Ruozhong Yang (ry8)

IE598 MLF F18

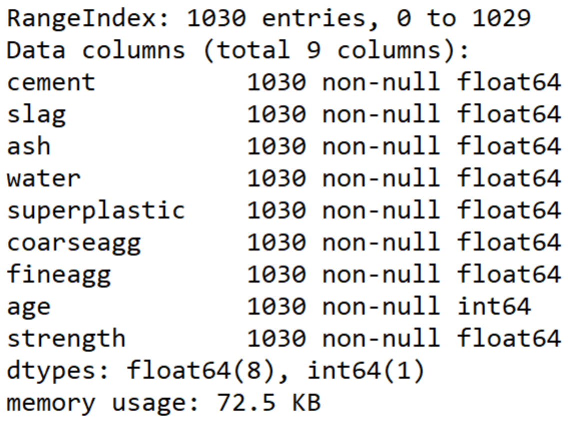
Module 4 Homework (Regression)

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.

The first step is to get the basic information of this data set and we an get the shape and information via codes:

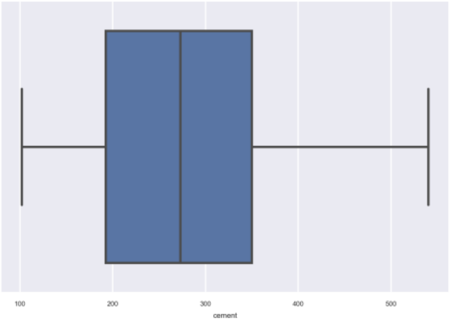


```
RangeIndex: 1030 entries, 0 to 1029
Data columns (total 9 columns):
cement          1030 non-null float64
slag            1030 non-null float64
ash             1030 non-null float64
water           1030 non-null float64
superplastic    1030 non-null float64
coarseagg       1030 non-null float64
fineagg         1030 non-null float64
age             1030 non-null int64
strength        1030 non-null float64
dtypes: float64(8), int64(1)
memory usage: 72.5 KB
```

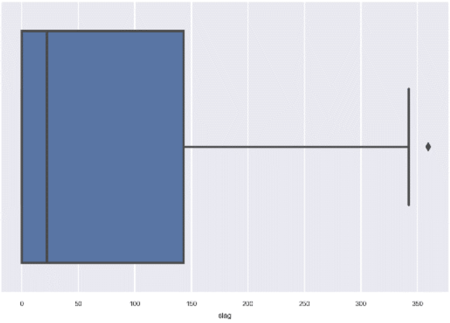
1.1 Basic information

Then, we can use the box plot to read to approximate median, range of the nine features. In this report, we will use strength as the feature we want to predict and the rest eight to predict it via different types of model in SKlearn. The box plot of each feature looks like this:

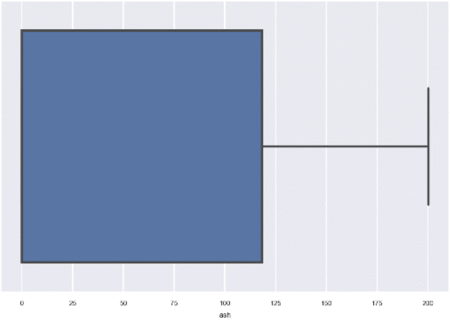
cement



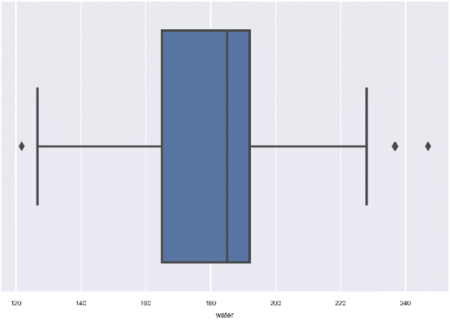
slag



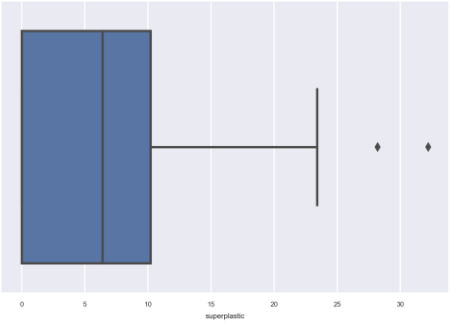
ash



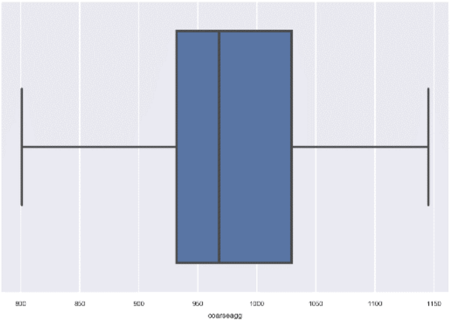
water



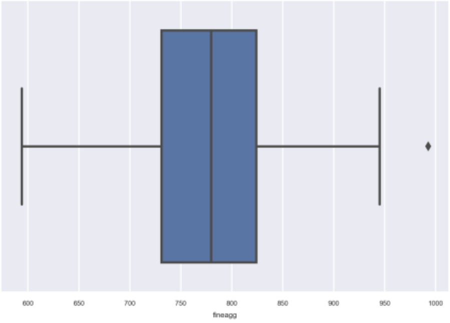
superplastic



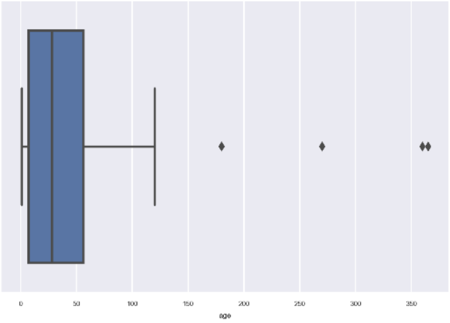
coarseagg

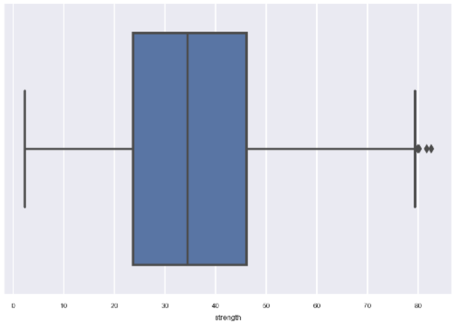


fineagg

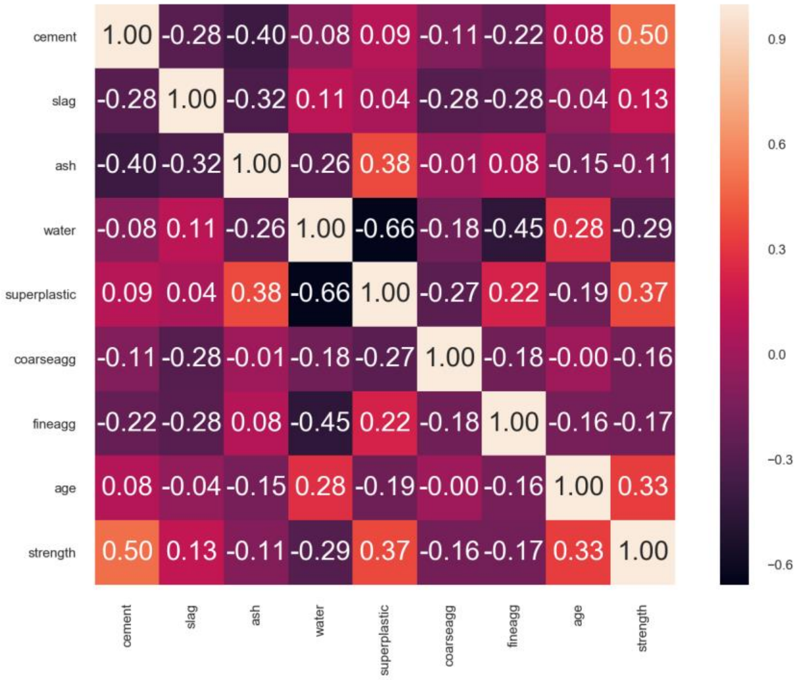


age

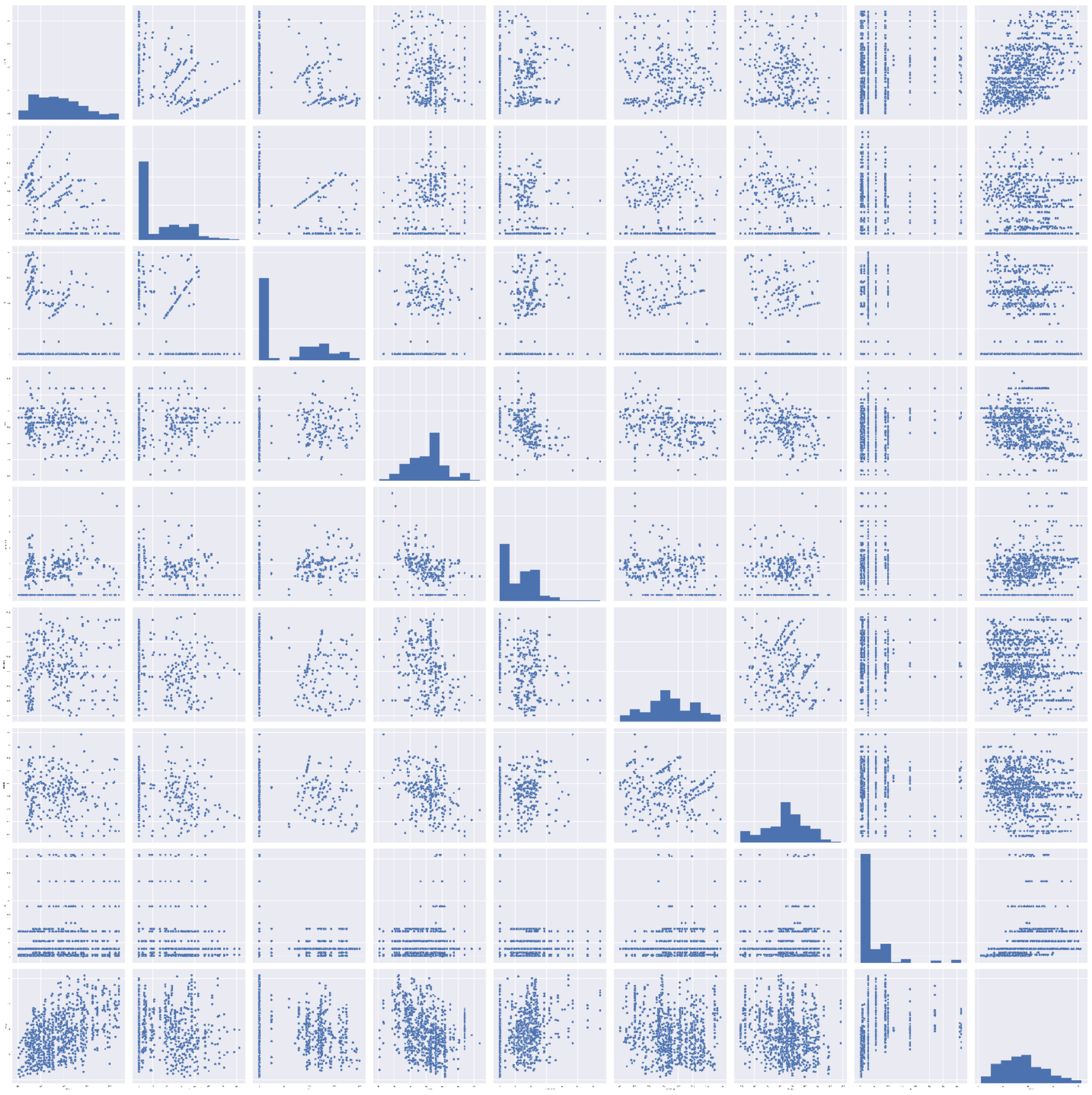


strength


We also print the correlation matrix and scatterplot matrix of this data set. The scatterplot matrix is such a huge matrix that I have to submit it in another file to get a clear view.



1.2 Correlation matrix

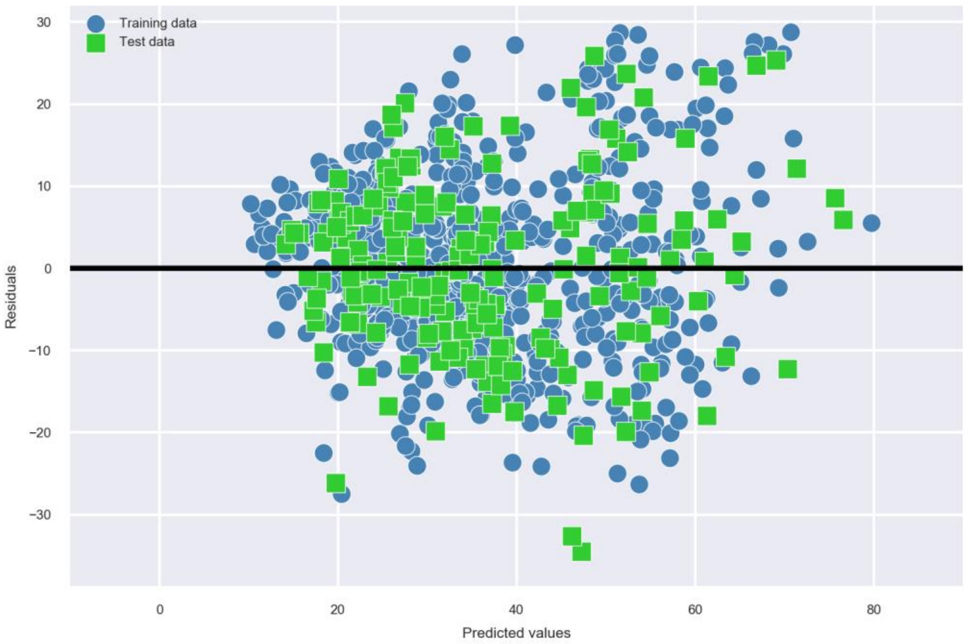


1.3 Scatterplot matrix(all)

Part 2: Linear regression

In this part ,we fit a linear model using SKlearn to all of the features of the dataset. After fitting ,we can describe the model (coefficients and y intercept), plot the residual errors, calculate performance metrics: MSE and R2.

No.	MSE train	MSE test	R^2 train	R^2 test	Slope	Intercept
LR	106.025	112.134	0.617	0.608	0.123	-30.959

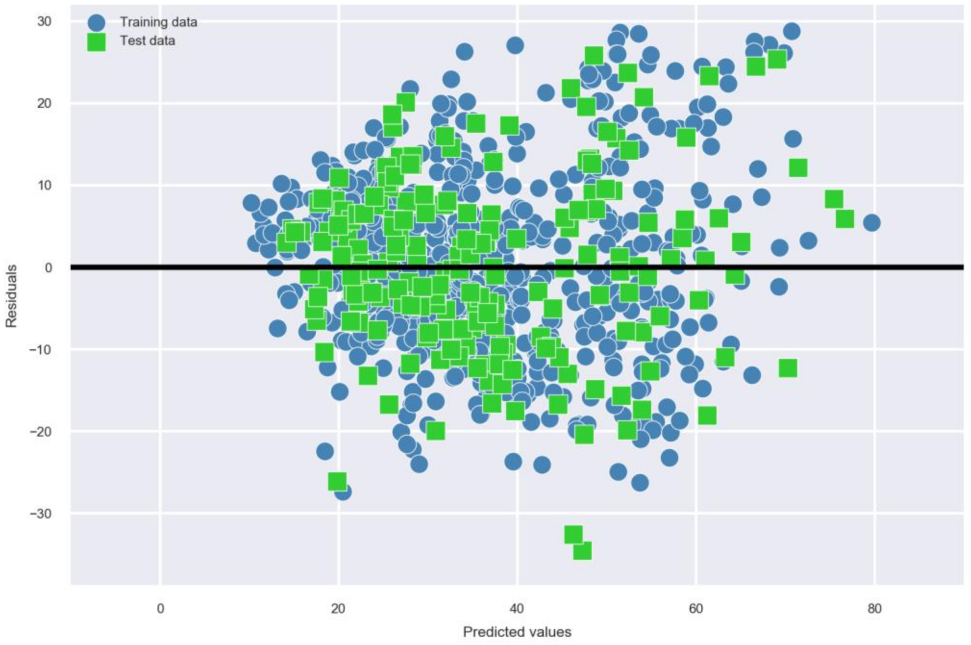
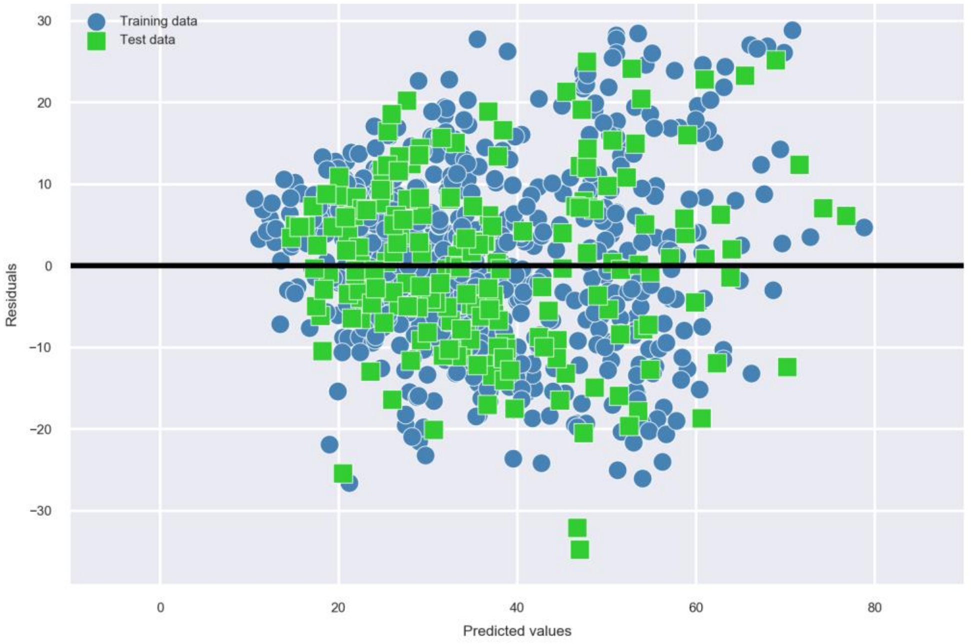


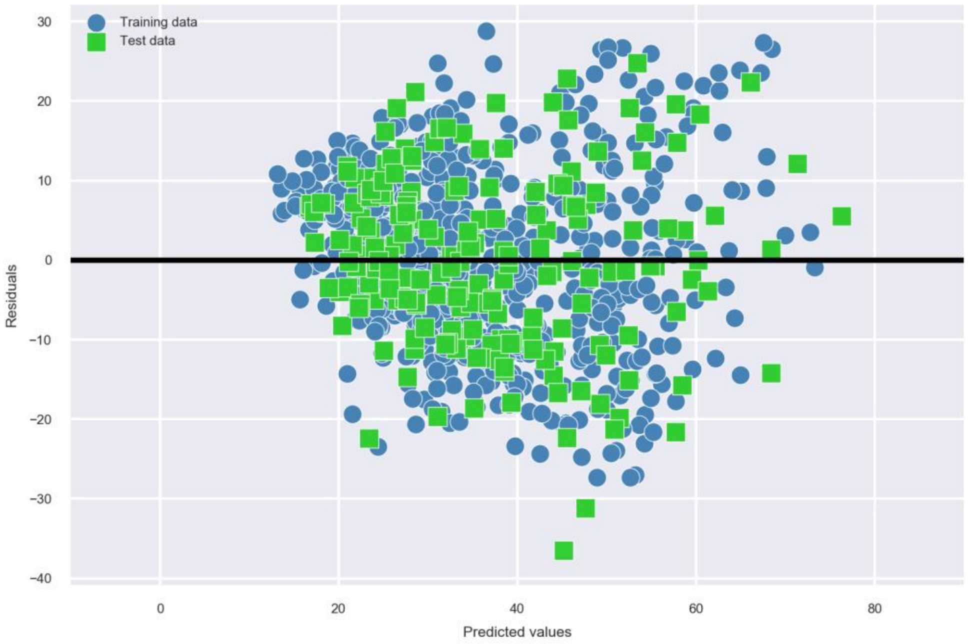
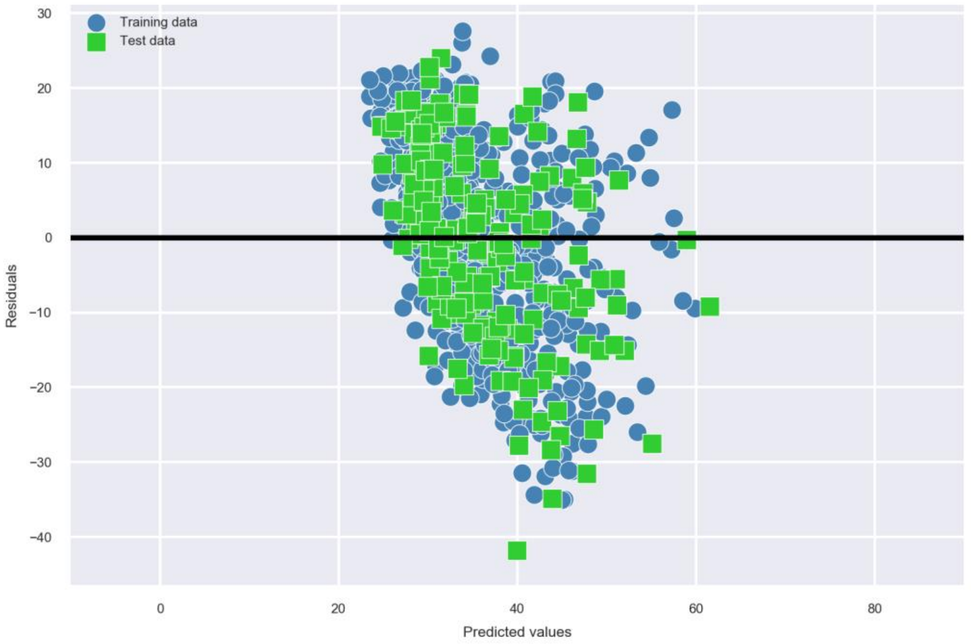
1.4 Residual errors(LR)

Part 3.1: Ridge regression

By fitting a Ridge model via SKlearn to all of the features of the dataset, we can test several settings for alpha. According to the MSE and R^2 , model Ridge(alpha=1e-02) gives a good description.

No.	alpha	MSE train	MSE test	R^2 train	R^2 test	Slope	Intercept
1	1e-03	106.030	112.044	0.617	0.608	0.122	-25.467
2	1e-02	106.364	111.738	0.617	0.609	0.109	10.437
3	1e-01	111.249	114.225	0.598	0.601	0.073	79.351
4	1	153.926	154.898	0.444	0.459	0.036	67.449

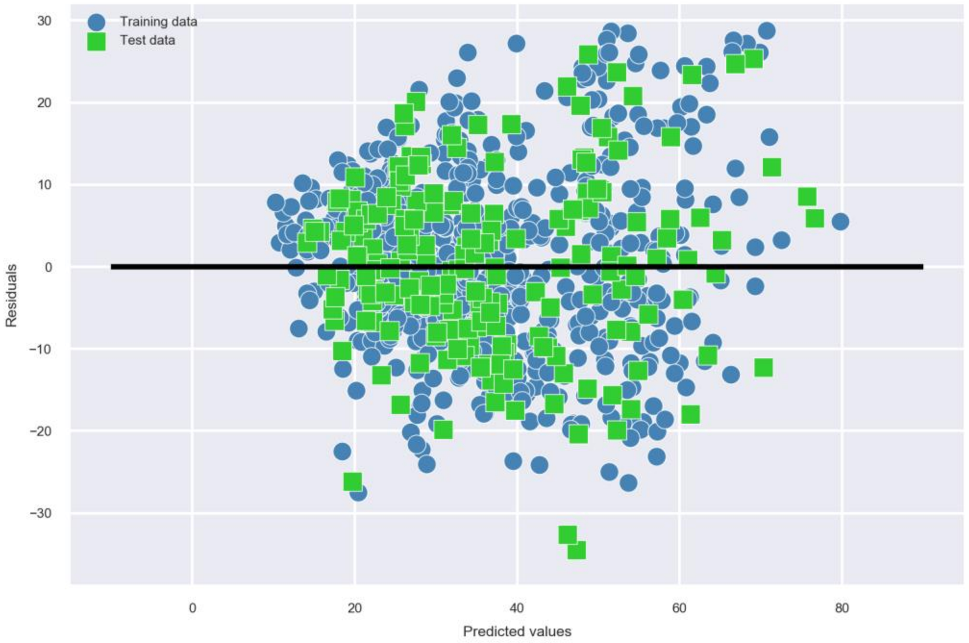
alpha =1e-03	
alpha =1e-02	

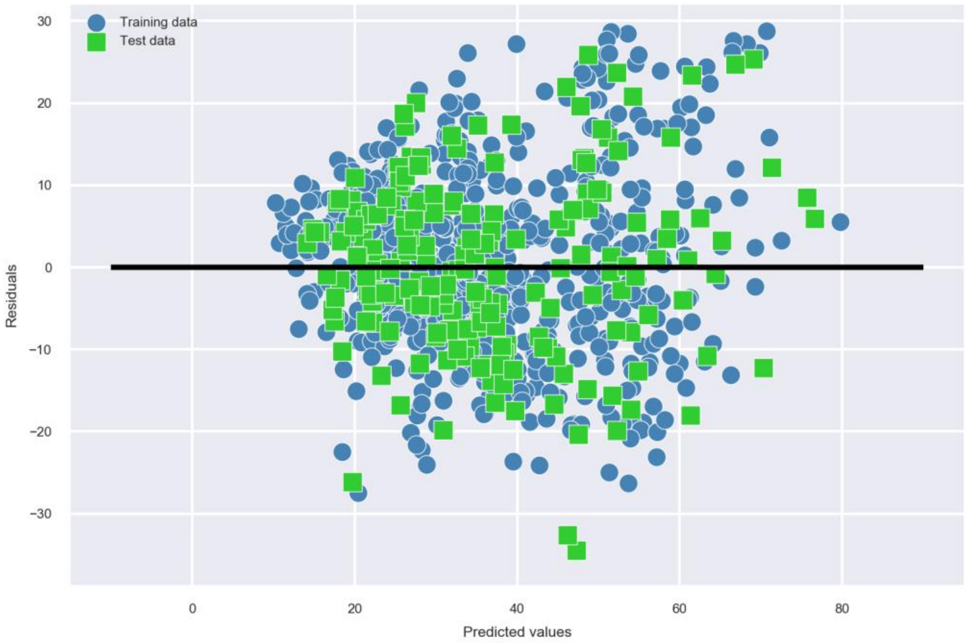
alpha =1e-01	
alpha =1	

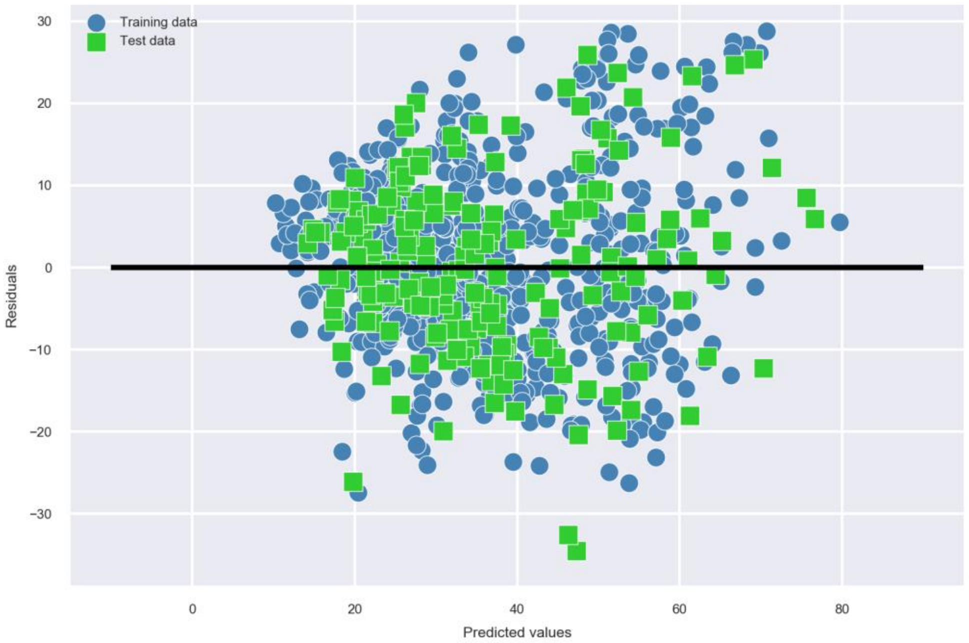
Part 3.2: LASSO regression

By fitting a LASSO model via SKlearn to all of the features of the dataset, we can test several settings for alpha. According to the MSE and R^2 , all those models seem to give a similar description.

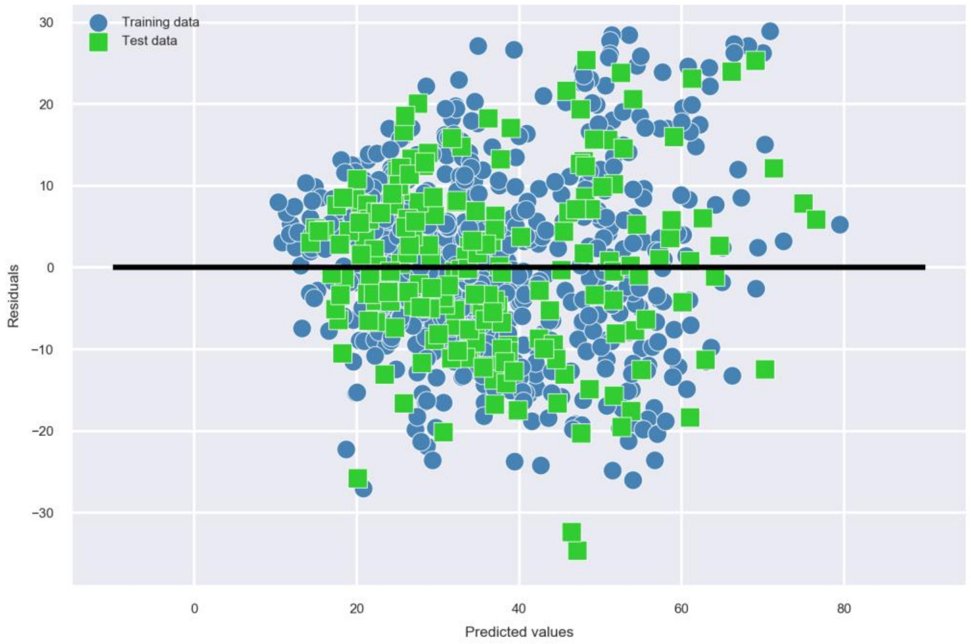
No.	alpha	MSE train	MSE test	R^2 train	R^2 test	Slope	Intercept
1	1e-06	106.025	112.134	0.617	0.608	0.123	-30.935
2	1e-05	106.025	112.131	0.617	0.608	0.123	-30.720
3	1e-04	106.026	112.100	0.617	0.608	0.123	-28.574
4	1e-03	106.116	111.888	0.617	0.609	0.116	-7.195

alpha =1e-06	

<table>
<tr>
<td>alpha =1e-05</td>
<td>


</td>
</tr>
<tr>
<td>alpha =1e-04</td>
<td>


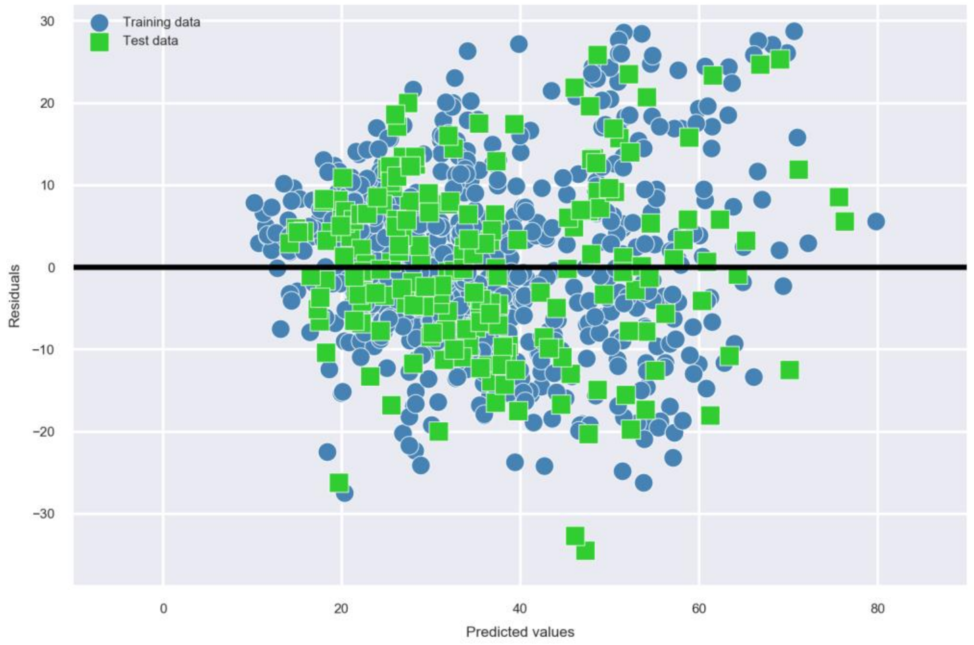
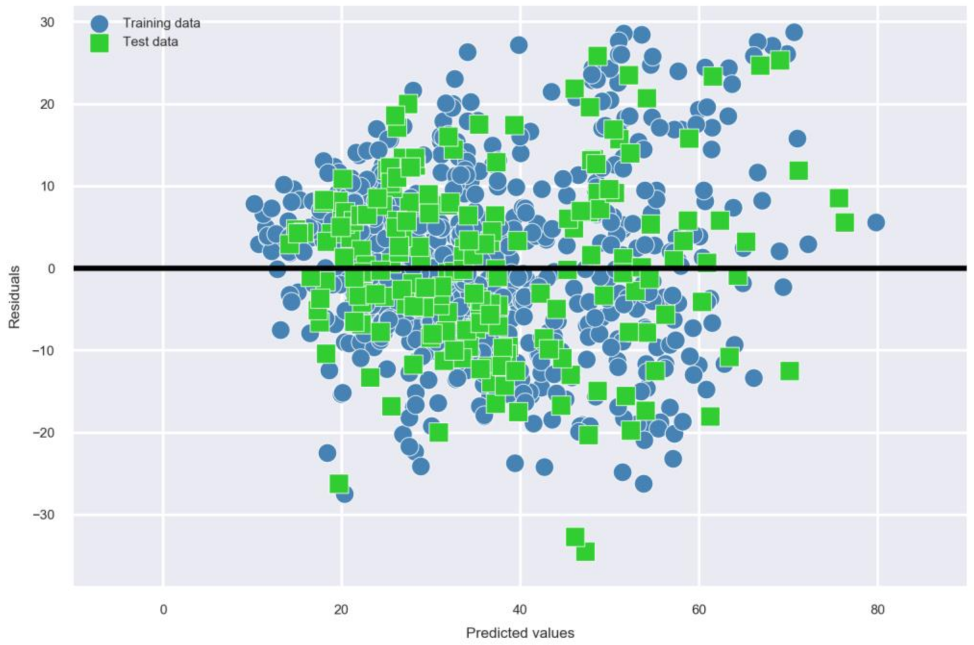
</td>
</tr>
</table>

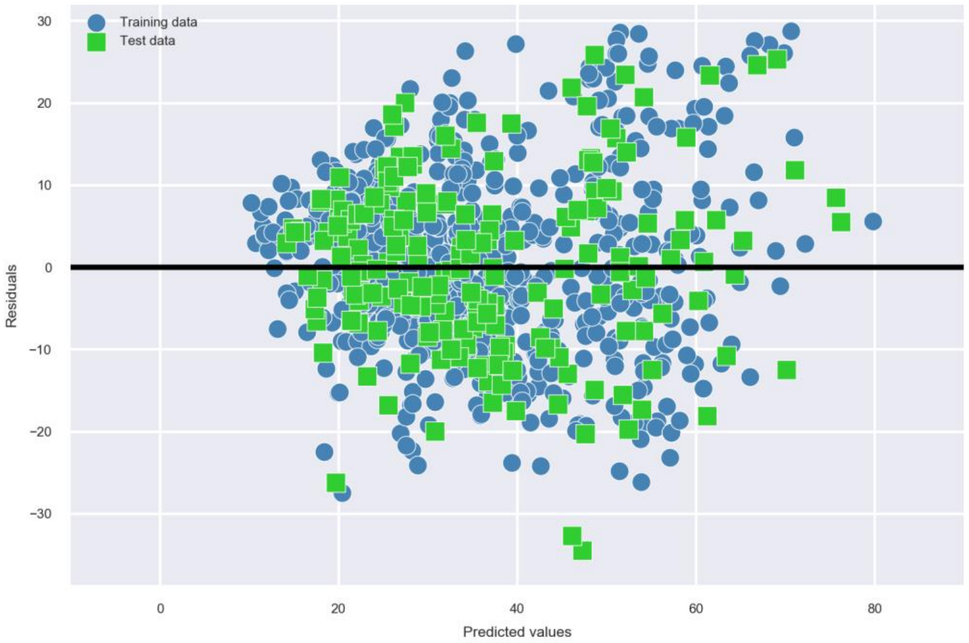
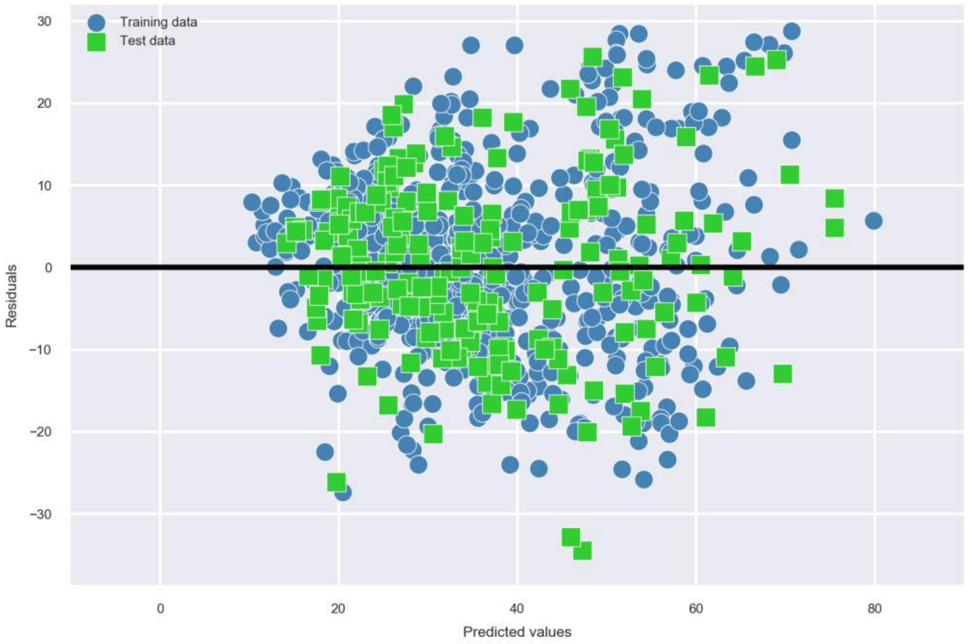
alpha =1e-03	

Part 3.3: Elastic Net regression

y fitting a LASSO model via SKlearn to all of the features of the dataset, we can test several settings for l1_ratio(alpha=1). According to the MSE and R^2 , all those models seem to give a similar description.

No.	l1_ratio=	MSE train	MSE test	R^2 train	R^2 test	Slope	Intercept
1	1e-03	106.029	112.155	0.617	0.608	0.123	-29.202
2	1e-02	106.030	112.154	0.617	0.608	0.123	-29.125
3	1e-01	106.032	112.153	0.617	0.608	0.123	-28.344
4	1	106.088	122.185	0.617	0.608	0.122	-20.045

l1_ratio=1e-03	
l1_ratio=1e-02	

l1_ratio=1e-01	
l1_ratio=1	

Part 4: Conclusions

Write a short paragraph summarizing your findings.

According to whole analysis we can see the most related feature of strength is cement, among all those models, the most effective model should be the model Ridge(alpha=1e-02).

Part 5: Appendix

https://github.com/yrz437396236/IE598_F18_HW1/tree/master/IE598_F18-HW4