Assignment-2 Applied Machine Learning

Regression on Ames Housing Dataset

For this assignment, we will use the Ames housing dataset. You can access the data from following link.

http://www.amstat.org/publications/jse/v19n3/decock/AmesHousing.xls

You can find a description of the variables here. Please pay attention to it.

http://jse.amstat.org/v19n3/decock/DataDocumentation.txt

Wherever needed, write down your observation as comments in the notebook.

Take note that for categorical variables, NA here does not mean a missing value, but should be treated as a separate category.

1. Visualize the univariate distribution of each continuous, and the distribution of the target. Do you notice anything? Is there something that might require special treatment? [10 Marks]

2. Visualize the dependency of the target on each continuous feature (2d scatter plot). [10 Marks]

3. Split data in training and test set. Do not use the test-set unless for a final evaluation in **question 6**. For each categorical (nominal) variable, cross-validate a Linear Regression model using just this variable (one-hot-encoded). Visualize the relationship of the categorical variables that provide the best R^2 value with the target. [20 Marks]

4. Use ColumnTransformer and pipeline to encode categorical variables. Evaluate Linear Regression (OLS), Ridge, Lasso and ElasticNet using cross-validation with the default parameters. Does scaling the data (within the pipeline) with StandardScaler help? Read about ColumnTransformer below. [30 Marks]
https://scikit-learn.org/stable/modules/compose.html#column-transformer

5. Tune the parameters of the models using GridSearchCV. Do the results improve? Visualize the dependence of the validation score on the parameters for Ridge, Lasso and ElasticNet. [20 Marks]

6. Visualize the coefficients of the resulting models. Do they agree on which features are important? [10 Marks]