Report on ML cycle

- First split the data to training data and test data
- Encoding the training data and make it ready using full pipeline
- We train the data with linear regression model and measure the performance
- We found that the error is too big (underfitting) that is because the model is too simple for data
- We train the data with Decision Tree Regressor model and measure the performance
- We found that the root mean square error equals to zero that is because there is an overfitting but we will clear it with cross validation
- We apply the cross validation and found the mean of the score = 68848.18979613911 which is too big
- And the same with liner regression and cross validation
- We train the data with random forest regressor model and make a cross validation and found the mean score = 49835.16567770626 it is lower than the others
- We use the grid search to find the best model and found that the best estimator = random forest regressor
- And make grid search to find the best relative attributes which are(max_features=8, n_estimators=30, random_state=42)
- After we trained the model we test it with test data
- Making the test data prepared using full pipeline transformation
- And evaluating the model on the test data
- Using scipy library we compute the 95% confidence intervals computing the lower and upper bounds