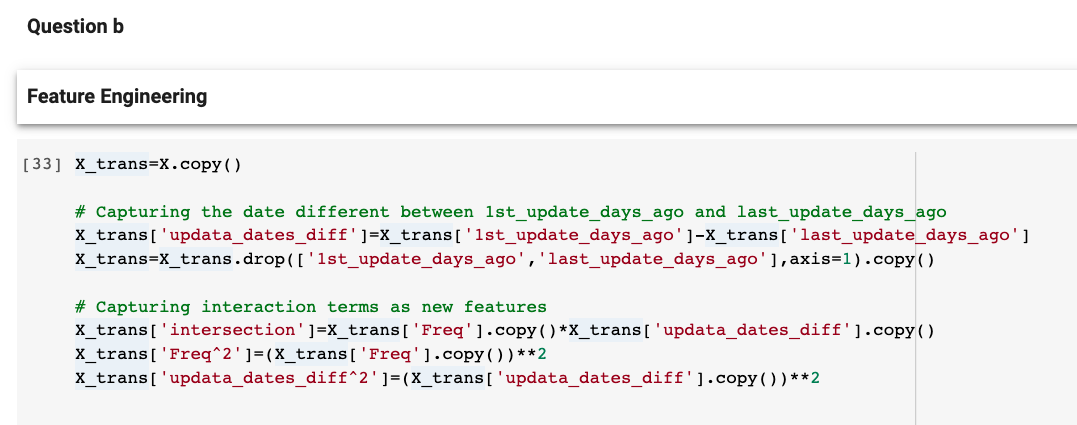




So far, we would prefer the Simple Linear Regression model because it has the lowest MAE score (based on cross-validation) among the three models.

1. **Engage in feature engineering (i.e., create new features based on existing features) to optimize the performance of linear regression, k-NN, and regression tree techniques. Present the results for each of the three techniques (choose the best performing model for each technique in case you try multiple models) and discuss which of the three yields the best performance. Use cross-validation with 10 folds to estimate the generalization performance. Discuss whether and why the generalization performance was improved or not.**

There are many possibilities to engage in feature engineering. In the following sample answer, we provide a couple of examples.



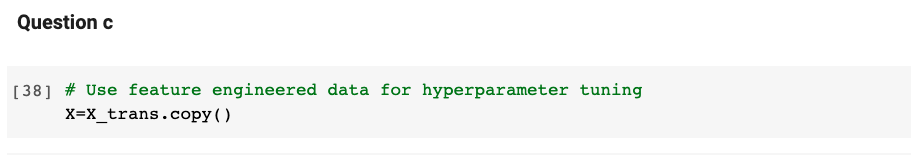
In the next steps, please make sure they use the feature engineered data (In my case it has been stored as X\_trans, y)



We notice that across all the three models, the performance improved as a result of feature engineering, particularly for the simple linear regression model and the kNN regressor model. The best model is still the Linear Regression model with a generalization performance (MAE) of 73.76. The generalization performance improved because we have captured some complex patterns in the data with feature engineering. In other words, it seems that increasing the complexity of the model with non-linear features, has improved the generalization performance of the models.

1. **Engage in parameter tuning to optimize the performance of linear regression, k-NN, and regression tree techniques. Use cross-validations with 10 folds to estimate the generalization performance. Present the results for each of the three techniques and discuss which one yields the best performance.**

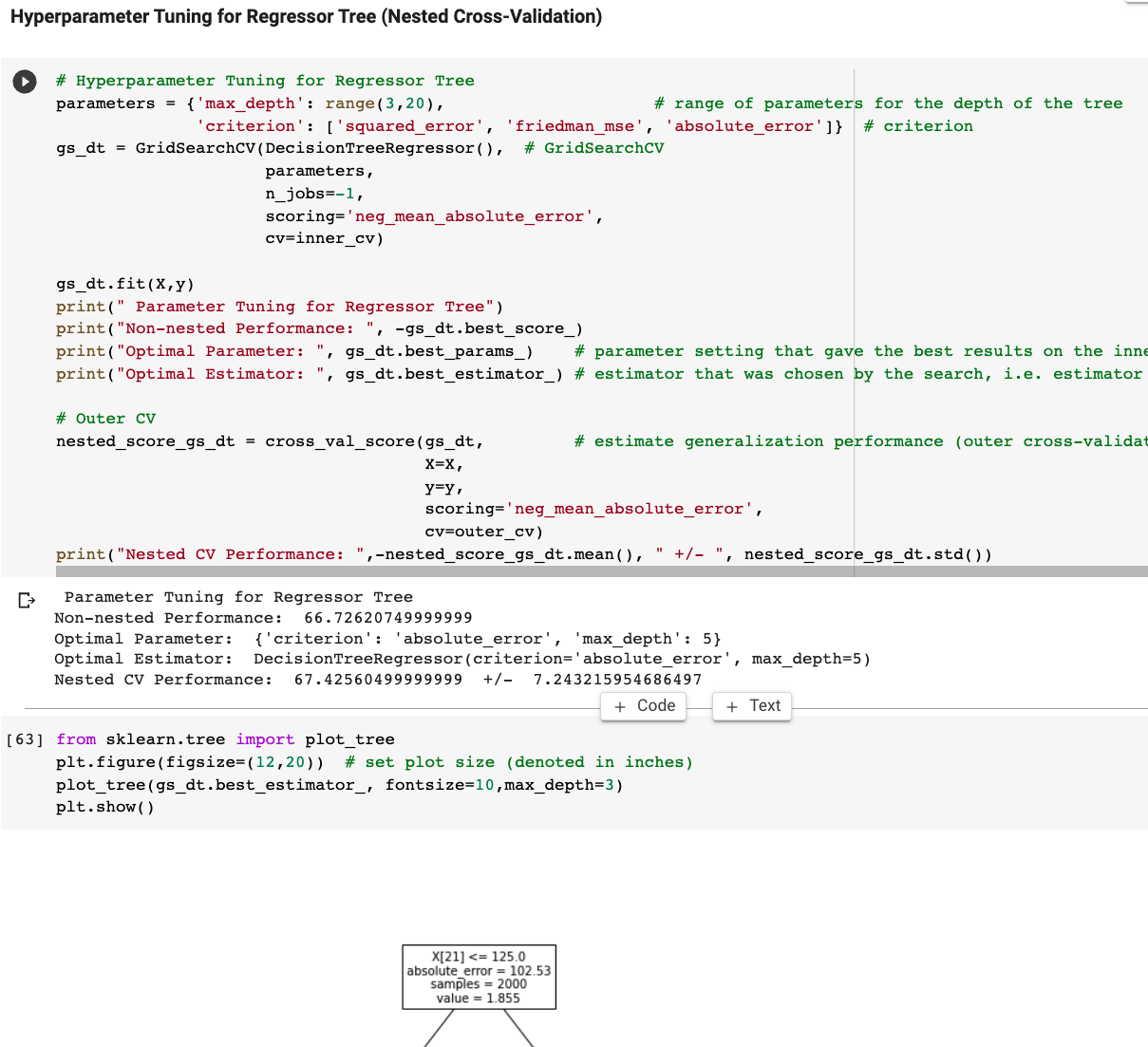
Since feature engineering improved the generalization performance of the models, we will proceed with the hyperparameter tuning on the feature engineered data.



Text

Description automatically generated with medium confidence





After engaging in hyperparameter tuning using the nested cross validation technique, the best performing model is now the Regressor tree model with a generalization performance of a mean MAE of 67.43.