Problem 2©

Linear Regression Coefficients:

index ▲	Feature	Coeff
0	CRIM	-0.0993237493
1	ZN	0.0522513375
2	INDUS	0.004515606
3	CHAS	2.9572610163
4	NOX	1.1279377483
5	RM	5.8541984993
6	AGE	-0.01495686
7	DIS	-0.9208437549
8	RAD	0.1595191043
9	TAX	-0.0089342717
10	PTRATIO	-0.4356744352
11	B	0.0149052352
12	LSTAT	-0.4747505148

Ridge Regression Coefficients when eta = 15:

Filter Rows

index ▲	Feature	Coeff
0	CRIM	-0.1006476846
1	ZN	0.0546323927
2	INDUS	0.0129580836
3	CHAS	2.2727834496
4	NOX	0.4576743373
5	RM	5.7281521134
6	AGE	-0.0100943724
7	DIS	-0.8969852778
8	RAD	0.1630844674
9	TAX	-0.0089823137
10	PTRATIO	-0.4061490574
11	B	0.0155177873
12	LSTAT	-0.484273584

Ridge Regression Coefficients when eta = 0.02:

index ▲	Feature	Coeff
0	CRIM	-0.0993282363
1	ZN	0.0522540587
2	INDUS	0.0045516127
3	CHAS	2.9561707142
4	NOX	1.1226184046
5	RM	5.8542341864
6	AGE	-0.0149461895
7	DIS	-0.9208461031
8	RAD	0.1595292071
9	TAX	-0.0089338279
10	PTRATIO	-0.4356325653
11	B	0.0149065509
12	LSTAT	-0.4747422626

Ridge regression puts constraint on the coefficients w. The penalty term (eta/2= lambda) regularizes the coefficients such that if the coefficients take large values the optimization function is penalized. Lower the constraint (low eta) on the features, the model will resemble linear regression model (for example when (eta/2 = 0.01). For higher value of eta (15), the magnitudes are considerably less compared to linear regression case.

Problem 2(D):

After implementing the prediction function and root mean square error, I got following results:

The RMSE of linear regression on train set (after concatenating a constant value (1) to each feature vector to learn some linear offset): 4.63142853427834
The RMSE of linear regression on test set (after concatenating a constant value (1) to each feature vector to learn some linear offset): 4.8699261725702

The RMSE of Ridge regression on train set (after concatenating a constant value (1) to each feature vector to learn some linear offset): 4.795434059479303 [When eta = 15 (eta= 15)]

The RMSE of Ridge regression on test set (after concatenating a constant value (1) to each feature vector to learn some linear offset): 5.1603378230671995 [When eta = 15 (eta = 15)]

Discussion:

Ridge regression puts constraint on the coefficients w. The penalty term eta regularizes the coefficients such that if the coefficients take large values the optimization function is penalized. When eta= 15, in the training dataset, eta puts more restriction on the coefficients by shrinking their magnitude that's why RMSE error is greater (4.80) in Ridge regression than linear regression (4.63). It is evident by noticing the RMSE value of test sets in both cases that Ridge regression underfits the Boston housing price model with a 5.16 RMSE where in linear regression the RMSE is 4.87.

Lower the constraint (low eta) on the features, the model will resemble linear regression model (for example when eta = 0.02).

Problem 2(e):

The RMSE of linear regression on train set with top three features: 5.273361751695365
The RMSE of linear regression on test set with top three features: 5.494723646664577

The RMSE of ridge regression on train set with top three features: 5.275045693942413
The RMSE of ridge regression on test set with top three features: 5.481154712581162

We can see that this time in both cases RMSE error is greater (with only top 3 features) than before (trained with all its feature). The features we deleted at least some of them played critical roles in the classification model. Moreover, if we observe the heat matrix, we can see there are some collinearity between RM and LSTAT (-0.6, negative collinearity) which also influenced the accuracy of the model.

Problem 2(f)

Feature Engineering:

I was able to reduce RMSE from 4.87 (2(d)) to 4.77

The techniques I used for feature engineering are following:

1. Zn and INDS are very closely related to each other. So, I added these two features and made them one feature
2. PTRATIO, TAX and B - made these features linear by using a logarithmic function

3. Took the variance of age and add this as a feature and at this point the correlation with MEDV increased then the previous feature AGE

4. Took square root of CRIM, NOX and RAD.