Columbia University in the City of New York
MATHGR5430 Machine Learning for Finance, Spring 2022

Homework 5

Prof. Silva

Due by 05/01/2022

Please submit a single .pdf-file generated from your Jupyter-Notebook on CourseWorks.

1 Background

This famous dataset contains housing prices from the boston area with their respective features. The feature 'MEDV' is the response variable. The predictors have different weight and we will revisit multiple regression approaches as well as feature engineering to get a low MSE on the test set. For this homework, we will give you less clear instructions on what to do but more so use your own developed skills in data analysis to get a low error. This homework should relatively straight forward to do but will make you rethink important concepts from the lectures.

2 Exercises

For all of the regression techniques you run: Please report the MSE (other information is optional)

You can use the following code to report MSE: np.square(np.subtract(y_test, y_pred)).mean()

1. Run a univariate Linear Regression with the feature that should have the best predictive power.

2. Run a m multivariate Linear Regression with the two features that should have the best predictive power.

3. Using the metric you have used before for feature selection: Determine whether you would want to run a Ridge, Lasso or ElasticNet regression next. Run the regression and comment on your choice.

4. Now, perform EDA and Feature Engineering. You are free to do whatever you think might improve performance, but at least:

   (a) Examine, and possibly influence distribution of the response variable, look at outliers. (Do you do that only for y_train or the entire y?

   (b) Handle outliers of each feature when plotted against the response using seaborn's jointplot: sns.jointplot(x = df['feature'], y = df['label'], kind = "reg") (Do you do that only for X_train or the entire X?

5. Run a Random Forest Regressor

   Feel free to run any other regression models that you might think perform better on the data. You should easily be able to get an **MSE below 0.07** (using the above formula).