EMRAH SARIBOZ

HW-5 Report

MSE and run-time analysis

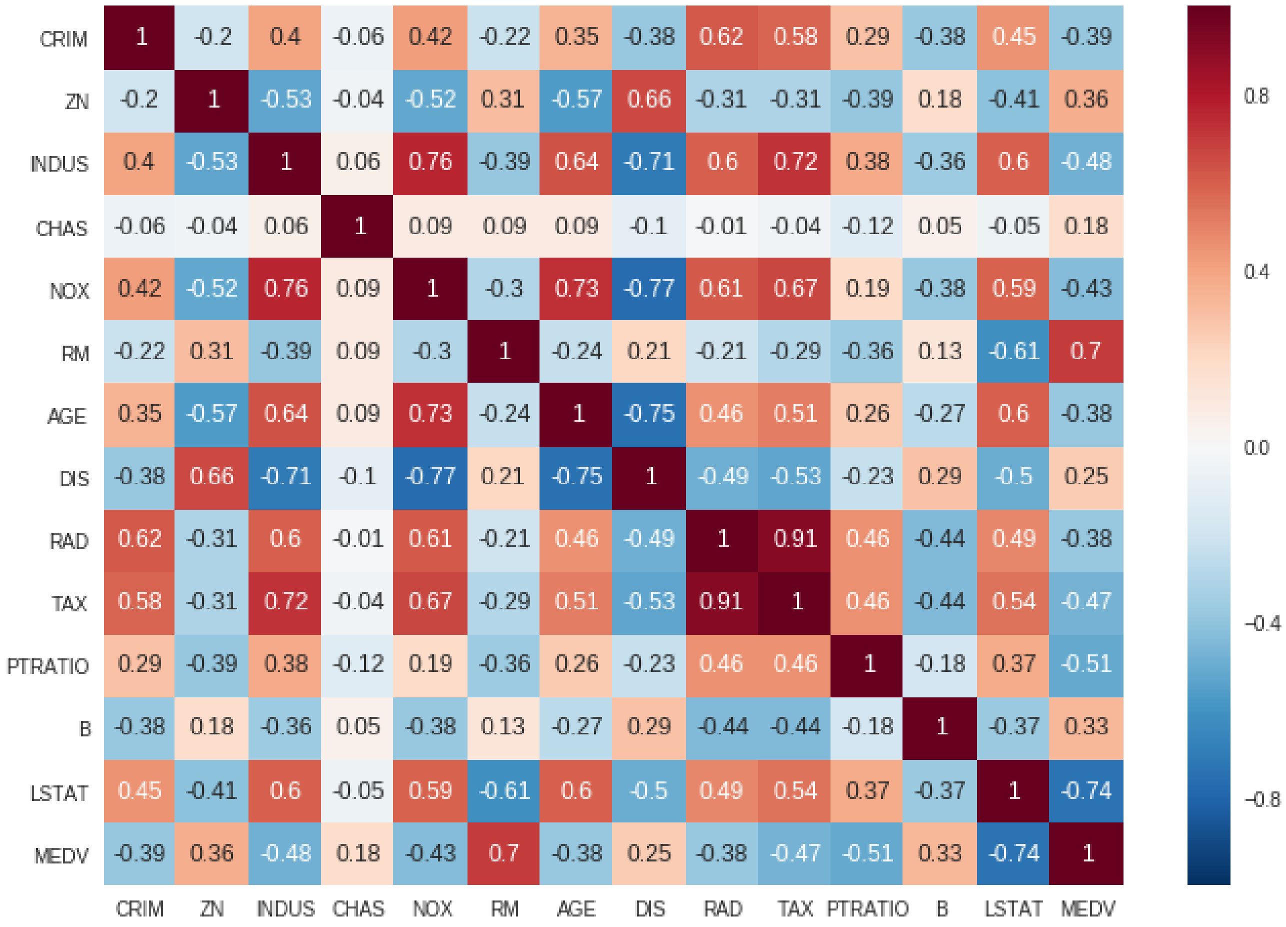
Model	Dataset	Run-time	Training MSE	Testing MSE
Normal Reg	Housing	0.01 seconds	38.71	38.41
Linear Reg	Housing	0.02 seconds	38.71	38.41
Linear Reg	Cali R.	0.004 seconds	61875.17	50737.021
Ridge Reg	Housing	0.01 seconds	38.71	38.40
Ridge Reg	Cali R.	0.004 seconds	61875.16	50737.017
Lasso Reg	Housing	0.008 seconds	38.730	38.251
Lasso Reg	Cali R.	0.005 seconds	61875.167	50736.90
Ransac Reg	Housing	0.02 seconds	50.121	41.774
Ransac Reg	Cali R.	0.0181 seconds	63074.088	55914.60
Poly + Linear Reg	Housing	0.012 seconds	29.442	32.511
Poly + Linear Reg	Cali R.	0.004 seconds	61714.538	46777.02

Table 1

Data Preprocessing

- I have added 30 days rows in order to form the "cali_renaw_cleaned" dataset. In the regression analysis of this dataset, I tried to find out whether it is possible to predict "Small Hydro" amount using the Biomass. Interestingly, I found out a high negative correlation between them.

- For the boston dataset, I used 'LSTAT' feature to predict 'MEDV'. The reason I chose this is because of the correlation of MEDV and LSTAT. It is highly correlated which can give us good data distribution for the regression analysis. Also, using only one feature, I can plot the regression line as well which gives a good visualization about the performance of the regressor. Correlation matrix can be found below.

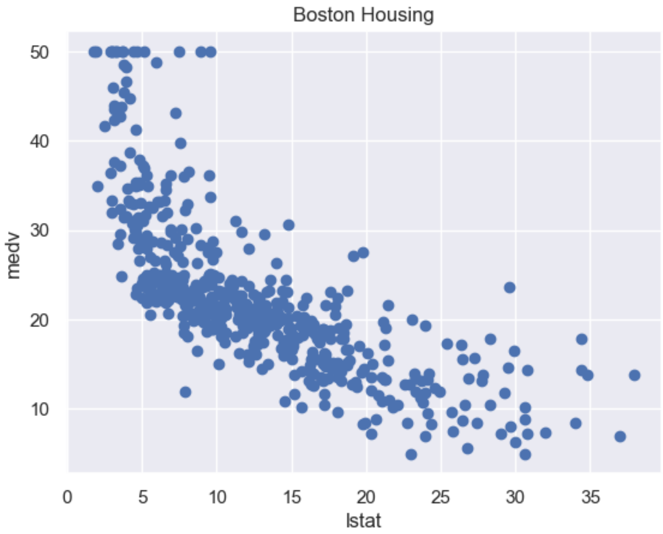


	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
CRIM	1	-0.2	0.4	-0.06	0.42	-0.22	0.35	-0.38	0.62	0.58	0.29	-0.38	0.45	-0.39
ZN	-0.2	1	-0.53	-0.04	-0.52	0.31	-0.57	0.66	-0.31	-0.31	-0.39	0.18	-0.41	0.36
INDUS	0.4	-0.53	1	0.06	0.76	-0.39	0.64	-0.71	0.6	0.72	0.38	-0.36	0.6	-0.48
CHAS	-0.06	-0.04	0.06	1	0.09	0.09	0.09	-0.1	-0.01	-0.04	-0.12	0.05	-0.05	0.18
NOX	0.42	-0.52	0.76	0.09	1	-0.3	0.73	-0.77	0.61	0.67	0.19	-0.38	0.59	-0.43
RM	-0.22	0.31	-0.39	0.09	-0.3	1	-0.24	0.21	-0.21	-0.29	-0.36	0.13	-0.61	0.7
AGE	0.35	-0.57	0.64	0.09	0.73	-0.24	1	-0.75	0.46	0.51	0.26	-0.27	0.6	-0.38
DIS	-0.38	0.66	-0.71	-0.1	-0.77	0.21	-0.75	1	-0.49	-0.53	-0.23	0.29	-0.5	0.25
RAD	0.62	-0.31	0.6	-0.01	0.61	-0.21	0.46	-0.49	1	0.91	0.46	-0.44	0.49	-0.38
TAX	0.58	-0.31	0.72	-0.04	0.67	-0.29	0.51	-0.53	0.91	1	0.46	-0.44	0.54	-0.47
PTRATIO	0.29	-0.39	0.38	-0.12	0.19	-0.36	0.26	-0.23	0.46	0.46	1	-0.18	0.37	-0.51
B	-0.38	0.18	-0.36	0.05	-0.38	0.13	-0.27	0.29	-0.44	-0.44	-0.18	1	-0.37	0.33
LSTAT	0.45	-0.41	0.6	-0.05	0.59	-0.61	0.6	-0.5	0.49	0.54	0.37	-0.37	1	-0.74
MEDV	-0.39	0.36	-0.48	0.18	-0.43	0.7	-0.38	0.25	-0.38	-0.47	-0.51	0.33	-0.74	1

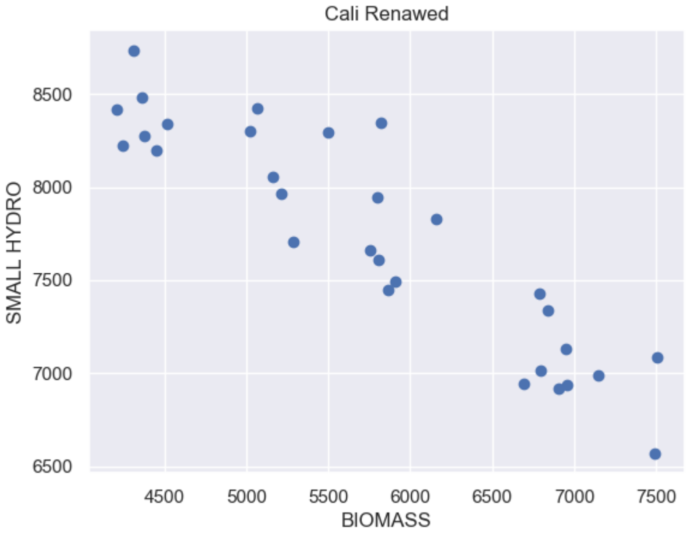
Distribution of Datasets

Before going into details of the regression analysis, I believe it is good idea to see distribution of the datasets.

Boston Housing Dataset

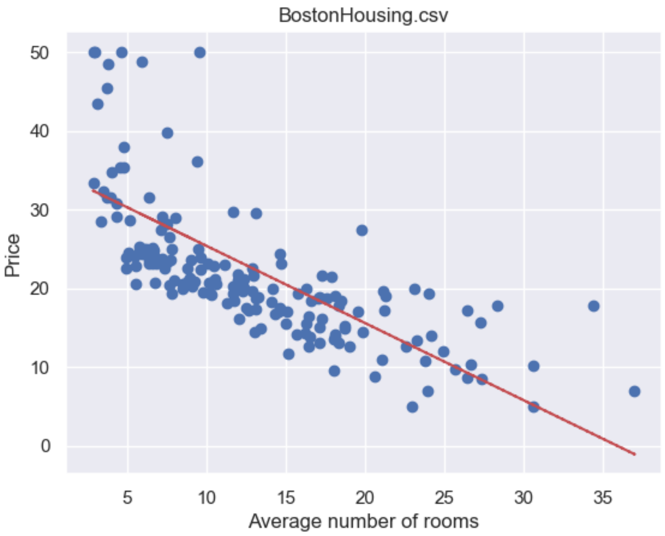


California Renewable Production

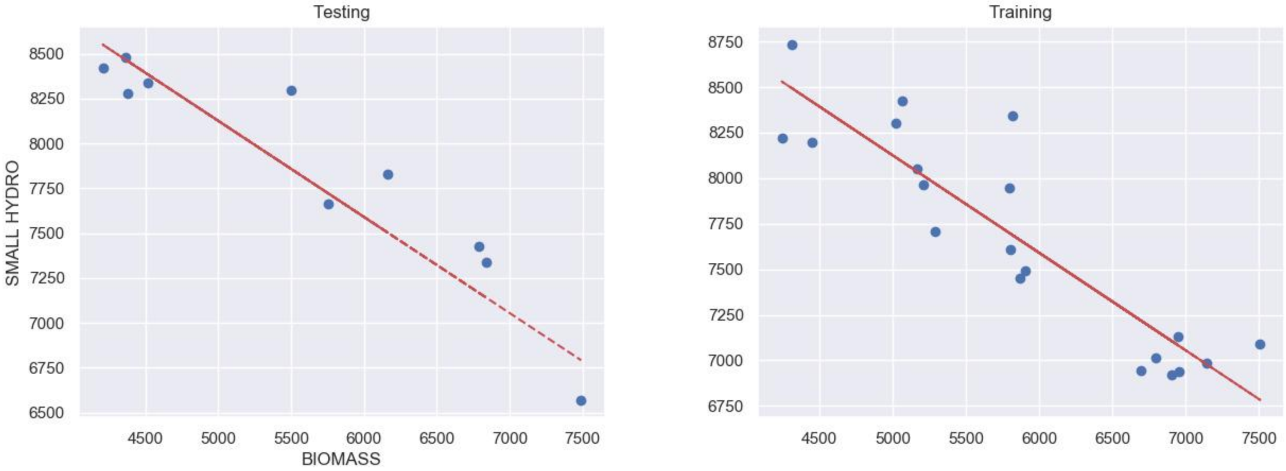


Normal Equation and Linear Regression Analysis

- In this homework, I only analyzed the normal equation on the housing dataset; however, the code is written in a way to handle different datasets as well.
- In my analysis, I found out the normal equation and linear equation are performing in same way. Both of them are fast and difference between the training and testing MSE are small. This indicated that the model fits well to the given dataset and **there isn't sign of any overfitting for both regressors.**
- However, when I train my LR on Cali. R. dataset, as in most of the regression analysis, training MSE is higher than the testing MSE. In other words, testing dataset does better job at fitting. Upon search online, I came up with the following reason why this is the case **(and this applies for all same regressors that performs better on testing then the training).**
- Following plot shows the regression line on the Normal Equation.



- Testing dataset easy to predict compare to training. Another reason is I need more dataset than what I have right now (I only have 30 rows). This situation also known as unknown fit.
    - Please note that data normalization didn’t fix this issue. I also tried that and there is still this situation takes place.
- Following two plots shows the regression line for both training and testing datasets.



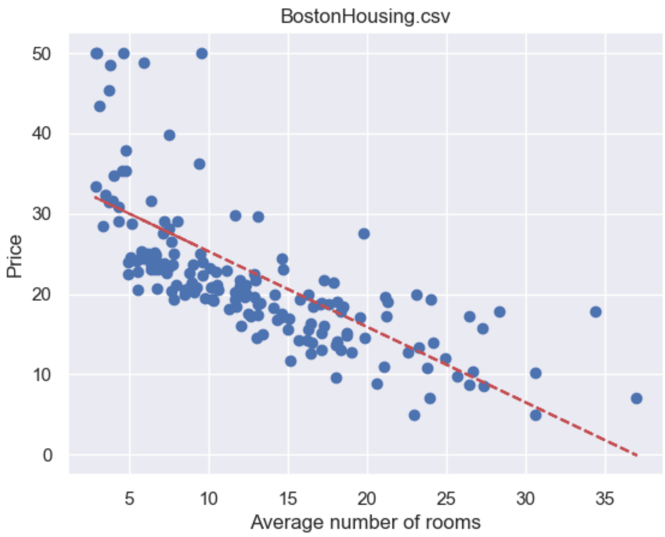
Ridge Regression analysis

- Ridge regression on the Boston Housing dataset did good job. Difference between the training and testing MSE’s are almost identical which **means it doesn’t overfit.**
- For the Cali. R. dataset, it performs as same as normal and linear regression.
- I changed the value of alpha from 2 to 3 and 4; however, I didn’t get any improvement on both datasets. I used ‘***svd*** and ‘***cholesky’*** on both datasets; however, none of the MSE scores changed. The run time little increased on ‘***cholesky’***.

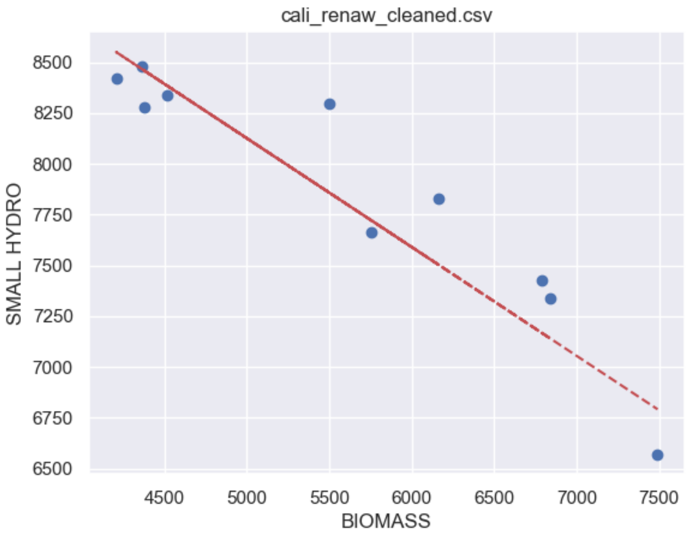
Lasso Regression analysis

- **For Boston Dataset**
    - So far, lasso regression did the worst job in terms of time it takes to fit. However, MSE scores for both training and testing datasets are almost identical as other regressors. **There is no overfitting issue.**
    - Changing alpha from 2 to 4 and 5 reduced the MSE by 0.01. I believe it is still improvement.

- Following figure shows the regressor line for the testing dataset on Housing dataset.

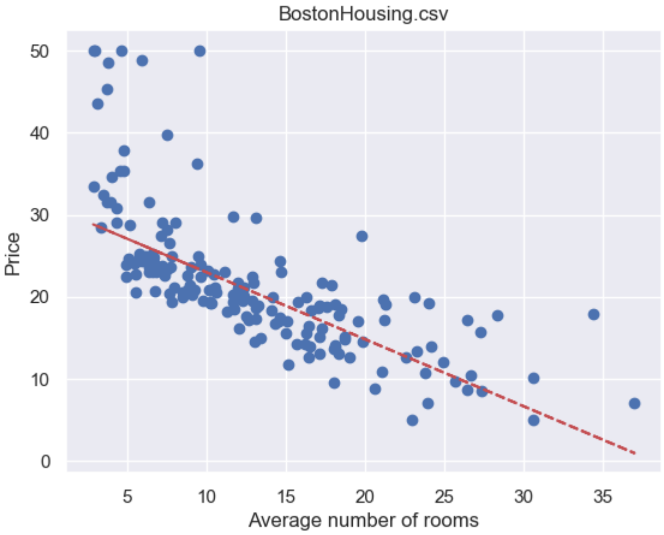


- **For Cali. R. Dataset**
  - Changing the degree for the Lasso Regression did 0.01 improvement on the testing dataset; however, testing dataset did better job compare to training dataset.
  - Following figure shows the regressor line for testing dataset.

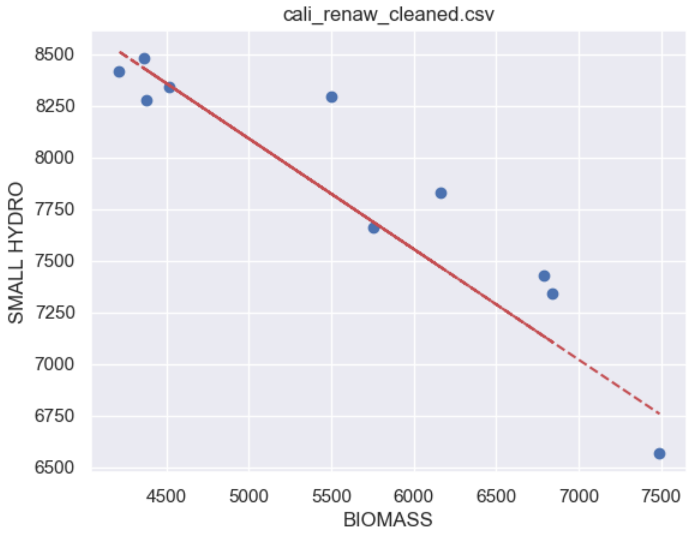


Ransac Regression Analysis

- **For Boston Dataset**
  - So far, all the regressors except polynomial featured linear regression, resulted same MSE for both training and testing datasets. However, this regressor resulted worse compare to others. Upon checking online resources, I came with the following reason.
    - Dataset distribution is not suitable for the RANSAC analysis. Yet we will see in the next paragraph polynomial featured LR does the better job. It still does unknown fitting (Lower MSE on testing dataset).
    - **No sign of overfitting.**
    - Here the regressor plot for the Ransac Regression on the Boston Dataset.



- **For Cali. R. Dataset**
  - Again, it does the worst job in terms of MSE scoring compare to other regressors.
  - **I don't see any overfitting issue.**
  - The below plot shows the regressor on the testing dataset.



Polynomial Featured Linear Regression

For Boston Dataset

- So far, this is the first time I encountered a hint for the **overfitting problem.** Testing MSE is higher compare to training MSE score. Although there is a small overfitting issue (might not be overfitting issue), it does better job compare to all regressors.
    - **Different Degree**
        - I got the above results with the default degree, 2.
        - **Changing this degree to 3 gave me lower MSE score**. (30.07 MSE on testing and 28.1 MSE on training)
        - **I received best tradeoff at degree 6 with 26.03 MSE on training and 28.10 MSE on testing dataset.)**

For Cali. R. Dataset

- As it is clear from the table 1, it does the worst job in terms of MSE on this dataset compare to all regressors which was quite expected as dataset distribution was linear.
- Increasing degree worsened the MSE again as expected.