Assignment 4

Negin Baghbanzadeh

Question 1

Changes that I made:

1. Added a pca and a min-max-scaler to the stages of the existing pipeline.
2. Created a new pipeline using pca, min-max-scaling and linear regression model.
3. Changed tuning parameters.

(It took over 7 hours to run the first pipeline, I didn't have time to run the second one before publishing the notebook, I'm really sorry)
https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/290121495876716/3477380901255997/4913019624089264/latest.html

Question 2

I tried to use two different models for this question, a random forest model and a linear regression model.

For each model, I created a pipeline. The pipelines have some stages.

Since we have some categorical features such as "clarity", we need to have an one-hot-encoder stage. The input value of OneHotEncoder should be an Integer so in order to use this transformation, we need to convert string to int, so we need an indexer stage which is StringIndexer.

The numerical features need standardization so we need a min-max-scaler stage too.

I also wanted to add a feature reduction stage to the pipeline, but the performance decreased a lot when using feature reduction so, I decided not to use it in the random forest pipeline but I used it in the linear regression pipeline.

https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/290121495876716/2656493038758O1/4913019624089264/latest.html