My Name (myNetID)

IE598 MLF F18

Module 4 Homework (Regression)

Part 1: Exploratory Data Analysis

Describe the data sufficiently using the methods and visualizations that we used previously in Module 3 and again this week. Include any output, graphs, tables, heatmaps, box plots, etc. Label your figures and axes. DO NOT INCLUDE CODE!

Split data into training and test sets. Use random\_state = 42. Use 80% of the data for the training set. Use the same split for all models.

A close up of a map

Description automatically generated

LASAT has a negative relationship with MEDV

LASAT has a negative relationship with RM

RM has a positive relationship with MEDV

A picture containing drawing

Description automatically generated

RM has a positive relationship with MEDV

LASAT has a negative relationship with MEDV

NOX and INDUS have a positive relationship with MEDV

NOX have a positive relationship with INDUS

RM has a positive relationship with MEDV

A close up of a map

Description automatically generated

RM does not follow exactly normal distribution, it has fatter tails.

Part 2: Linear regression

Fit a linear model using SKlearn to all of the features of the dataset. Describe the model (coefficients and y intercept), plot the residual errors, calculate performance metrics: MSE and R2.

MSE train: 20.613, test: 26.630

R^2 train: 0.763, test: 0.637

coefficients [ 0.81576845 -0.07167494 0.1626726 -0.03095303 -0.21689699 -0.17436198

0.18945189 -0.07986522 -0.04284088 -0.29269901 -0.27839046 0.13052308

-0.03917064 -1.13411737 0.74827014 0.27727456 0.70802424 -2.00127056

3.17561044 -0.17974237 -3.12986203 2.20304907 -1.74483928 -1.98596453

1.19684554 -3.55844917]

y intercept 22.796534653465383

A screenshot of a cell phone

Description automatically generated

Part 3.1: Ridge regression

Fit a Ridge model using SKlearn to all of the features of the dataset. Test some settings for alpha. Describe the model (coefficients and y intercept), plot the residual errors, calculate performance metrics: MSE and R2. Which alpha gives the best performing model?

MSE train: 20.695, test: 26.742

R^2 train: 0.762, test: 0.635

coefficients [ 0.78962525 -0.09285859 0.1494771 -0.03755549 -0.23018224 -0.15621739

0.19918585 -0.08477565 -0.05085311 -0.29025028 -0.26417314 0.13093716

-0.05076515 -1.04413409 0.60300853 0.09149889 0.73612899 -1.7360696

3.21130716 -0.18547792 -2.80822413 1.63343591 -1.25197363 -1.90846369

1.17161784 -3.45485259]

y intercept 22.796534653465383

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

the best alpha is 9.385

Part 3.2: LASSO regression

Fit a LASSO model using SKlearn to all of the features of the dataset. Test some settings for alpha. Describe the model (coefficients and y intercept), plot the residual errors, calculate performance metrics: MSE and R2. Which alpha gives the best performing model?

MSE train: 22.239, test: 28.774

R^2 train: 0.744, test: 0.608

coefficient [ 0.56292289 -0. 0. -0. -0.09121884 -0.

0.03180995 -0. -0. -0.1363641 -0.04606502 0.

-0. -0.57702188 0. -0. 0.63214341 -1.01476427

3.3878616 -0. -1.59174286 0. -0. -1.72618034

0.97731782 -3.54773182]

intercept 22.796534653465383

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

the best alpha is 0.208

Part 4: Conclusions

Write a short paragraph summarizing your findings.

According to EDA, certain attributes have certain relationship with response MEDV. From MSE test>train, and R^2 train>test, so the models do not overfit. But it seems Lasso can be used to select features, but I used my best alpha to fit the model, it cannot thin out all of these features, instead, I just thin out some other features. I don't know why. And from these residual plots, I cannot see certain types which means I do not omit any important information.

Part 5: Appendix

Link to github repo