PREDICTING BOSTON HOUSE PRICES USING DATA MINING TECHNIQUES

-BY ARUN PALLATH

EXECUTIVE SUMMARY

Boston Housing Data

PROBLEM & APPROACH:

Boston housing data is a data set in package MASS. The data set has 506 rows and 14 columns. This report provides an analysis and evaluation of the factors affecting the median value of the owner-occupied homes in the suburbs of Boston. The in-built data set of Boston Housing Data in package MASS is used for this analysis and various factors about the structural quality, neighbourhood, accessibility, and air pollution such as per capita crime rate by town, proportion of non-retail business acres per town, index of accessibility to radial highways etc are considered for this study.

Methods of analysis includes the following:

- Summary statistics of the variables and finding correlation between variables
- Exploratory data analysis using visualization
- Random sampling of data set into 80/20 training and testing data set
- Fitting a linear regression model and performing various variable selection methods
- Performing Cross Validation
- Fitting a Regression Tree
- Finally, comparing the models based on in-sample (MSPE) and out-of-sample prediction errors (MSPE).
- Repeat all the modelling techniques using another random sample and compare the results.
- Compare various Tree models with Linear Regression model

RESULTS:

We performed all the analysis as mentioned above in the approach and below is the summary the results of our analysis.

Parameters	Random Sample 1	RandomSample2
Final Model	**medv ~ lstat + rm + ptratio + dis + nox + black + chas + zn + rad + tax + crim**	**medv ~ lstat + rm + ptratio + dis + nox + black + chas + zn + rad + tax + crim**
AIC	2431.085	2388.7
In-Sample MSE	23.230	20.917
Out-of-Sample MSE	20.767	28.936
CV Score	23.748	23.37
Regression Tree Out-of-Sample MSE	18.506	28.531

Table 1: Summary of Boston Housing data model

We can see that we obtain almost similar results when we do model on both the random samples. We have concluded that the best model to predict Boston Housing prices is:

medv ~ lstat + rm + ptratio + dis + nox + black + chas + zn + rad + tax + crim

A comparison between Linear Regression and Tree-Based Models:

Model	In Sample MSE	Out-of-Sample MSE
Linear Regression	22.91754	23.06699
Regression Tree	13.18067	23.74355
Bagging	10.79223	18.58307
Random Forest	10.94049	10.84043
Boosting	0.01637	10.89395

We can observe that:

- The tree models have a lower test MSE and thus they perform better than a linear regression model.
- The test MSE with a bagged regression tree is lower than a regression tree and thus performs better.
- Random Forest yields an improvement over bagging.
- The boosted model test MSE is similar to the test MSE for random forests and superior to that for bagging.

BOSTON HOUSING DATA

1.Exploratory Data Analysis:

- This dataset contains a set of 506 observations under 14 attributes - crim, zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, b, lstat, medv.
- After random sampling the dataset to train and test, we have 404 observations and 14 attributes in the training dataset.
- Data Types: All the 14 attributes are of float data type.

boston_train.median()		boston_train.mean()		boston_train.std()	
crim	0.253715	crim	3.679883	crim	8.691423
zn	0.000000	zn	10.983051	zn	22.966498
indus	9.690000	indus	11.129181	indus	6.819598
chas	0.000000	chas	0.064972	chas	0.246825
nox	0.538000	nox	0.552345	nox	0.110090
rm	6.241000	rm	6.313678	rm	0.687150
age	76.700000	age	68.350282	age	27.945635
dis	3.142300	dis	3.732462	dis	2.019922
rad	5.000000	rad	9.737288	rad	8.834418
tax	335.000000	tax	412.254237	tax	169.211227
ptratio	19.100000	ptratio	18.542938	ptratio	2.086872
b	391.280000	b	352.176808	b	98.330740
lstat	10.685000	lstat	12.468757	lstat	7.056974
medv	21.700000	medv	22.527401	medv	9.037079
dtype: float64		dtype: float64		dtype: float64	

Table 4: Mean, Median, Standard Deviation

Observations:

- For most of the parameters, the mean is greater that the median which indicates positive skewness.
- But for age, ptratio and b, the mean is lesser than the median which indicates a slight negative skewness.

Outliers:

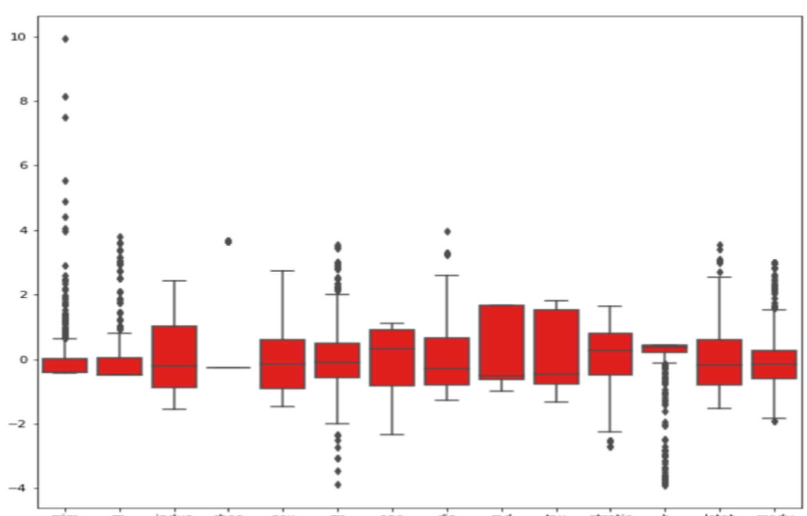


Fig 1: Boxplot of variables

We have used box plots to understand the distribution of each variable. From the boxplots of indus, nox, age, rad, tax, we can see the presence of outliers. Nox and rm variables looks to have a symmetrical distribution. We can ignore the distribution of chas variable as it is a binary variable. The distributions of rad and crim has a high negative skewness. The variables crim, zn, rm, b and medv has a lot of outliers.

Histograms:
We will visualize the distributions of all variables using Histograms.

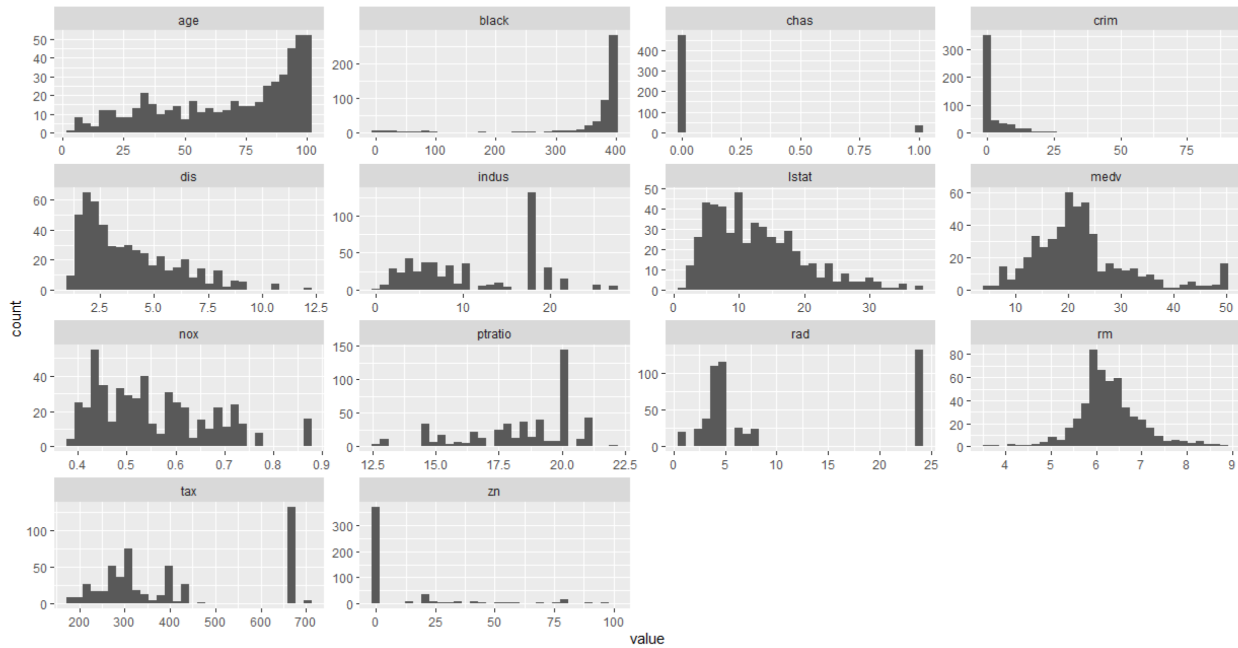


Fig 2: Histograms

We can observe that rm variable is almost normally distributed and all other variables have a skewed distribution.

Correlation between different variables:

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	medv
crim	1.000000	-0.198903	0.412795	-0.058451	0.443012	-0.180575	0.351187	-0.384644	0.618935	0.579926	0.293060	-0.408542	0.434678	-0.398383
zn	-0.198903	1.000000	-0.521400	-0.031289	-0.530677	0.282557	-0.563924	0.649600	-0.312851	-0.310144	-0.365969	0.180331	-0.398189	0.343446
indus	0.412795	-0.521400	1.000000	0.063564	0.779504	-0.358831	0.644122	-0.703315	0.616156	0.716468	0.369002	-0.368700	0.585845	-0.488234
chas	-0.058451	-0.031289	0.063564	1.000000	0.079657	0.125978	0.077147	-0.095396	-0.016834	-0.040483	-0.145124	0.075500	-0.067252	0.209895
nox	0.443012	-0.530677	0.779504	0.079657	1.000000	-0.261569	0.737838	-0.775564	0.663895	0.706992	0.259166	-0.396630	0.586547	-0.447606
rm	-0.180575	0.282557	-0.358831	0.125978	-0.261569	1.000000	-0.205065	0.155265	-0.222319	-0.290595	-0.347913	0.101442	-0.596364	0.697851
age	0.351187	-0.563924	0.644122	0.077147	0.737838	-0.205065	1.000000	-0.728703	0.463305	0.510128	0.274035	-0.282774	0.586916	-0.390828
dis	-0.384644	0.649600	-0.703315	-0.095396	-0.775564	0.155265	-0.728703	1.000000	-0.505366	-0.527397	-0.249291	0.297915	-0.474710	0.235943
rad	0.618935	-0.312851	0.616156	-0.016834	0.663895	-0.222319	0.463305	-0.505366	1.000000	0.917384	0.470680	-0.475309	0.527026	-0.438319
tax	0.579926	-0.310144	0.716468	-0.040483	0.706992	-0.290595	0.510128	-0.527397	0.917384	1.000000	0.447915	-0.468355	0.563245	-0.514200
ptratio	0.293060	-0.365969	0.369002	-0.145124	0.259166	-0.347913	0.274035	-0.249291	0.470680	0.447915	1.000000	-0.211632	0.391535	-0.527249
b	-0.408542	0.180331	-0.368700	0.075500	-0.396630	0.101442	-0.282774	0.297915	-0.475309	-0.468355	-0.211632	1.000000	-0.388630	0.365048
lstat	0.434678	-0.398189	0.585845	-0.067252	0.586547	-0.596364	0.586916	-0.474710	0.527026	0.563245	0.391535	-0.388630	1.000000	-0.733571
medv	-0.398383	0.343446	-0.488234	0.209895	-0.447606	0.697851	-0.390828	0.235943	-0.438319	-0.514200	-0.527249	0.365048	-0.733571	1.000000

Fig 3: Correlation matrix

Before starting any analysis, we need to understand what relationship the response variables have on the predictor variables. We can understand this relationship by using correlation matrix. From the above correlation matrix, we can see that the response variable medv has a positive correlation with zn, chas, rm, dis and b. Medv also has a negative correlation with crim, indus, nox, age, rad, tax, ptratio and lstat. Rm, tax, ptratio and lstat has the highest effect on the response variable medv as the correlation value is greater than 0.5. We can start building a model based on the strength of the effect of the predictor variables on the response variable.

2. Modelling :

Here, we are trying to model the relationship between a scalar response(medv) and one or more explanatory variables (also known as dependent and independent variables).
Before finding out a model, we first prepare the data: ie, splitting data into training and testing samples. We have sampled the data in a 80:20 ratio, ie 80% data for training and 20% data for testing. The regression model will be built on the training set and future performance of our model will be evaluated with the test set. A general linear regression is done on the data. Then, various other variable selection techniques (Best subset, backward elimination, forward selection, and stepwise selection) are used to choose the best model.

<u>Random Sample 1</u>

Linear Regression:

We have created a linear regression model on Boston Housing Dataset using all covariates and response variable as medv with no variable transformation. Below is the summary statistics of this linear regression.

```
Call:
lm(formula = medv ~ ., data = Boston_train)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.8618  -2.7972  -0.6081   1.8075  25.8580 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.694e+01  5.438e+00   6.793 3.58e-11 ***
crim        -8.134e-02  4.008e-02  -2.030 0.043002 *  
zn           4.013e-02  1.489e-02   2.696 0.007290 ** 
indus        2.903e-02  6.608e-02   0.439 0.660680    
chas         3.109e+00  8.987e-01   3.459 0.000594 ***
nox         -1.844e+01  3.988e+00  -4.624 4.95e-06 ***
rm           3.710e+00  4.404e-01   8.424 5.18e-16 ***
age          1.237e-04  1.397e-02   0.009 0.992941    
dis         -1.467e+00  2.149e-01  -6.825 2.92e-11 ***
rad          2.804e-01  7.074e-02   3.964 8.61e-05 ***
tax         -1.064e-02  4.011e-03  -2.652 0.008292 ** 
ptratio     -9.619e-01  1.415e-01  -6.797 3.49e-11 ***
black        9.979e-03  2.867e-03   3.481 0.000550 ***
lstat       -5.352e-01  5.362e-02  -9.981  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.783 on 441 degrees of freedom
Multiple R-squared:  0.7341,	Adjusted R-squared:  0.7262 
F-statistic: 93.63 on 13 and 441 DF,  p-value: < 2.2e-16
```

Table 5: Summary of the model

The variables indus and age seems to be statistically insignificant as they have a p-value greater than 0.05. All other variables are significant as they have a p-value less than 0.05. The adjusted R-squared value is 72.62%, which means that 72.62% of the variation in response variable can be explained by the model.

Variable Selection:

We used various variable selection criteria like AIC, BIC and LASSO regression to find the best model. The models suggested by each of the selection methods are as follows:

Model Type	Model	AIC value
AIC	**medv ~ lstat + rm + ptratio + dis + nox + black + chas + zn + rad + tax + crim**	**2431.085**
BIC	**medv ~ lstat + rm + ptratio + dis + nox + black + chas + zn**	**2440.014**
LASSO	**medv ~ crim + lstat + rm + ptratio + dis + nox + black + chas**	**2453.143**

Table 6: Variable Selection model parameters (Random Sample 1)

We can find that the least AIC value was observed for the model suggested by AIC selection criteria. So, we consider it as our final model. The final model we selected is:

medv ~ lstat + rm + ptratio + dis + nox + black + chas + zn + rad + tax + crim

We calculated the model mean squared error (MSE) to be **23.230**.

We used the testing dataset of 20% data to find out the model's out of sample performance. The out-of-sample MSPE is calculated to be **20.767.**

Cross Validation:

Cross validation is an alternative approach to training/testing split. We performed 5-fold Cross validation on the original data and found out the performance characteristics of the model. The model MSE came out to be **23.748.**

We can see that the cross validation MSE is little higher than the MSPE observed from the linear regression model. This is because Cross validation provides a less sample specific estimate of the MSE. Cross Validation also reduces chances of overfitted data.

Regression Tree:

We used the training dataset to plot a regression tree.

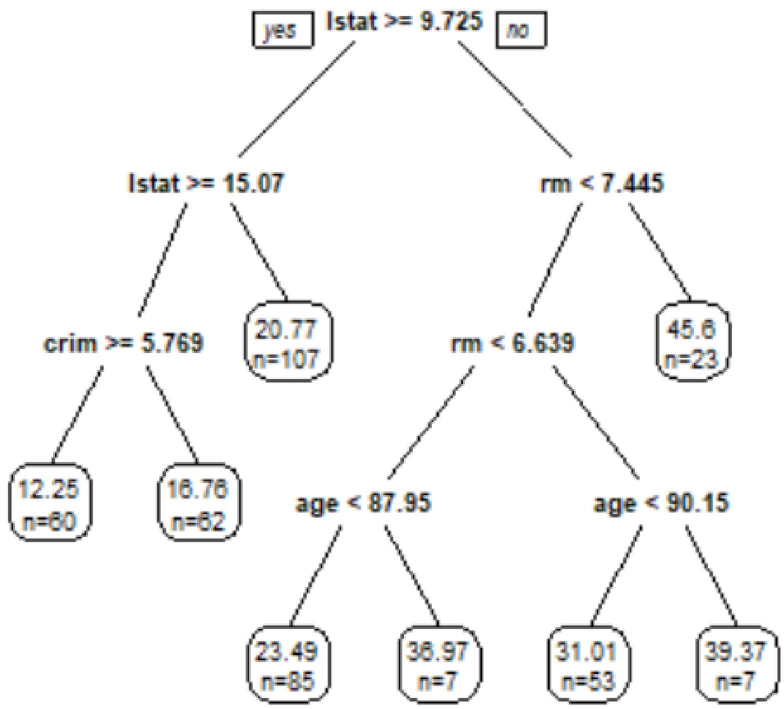


Fig 4: Regression Tree (Random Sample 1)

We can use this Regression tree to predict the response values .

The in-sample and out-of-sample prediction for regression trees is also similar to *lm* and *glm* models. The in-sample MSE was calculated to be **15.614** and the out of sample MSPE came out to be **18.506.**

If we compared the performance of Regression Tree to the performance of the final model we obtained after variable selection procedures, we can see that Regression Tree gives a **smaller MSE** and thus it performs better than the linear regression model.

<u>Random Sample 2</u>

We repeated all these modelling methods again but this time with another random 80:20 sample of training and testing data.

We performed all the variable selection methods on the new dataset and the model suggestions and performances were as below:

Model Type	Model	AIC value
AIC	medv ~ lstat + rm + ptratio + dis + nox + black + chas + zn + rad + tax + crim	2388.7
BIC	medv ~ lstat + rm + ptratio + dis + nox + black + chas + zn	2404.011
LASSO	medv ~ crim + lstat + rm + ptratio + dis + nox + black + chas	2402.691

Table 7: Variable Selection model parameters(Random Sample 2)

We can find that the least AIC value was observed for the model suggested by AIC selection criteria. So, we consider it as our final model. The final model we selected is:

medv ~ lstat + rm + ptratio + dis + nox + black + chas + zn + rad + tax + crim

Regression Tree:

We used the training dataset to plot a regression tree.

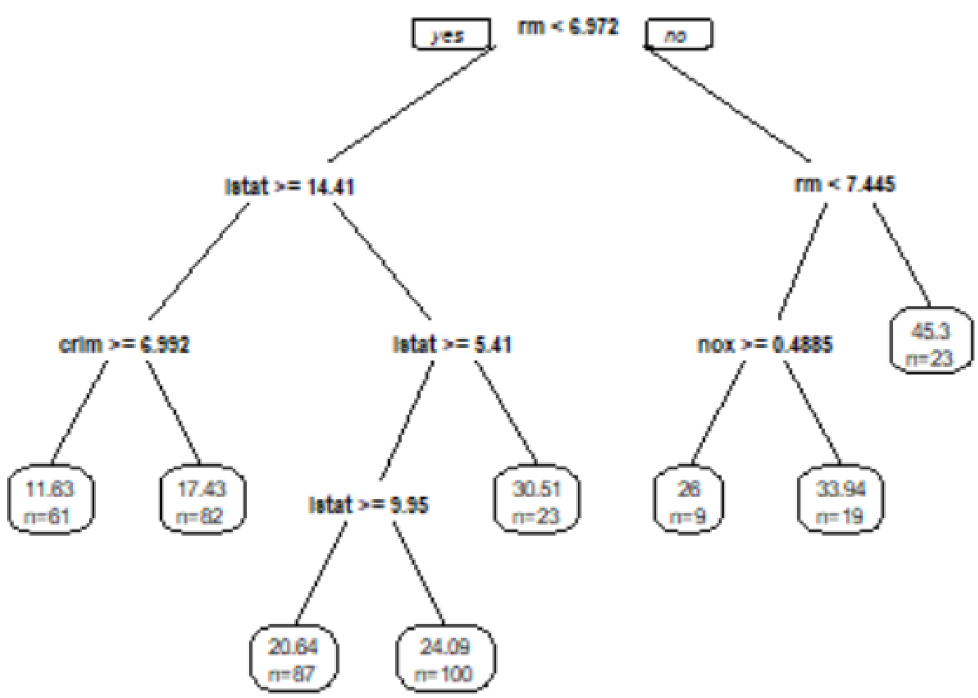


Fig 5: Regression Tree (Random Sample 2)

We compared both the model performances and the results are as below:

Parameters	Random Sample 1	Random Sample2
Final Model	**medv ~ lstat + rm + ptratio + dis + nox + black + chas + zn + rad + tax + crim**	**medv ~ lstat + rm + ptratio + dis + nox + black + chas + zn + rad + tax + crim**
AIC	2431.085	2388.7
In-Sample MSE	23.230	20.917
Out-of-Sample MSE	20.767	28.936
CV Score	23.748	23.37
Regression Tree Out-of-Sample MSE	18.506	28.531

Table 8: Comparison of models with different random samples

We can see that we obtain almost similar results when we do model on both the random samples. We have concluded that the best model to predict Boston Housing prices is:

medv ~ lstat + rm + ptratio + dis + nox + black + chas + zn + rad + tax + crim

<u>Comparing various Tree models:</u>

BAGGING: We performed bagging regression trees with 100 bootstrap replications. The In-Sample performance was calculated to be 10.79223. The out-of-sample MSE came out to be 18.58307.

RANDOM FOREST: It works exactly the same way as bagging except that it uses a smaller value of the 'mtry' argument. We are using p/3 variables which is the default for building a randomForest() of regression trees. The in sample MSE of the random forest model is 10.94049.

The test MSE of the model is 10.84043. This indicates that random forests yielded an improvement over bagging in this case.

BOOSTING: The relative influence plot from fitted boosted regression tree is shown below:

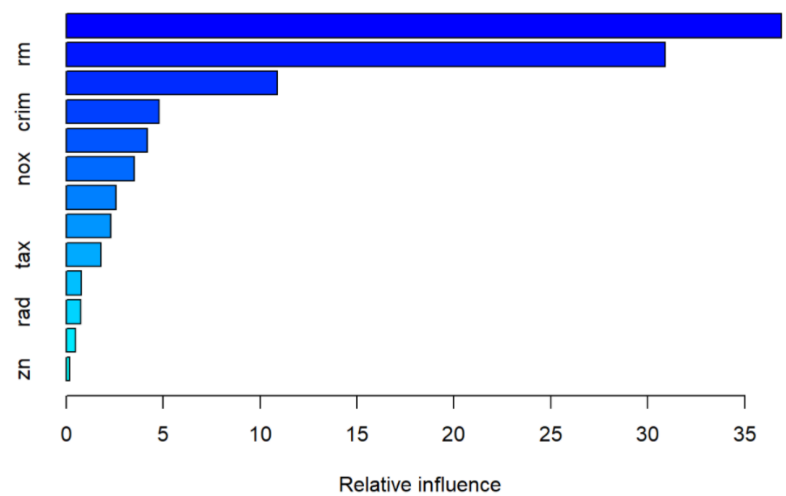
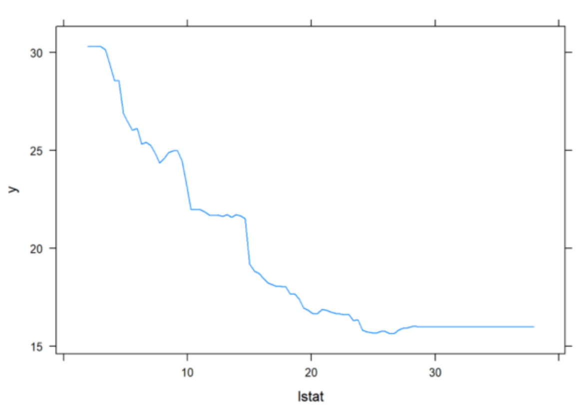


Fig 6: Relative influence plot (Boosting)

We can observe a relative influence plot and the relative influence statistics. We can see that lstat and rm are the most important variables. We can also produce partial dependence plots of these two variables.



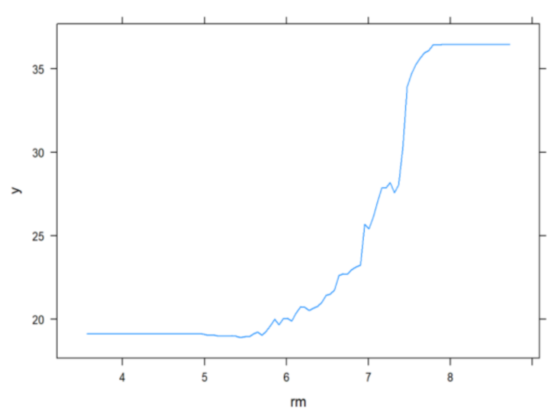


Fig 7: Partial dependence plots (Boosting)

We can see that median house prices are increasing with rm and decreasing with lstat as expected.

The in sample MSE is calculated to be 0.01637563.

The test MSE obtained is 10.89395 almost similar to the test MSE for random forests and better than that for bagging.