CIS053 Final Project

Mission College Summer 2023

Instructor: Jahan Ghofraniha

1. In this problem your goal is to build a predictive model for the real estate dataset (Real estate.csv). Before any modeling attempt, you will need to perform EDA on the dataset to get some insight into the data. The price per square unit area is the target (y values/output) and the rest of the columns should be treated as inputs. Perform the following steps (total 100 pts):
    a. EDA of the original dataset (load, clean, separate into X & Y, descriptive stats, histograms, correlation analysis, scatter plots)
    b. Based on the findings/observations in step a, decide whether the dataset requires any preprocessing (normalization or standardization). State the reason for your choice of preprocessing technique or if you decide not to do anything at this step.
    c. Build a multilinear regression model (you can use RFE or other methods. Make sure your assessment of the model quality is based on a test performance measure such as test MSE or test RSS)
    d. Build a regularized version of the regression model (use both Lasso and Ridge methods) and compare the results with step c.
    e. Use cross-validation to compare the results of all models and choose the best model based on test performance measure.
    f. Justify your answer based on your understanding of how cross-validation works and in the context of bias-variance trade off.
    g. Explain if you find any discrepancy between the results in the cross-validation step and steps c and d.
    h. Upload your Python code and the explanation of the results plus graphics in a pdf file (you will submit two files, a python file and a pdf file). You can alternatively submit a Jupyter notebook plus a pdf version of the notebook (two files)