Lai Chung Sing (1155031274)
CSCI3320 Assignment 1

3 Linear Regression in Scikit-learn

3.1 Explore Scikit-learn Dataset

3.1.1 Get n_features and n_samples

Number of features in the Boston dataset is: 13
Number of samples in the Boston dataset is: 506

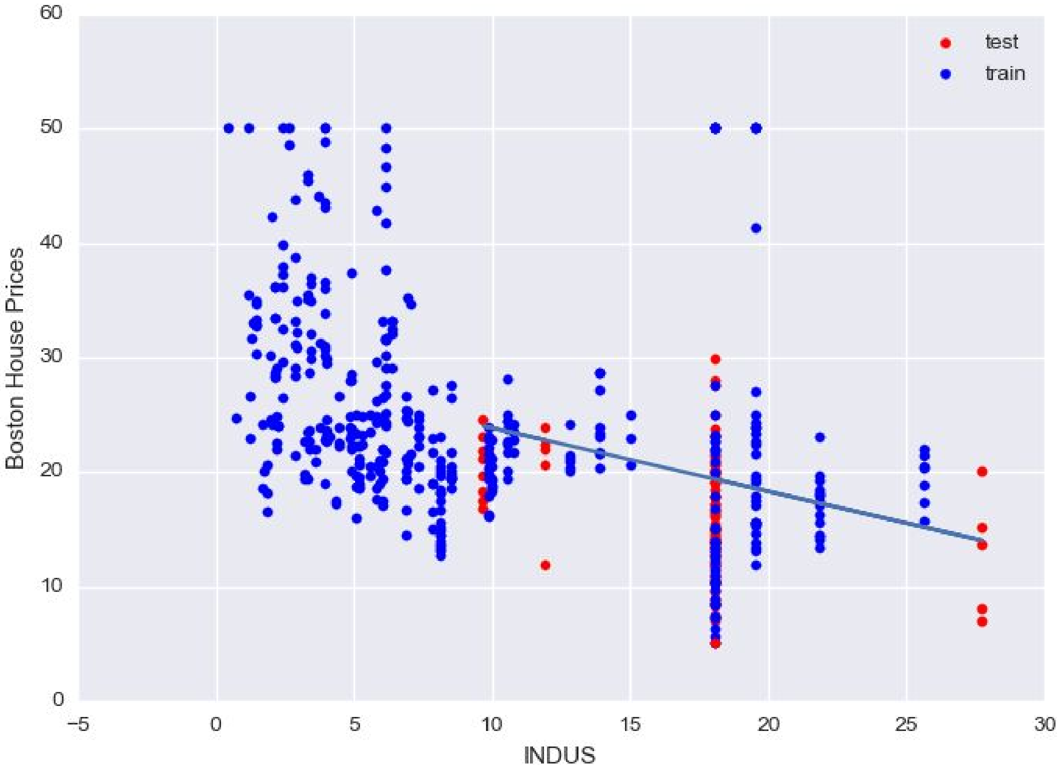
3.1.2 Find best fitted feature

Best fitted feature name is: INDUS
Best fitted model score is: 0.20596851298

3.1.3 Calculate the loss function

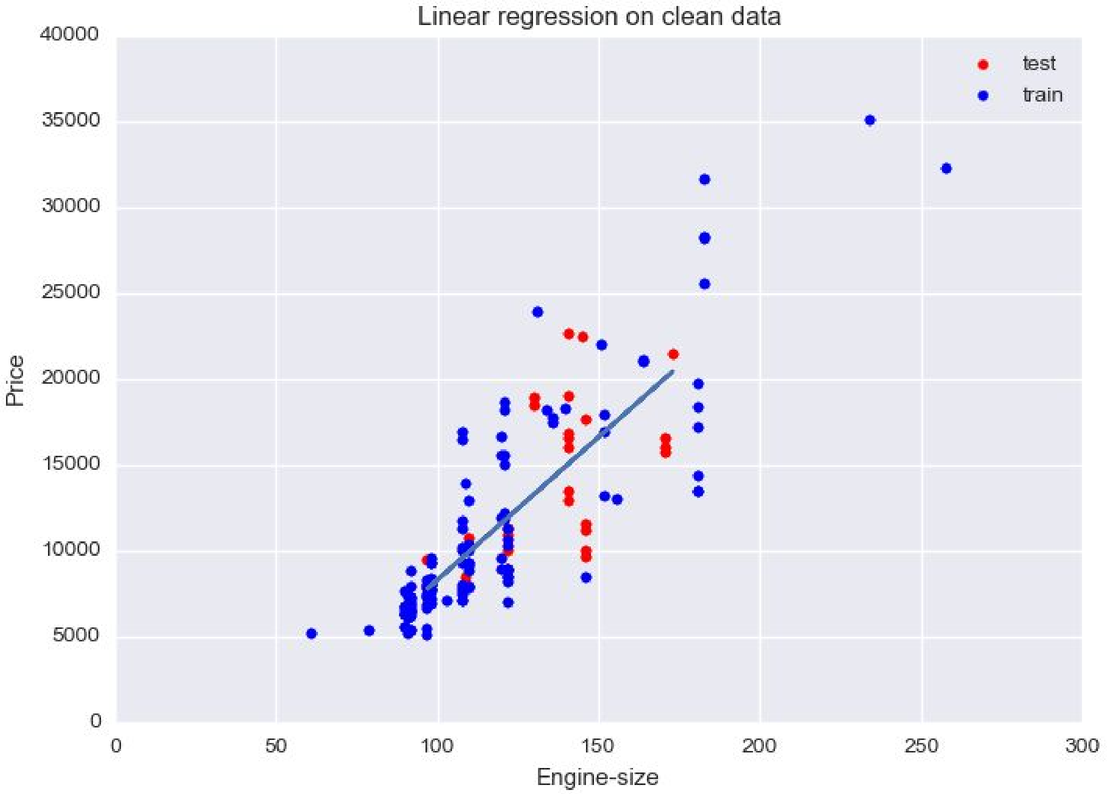
Value of the loss function for the best fitted model is: 18.5645355697

3.1.4 Plot the predictions and test data



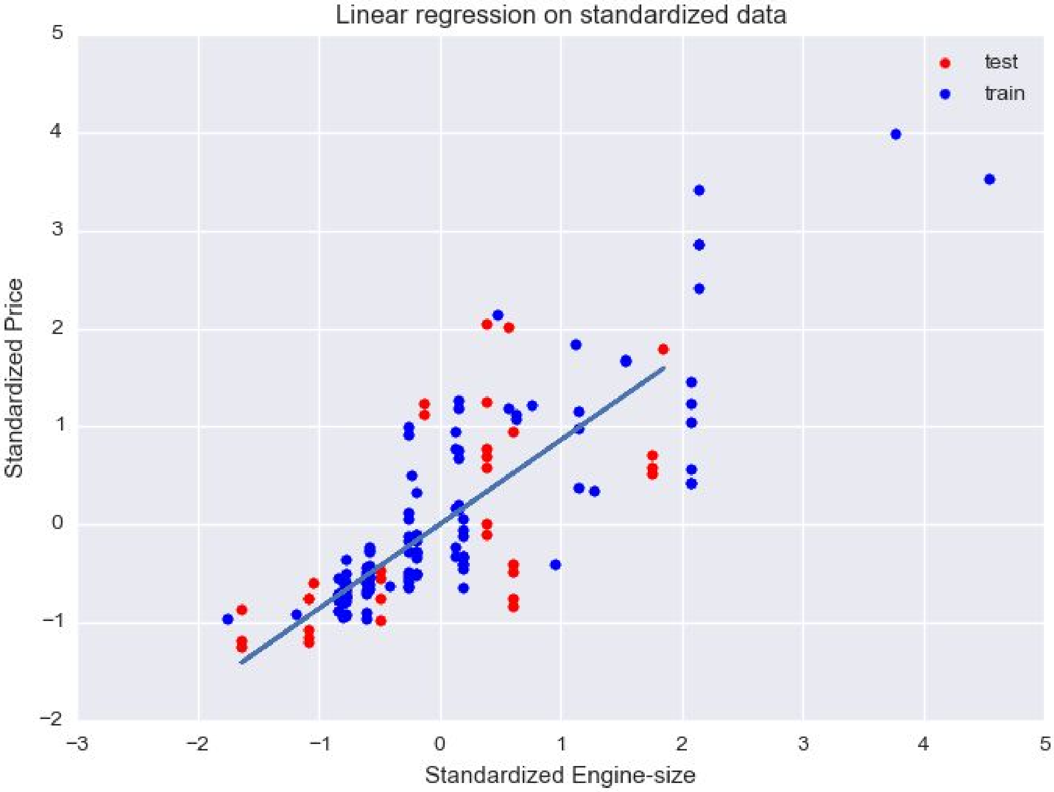
3.2 Explore Raw Dataset

3.2.3 Linear regression on the cleaned data



Price prediction for engine size equals to 175 is: 20793.532819

3.2.4 Linear regression on the standardized data



3.2.5 Linear regression with multiple features

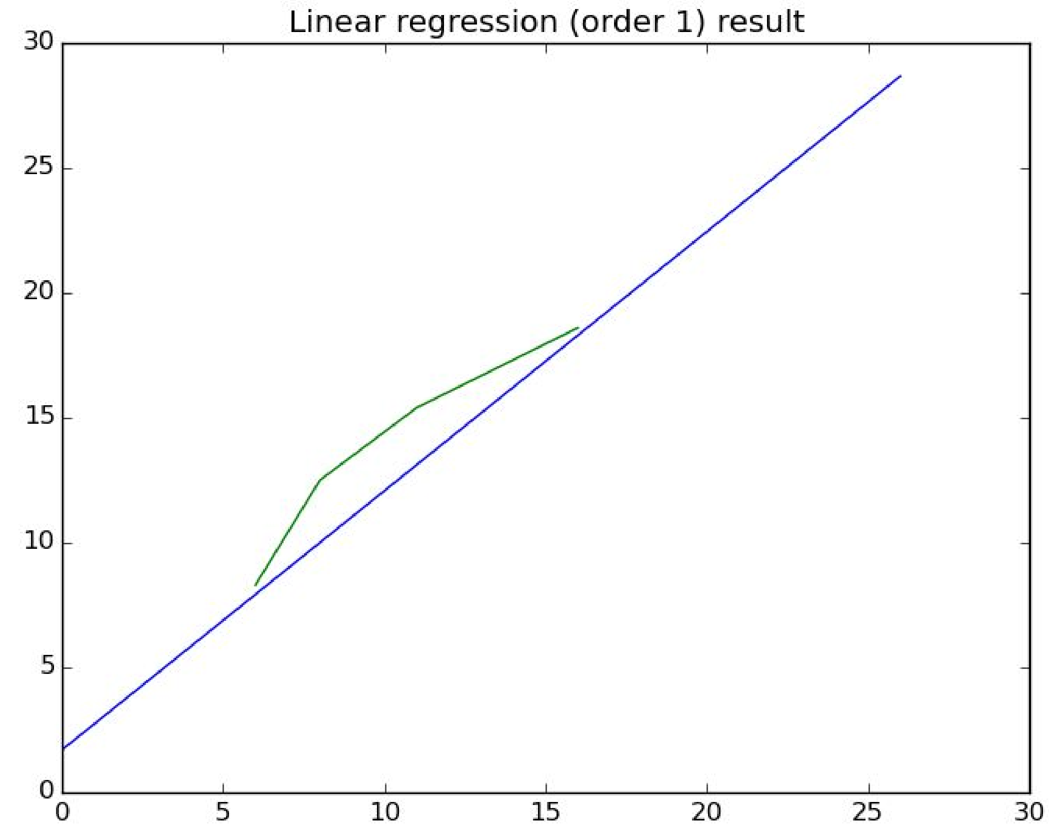
```
Parameter theta calculated by normal equation:
[[ -4.51028104e-17]
 [  8.62451816e-01]
 [  7.36117228e-02]]
Parameter theta calculated by SGD:  [-0.00054594] ,  [ 0.72440641 -0.00671158]
```

3.3 Understand Regularization

3.3.1 LR regression on polynomial data

y1 = 1.7 + 1.0375x

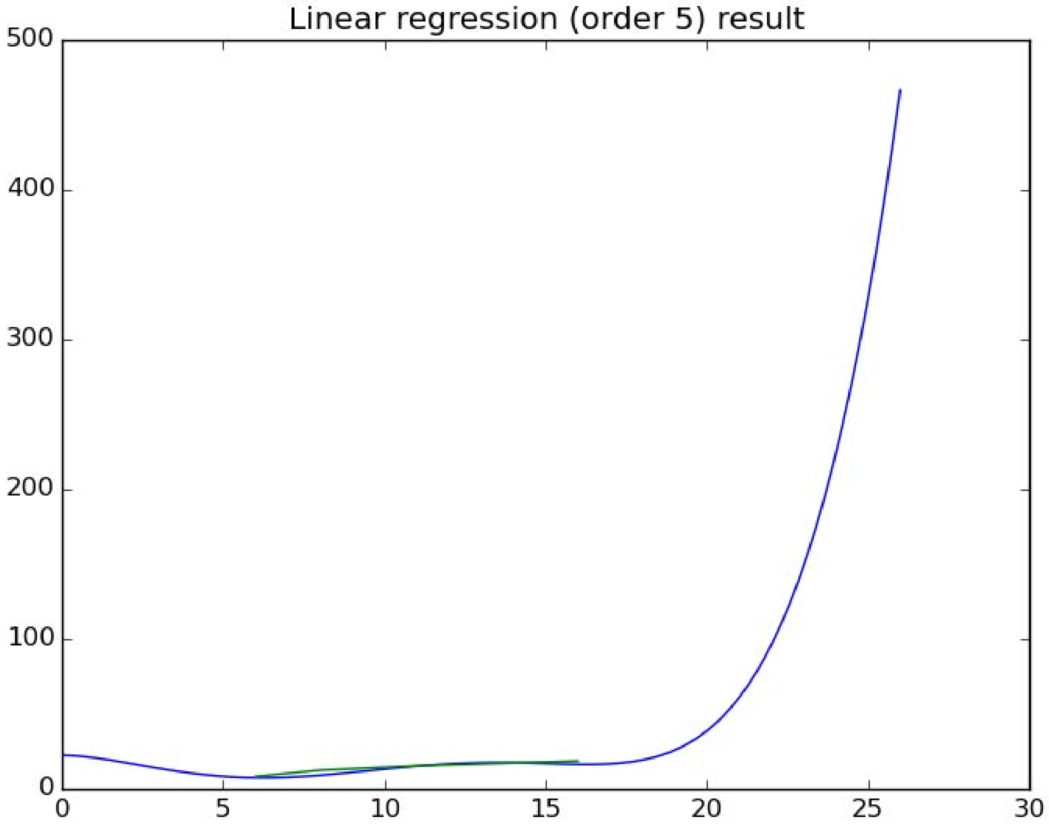
Linear regression (order 1) model score is: 0.79629076087



3.3.2 Polynomial regression on training data

y2 = (22.51893655) +
(2.34036485e-11)x +
(-3.45575425e-01) x*x +
(-1.69531630e+00)x*x*x +
(3.41630495e-01)x*x*x*x +
(-2.28934322e-02) x*x*x*x*x +
(5.09722747e-04) x*x*x*x*x*x

Linear regression (order 5) score is: 0.706486295693



3.3.3 Ridge Regression

y3 = (8.04794217) +
(0.00000000e+00)x +
(-7.08628224e-02) x*x +
(-3.45903785e-01)x*x*x +
(8.79869408e-02)x*x*x*x +
(-6.15336314e-03) x*x*x*x*x +
(1.36225166e-04) x*x*x*x*x*x
Ridge regression (order 5) score is: 0.82053974087

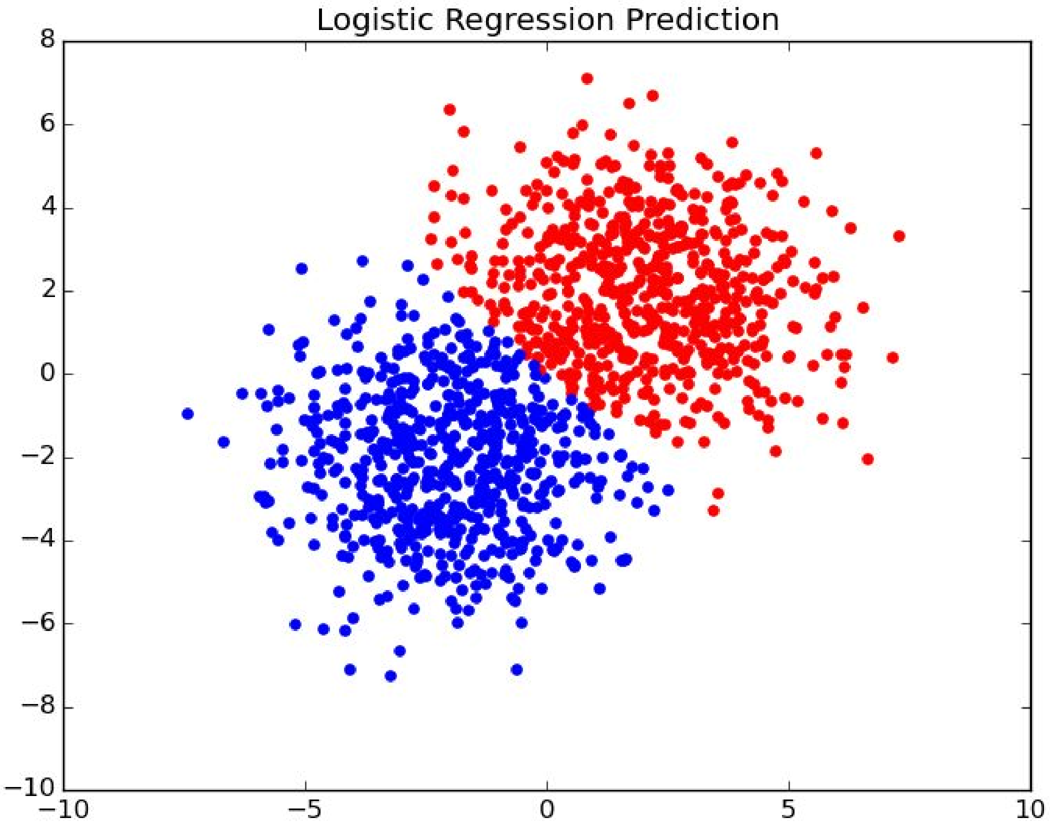
3.3.4 Comparisons

The model with the highest score is: Ridge Regression
Ridge model can prevent over-fitting: Yes
Ridge model is nearly equivalent to LR model (order 5) if alpha = 0: Yes
A larger alpha results in a larger coefficient for x*x*x*x*x: Yes

4 Linear Discrimination/Classification

4.1 Binary Classification

The predictions only have 0 and 1: Yes



4.2 Classification Statistics

Number of wrong predictions is: 73