Stats 151A: Midterm #2
Predicting the Value of a Boston Home Using Neighborhood Features

I. Introduction

The purpose of this report is to analyze housing data from various census tracts of Boston from the 1970 census. Harrison and Rubinfeld collected the data in 1979 for the purpose of discovering whether or not clean air influenced the value of houses in Boston, Massachusetts. Their paper was titled "Hedonic prices and the demand for clean air" and it can be found in the *Journal of Environmental Economics and Management*, 5, 81–102.

This report attempts to answer the following two questions: which variables mainly determine the median housing price in a tract? How are the prices of houses affected by these different neighborhood characteristics?

II. Data Description

The BostonHousing dataset was obtained from the R library mlbench. It contains housing data from 506 census tracts of Boston, Massachusetts, which was obtained from the 1970 United States census. The data consists of 14 variables (i.e. features) and 506 observations. Each observation is a subdivision of a county in Boston. The 14 variables are:

1. **crim**: per capita crime rate by town
2. **zn**: proportion of residential land zoned for lots over 25,000 sq.ft
3. **indus**: proportion of non-retail business acres per town
4. **chas**: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. **nox**: nitric oxides concentration (parts per 10 million)
6. **rm**: average number of rooms per dwelling
7. **age**: proportion of owner-occupied units built prior to 1940
8. **dis**: weighted distances to five Boston employment centres
9. **rad**: index of accessibility to radial highways
10. **tax**: full-value property-tax rate per USD 10,000
11. **ptratio**: pupil-teacher ratio by town
12. **b**: 1000(B - 0.63)^2 where B is the proportion of blacks by town
13. **lstat**: percentage of lower status of the population
14. **medv**: median value of owner occupied homes in USD 1000's

The response variable for this analysis will be **medv**. Therefore, there are a total of 13 dependent variables and 1 dependent (response) variable that will be considered for the initial model.

Prior to fitting a model, the variables were individually analyzed in order to visually identify any potential outliers or issues with our data. To facilitate this task, the density plots of each predictor variables are plotted along with the summary statistics of each variable. Any long tails suggest the presence of extreme values relative to the rest of the group (i.e. potential *outliers*). For example, **crim**, **zn**, **dis**, and **lstat** all have very long right

tails, whereas **b** has a long left tail. The following observations had very large **crim** rates, relative to the rest of the observations: 381, 399, 405, 406, 411, 415, 419, and 428.

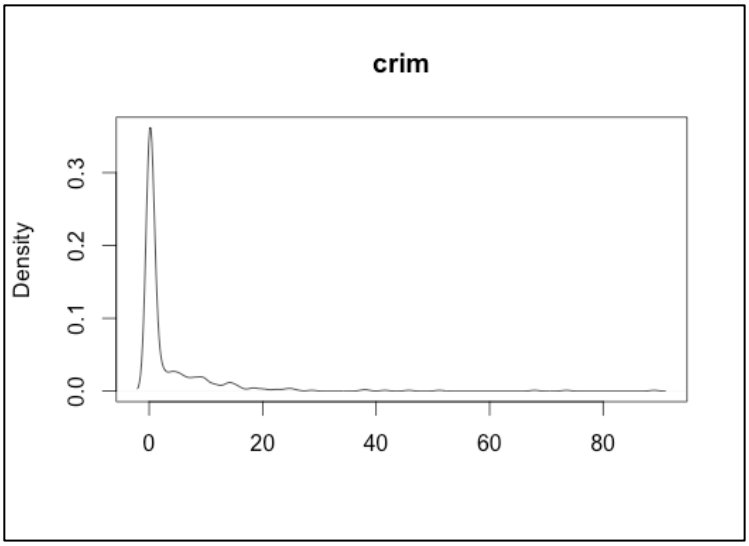

Figure 1: Density Plot for crim

However, it was very difficult to identify a small subset of potential outliers from other long-tailed variables because there were a large number of observations that were spread out on these tails. It was not just a small subset of observations that is causing these tails. This just shows the large amount of skewness for these predictors.

Another important observation from these density plots is that the variables **indus, rad**, and **tax** seem to suggest that there are two potential groups of data. Thinking back to the definition of these variables, it might be possible to split the BostonHousing data into *Urban* (larger proportion of nonretail business and closer access to highways etc.) and *Suburban tracts* and analyze them separately.

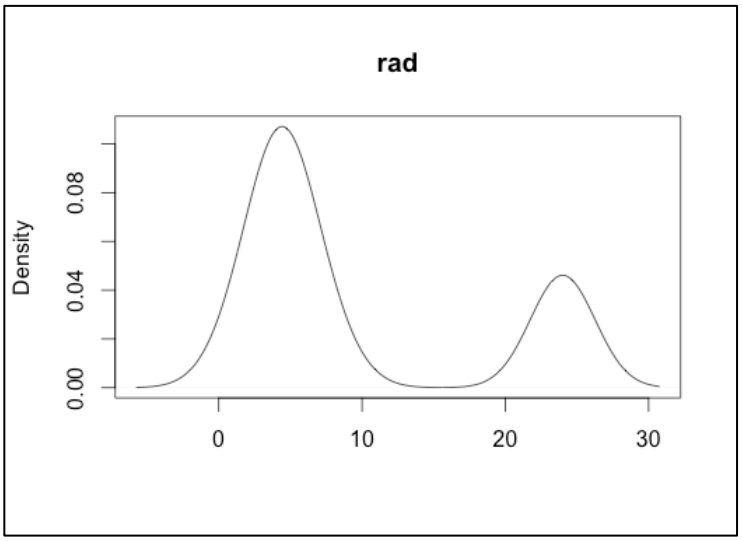

Figure 2 Density Plot for rad

Before creating a model, it is crucial to check how predictors are related with one another and more importantly, with the response variable **medv**. Looking at the last row of the correlation matrix for the predictors, **rm** is identified by its large positive correlation with a value of 0.6954. Intuitively, this makes sense because one can expect that **rm** serves as a proxy for determining the size of a house. Therefore, a larger average number of rooms can imply that the price of the house might be greater because it is a larger house. Also, **ptratio** (pupil-teacher ratio by town) and **lstat** (% of lower status of the population) have relatively large correlation coefficients of -0.5078 and -0.7377 respectively. Again, this does make some intuitive sense because one might expect the price of a house to decrease (i.e. negative correlation) if the area had a large % of lower status residents. Also, low-income areas tend to have larger classroom sizes. These are only preliminary results, so we must continue with caution. Lastly, it is important to note that **lstat** and **medv** do have a negative relationship, however, it shows a very strong non-linear relationship with the response variable, which suggests that a non-linear transformation of this predictor might be useful in our final model.

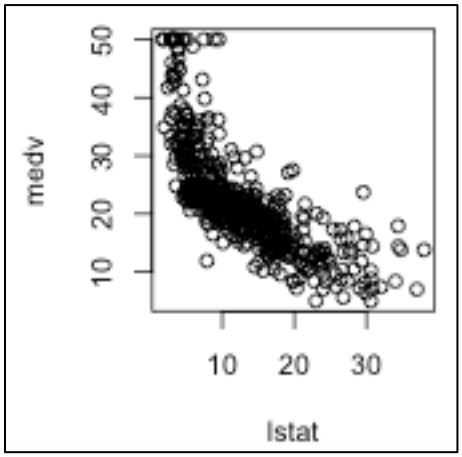

Figure 3 lstat marginal plot

III. Analysis & Results

Initial Full Regression Model

This analysis portion of the assignment involves the supervised learning method of multiple linear regressions paired with some variable selection procedures to try to hone in on the most relevant predictor variables for a model. The goal here is to try to find a model that is well balanced between *prediction* and *inference.* In other words, I want to create a model that can accurately predict **medv** that is not overly complicated in terms of large numbers of predictor transformation.

I begin by creating a *full linear model* of the form:

(1) **medv** = β_0 + β_1 (**crim**) + β_2 (**zn**) + β_3 (**indus**) + β_4 (**chas**) + β_5 (**nox**) + β_6 (**rm**) + β_7 (**age**) + β_8 (**dis**) + β_9 (**rad**) + β_{10} (**tax**) + β_{11} (**ptratio**) + β_{12} (**b**) + β_{13} **lstat** + **e.**

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.65E+01	5.10E+00	7.144	3.28E-12	***
crim	-1.08E-01	3.29E-02	-3.287	0.001087	**
zn	4.64E-02	1.37E-02	3.382	0.000778	***
indus	2.06E-02	6.15E-02	0.334	0.738288	
chas1	2.69E+00	8.62E-01	3.118	0.001925	**
nox	-1.78E+01	3.82E+00	-4.651	4.25E-06	***
rm	3.81E+00	4.18E-01	9.116	< 2e-16	***

age	6.92E-04	1.32E-02	0.052	0.958229	
dis	-1.48E+00	2.00E-01	-7.398	6.01E-13	***
rad	3.06E-01	6.64E-02	4.613	5.07E-06	***
tax	-1.23E-02	3.76E-03	-3.28	0.001112	**
ptratio	-9.53E-01	1.31E-01	-7.283	1.31E-12	***
b	9.31E-03	2.69E-03	3.467	0.000573	***
lstat	-5.25E-01	5.07E-02	-10.347	< 2e-16	***

This initial model has a 4.745 RSE and a R-squared of 0.7406. Notice that the variables **rm**, **lstat**, and **ptratio** appear to be the most significant coefficients when looking at the p-values. These were the exact same coefficients that we previously identified to have the largest (in absolute value) correlation coefficient with **medv**.

Numerous issues are identified when regression diagnostics is performed. For example, the "Residual vs. Fitted" and "Scale-Location" plots appear to have a slight U-shape, which provides is an indication of non-linearity in our data. The "Normal QQ" plot shows a large deviation from the straight line at the right side. This means that our standardized residuals are right-skewed implying that the normality assumption about errors is violated. The same three observations (369, 372, and 373) labeled on the "Residual vs. Fitted" plot are at the right tail end of the "Normal Q-Q" plot. These are three observations that should be tracked throughout the analysis because of their high residual values (i.e. potential outliers). From the "Residual vs. Leverage" plots, four observations are identified as being much larger than 2*(average of leverage) = 0.05533597. These observations are 411, 406, 419, and 38 (points are in ascending order).

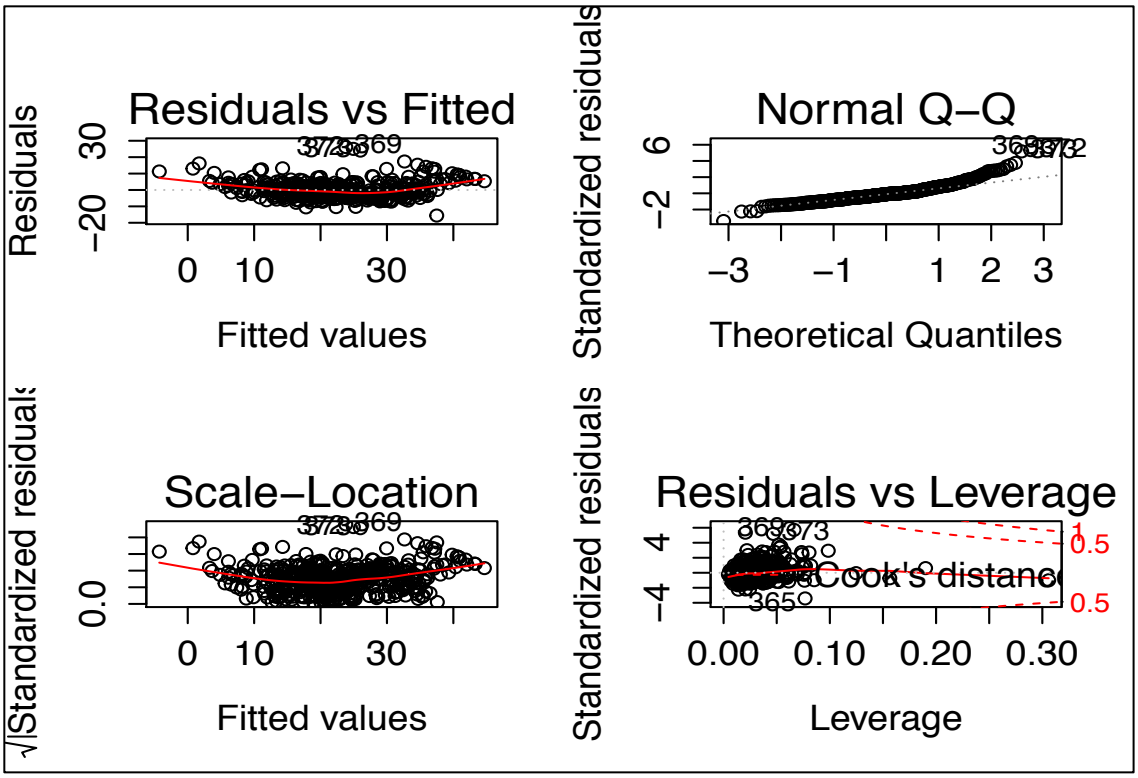

Figure 4: Regression Diagnostics Plot for Model (1)

Choosing an Optimal Model Size

In order to identify the subset of predictor variables that have a large association with the response variable, one can use *best subset selection* provided that our number of predictors *p* is not too large. Discarding those variables that have little association to our response can greatly increase the prediction accuracy of our model by getting rid of the extra noise and can lead to better model interpretability. In order to select the best model size from the best subset selection, my analysis *indirectly* estimated the test error (i.e. Mallow's Cp, AIC, BIC, and Adj R^2) and also *directly* estimated the test error via 10-fold cross- validation.

These five different methods yielded a linear model with 11 predictors, which is of the form:

(2) **medv** = β_0 + β_1 (**crim**) + β_2 (**zn**) + β_3 (**chas**) + β_4 (**nox**) + β_5 (**rm**) + β_6 (**dis**) + β_7 (**rad**) + β_8 (**tax**) + β_9 (**ptratio**) + β_{10} (**b**) + β_{11} (**lstat**) + **e.**

From this procedure, it is appears that **indus** and **age** are NOT significantly associated with **medv**. In other words, they provide very little information in trying to predict our response variable, which is why they were discarded. From the correlation matrix that was computed in 'Data Description' portion, it is important to note that both **indus** and **age** had some pretty high correlation coefficients with other predictors: **dis** vs. **age** -0.7478, **age** vs.

nox 0.7314, **tax** vs. **indus** 0.7207, **dis** vs. **indus** -0.7080, and **nox** vs. **indus** 0.7636. Because of this collinearity with other predictors, the best subset selection procedure dropped these variables because a large amount of their information was contained within other more important predictors for **medv.**

Recall that the marginal plots of our response variable **medv** against each of the predictor variables showed that **lstat** had a very strong non-linear relationship with **medv**, which suggests that a non-linear transformation of this predictor could enhance our model. This can also help to fix the issue of non-linearity of our residuals from the regression diagnostics of model (2), which was similarly present in the diagnostic plots of model (1) in Figure 4. Also, to fix the right-skewness that appeared in our "Normal Q-Q" plot, a log transformation of our response variable appears to be appropriate. The new model is now:

(3) $\log(\mathbf{medv}) = \beta_0 + \beta_1\,(\mathbf{crim}) + \beta_2\,(\mathbf{zn}) + \beta_3\,(\mathbf{chas}) + \beta_4\,(\mathbf{nox}) + \beta_5\,(\mathbf{rm}) + \beta_6\,(\mathbf{dis}) + \beta_7\,(\mathbf{rad}) + \beta_8\,(\mathbf{tax}) + \beta_9\,(\mathbf{ptratio}) + \beta_{10}\,(\mathbf{b}) + \beta_{11}\,(\mathbf{lstat}) + \beta_{12}\,(\mathbf{lstat\text{^}2}) + \mathbf{e}.$

After these two transformations, best subset selection is performed one last time to determine the optimal model size. The final result is a model of size 11 with the predictor **zn** being dropped. After the log transformation of **medv**, notice that our R-squared improved from 0.7868 to 0.8025, despite the fact that we also dropped a predictor term of **zn**. The predictive model that is chosen is:

(4) $\log(\mathbf{medv}) = \beta_0 + \beta_1\,(\mathbf{crim}) + \beta_2\,(\mathbf{chas}) + \beta_3\,(\mathbf{nox}) + \beta_4\,(\mathbf{rm}) + \beta_5\,(\mathbf{dis}) + \beta_6\,(\mathbf{rad}) + \beta_7\,(\mathbf{tax}) + \beta_8\,(\mathbf{ptratio}) + \beta_9\,(\mathbf{b}) + \beta_{10}\,(\mathbf{lstat}) + \beta_{11}\,(\mathbf{lstat\text{^}2}) + \mathbf{e}.$

The "Residual vs. Fitted" and "Scale-Location" plots for model (4) imply homoscedasticity (i.e. constant variance) because there is no 'funnel shape' in the residuals. Also, there is no longer an issue with non-linearity of our residuals because of the polynomial transformation of the **lstat** predictor. Three observations are labeled as potential outliers because of there high residual value. These points are 402, 373, and 372. Notice that the "Normal Q-Q" plot also improved after the transformations. The log transformation of the response got rid of the right-tail that was originally present in the residuals. Looking at the "Residual vs. Leverage" plot and also the "Leverage vs. Index" plot, the observations 381 and 419 are identified as high leverage points because they greatly exceed the 2*(average of leverages) mark and even the 4*(average of leverages) mark.

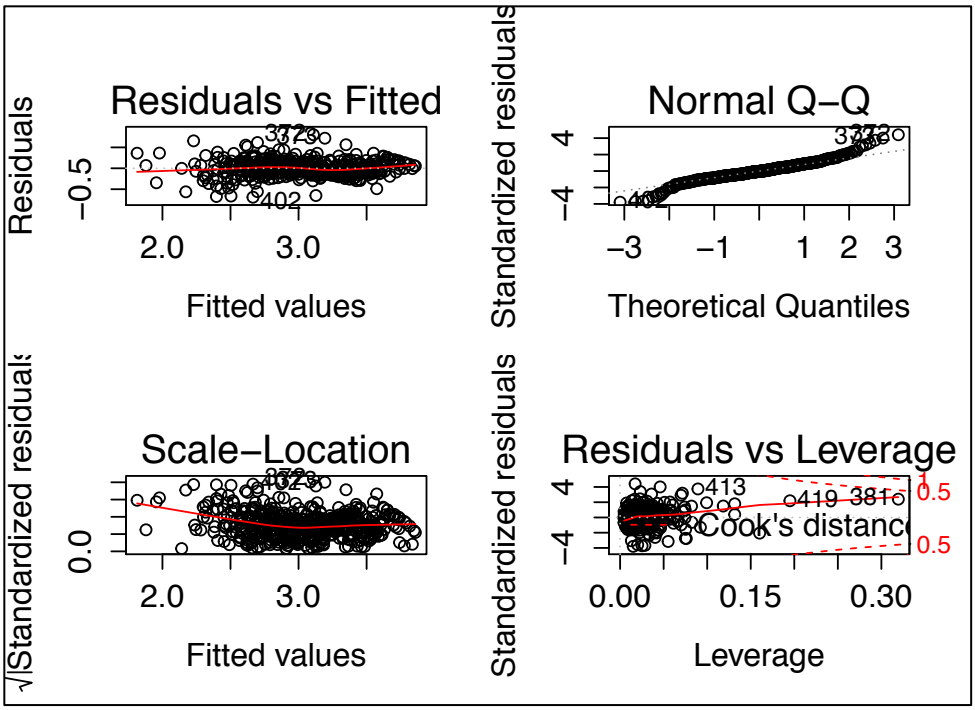

Figure 5: Regression Diagnostics Plot for Model (4)

Recall that Cook's Distance measures how much the regression would change if a point were deleted. Cook's distance is increased by leverage AND by large residuals: a point far from the centroid (of points) with a large residual can severely distort the regression. Three points with the largest Cook's Distance are 381, 413, and 419. However, none of them exceed the 0.5 Cook's distance mark, which is the rule-of-thumb boundary in trying to determine if points have significantly large Cook's Distance. The multiple t-test on the standardized predicted residual with the bonferroni correction could provide statistical proof as to whether or not some observations are outliers. Indeed, observations 401 and 402 are identified as outliers. The following is the output of model (4) after the removal of the outliers, which has an RSE of 0.1786 and an R-squared of 0.8074. Also, the calculation of our Variance Inflation Factor indicates an absence of collinearity between the chosen predictors, which is desired.

Table 1: Summary of Model (4)

	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	3.6264135	0.1862759	19.468	< 2e-16	***
crim	-0.0107555	0.0012483	-8.616	< 2e-16	***
chas1	0.0986104	0.0322255	3.06	0.002334	**
nox	-0.634825	0.1337848	-4.745	2.74E-06	***
rm	0.0839802	0.0154232	5.445	8.19E-08	***
dis	-0.0448149	0.0060681	-7.385	6.55E-13	***
rad	0.0129048	0.0023872	5.406	1.01E-07	***
tax	-0.0004793	0.0001248	-3.84	0.000139	***
ptratio	-0.0351216	0.0046665	-7.526	2.51E-13	***

b	0.0004465	0.0001016	4.397	0.0000135	***
poly(lstat,2)1	-4.8346222	0.2931063	-16.494	< 2e-16	***
poly(lstat,2)2	1.200432	0.1923433	6.241	9.39E-10	***

V. General Discussion/ Conclusion

Our final chosen predictive model is:

(4) log(**medv)** = β_0 + β_1 (**crim) +** β_2 (**chas) +** β_3 (**nox) +** β_4 (**rm) +** β_5 (**dis) +** β_6 **(rad) +** β_7 (**tax) +** β_8 (**ptratio) +** β_9 (**b) +** β_{10} **(lstat) +** β_{11} **(lstat^2) + e.**

The aim of this analysis was to identify which predictor variables are the most important when trying to predict the median value of a Boston house in the 1970s. At the start of this report, the linear model was fitted using all 13 variables. Through the usage of best subset selection, our model discarded 3 variables: **indus**, **age,** and **zn**. The two other big changes that occurred in the model was the addition of the **lstat^2** predictor variable and the log transformation of the response variable: log(**medv)**.

In multiple linear regression, we interpret β_j as the average effect on Y of a one unit increase in X_j , holding all other predictors fixed. Using this interpretation, one can say that -0.0107555 is the average effect on log(**medv)** of a one unit increase in **crim**, holding all other predictors fixed. Similarly, this linear model implies that 0.0839802 is the average effect on log(**medv**) of a one unit increase in **rm**. In other words, as the crime rate of a tract increases, the log of the median value of a house will decrease, provided that all other predictors are fixed. Also, as the average number of rooms per household in a tract increases, the log of the median value of a house will decrease; provide that all other predictors are fixed. Because the log function is monotonic, it does not create huge interpretation issues in our model. Therefore, the β.hat estimates of this model provide significant amount of information about how a particular predictor relates to the response variable of the model. However, we have to proceed with caution here.

Comparing all our β.hat estimates of model (4) to the correlation coefficients of **medv** and each individual predictor, it is clear that all the corresponding predictors have the exact same positive or negative sign except **rad** and **dis**. Notice how **dis** vs. **medv** is positively correlated, whereas the β_5.hat estimate for rad is negative! There appears to be contradictory conclusions being made. This is NOT the case, however. Suppose I had performed a simple linear regression of log(**medv)** ~ **dis.** I would have obtained a *positive* value for the β.hat estimate of **dis** because of the already identified positive correlation between the two variables. However, in the multiple linear regression setting of model (4), we obtained a *negative* β.hat estimate. This difference stems from the fact that in the simple regression case, the slope term represents the average effect of a 1-unit increase in **dis**, *ignoring* every other predictor. In contrast, in the multiple regression setting, the β coefficient for **dis** represents the average effect of increasing **dis** by a unit while holding the rest of the predictors in model (4) fixed. Paying attention to these small details is very important when trying to determine the meaning of β coefficients in multiple linear regressions.

In conclusion, this analysis shows how **crim, chas, nox, rm, dis, rad, tax, ptratio, b, lstat, and lstat^2** are the most important variables in predicting the median value of home in Boston in the 1970s. The prices of the houses in various tracts were affected in various different ways by these predictors. For example, our model suggests that **medv** decreases as **ptratio** increases. In other words, tracts that have a low pupil to teacher ratio seem to be areas that are more expensive. It was pointed out above that the median value of households tends to decrease when crime rates are high. These interpretations are a huge oversimplification of the relationship that all of these variables have with one another. However, this is the beauty of linear modeling. Linear regression is a relatively *inflexible* model that has high levels of interpretability, like we just showed above. However, this comes at the expense of less predictive power. I avoided a large amount of complicated transformations for my predictors because I placed more weight on *inference* for this analysis (i.e. trying to determine the relationship between my predictors and my response). The final model for this analysis had a good Test MSE, without having to sacrifice much interpretability.

Table 2: Correlation of medv With Predictors

	medv
crim	-0.3883046
zn	0.3604453
indus	-0.4837252
chas	0.1752602
nox	-0.4273208
rm	0.6953599
age	-0.3769546
dis	0.2499287
rad	-0.3816262
tax	-0.4685359
ptratio	-0.5077867
b	0.3334608
lstat	-0.7376627
medv	1