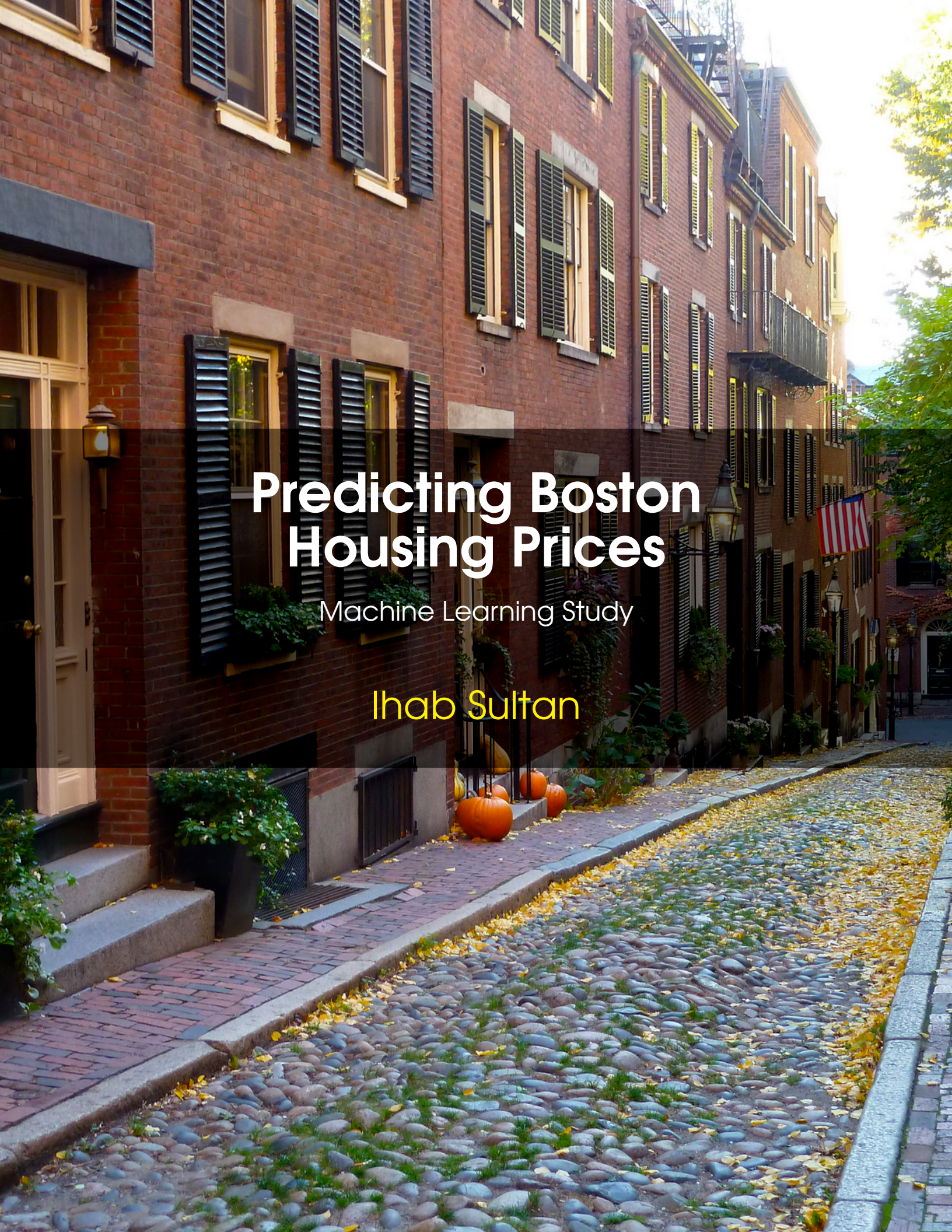
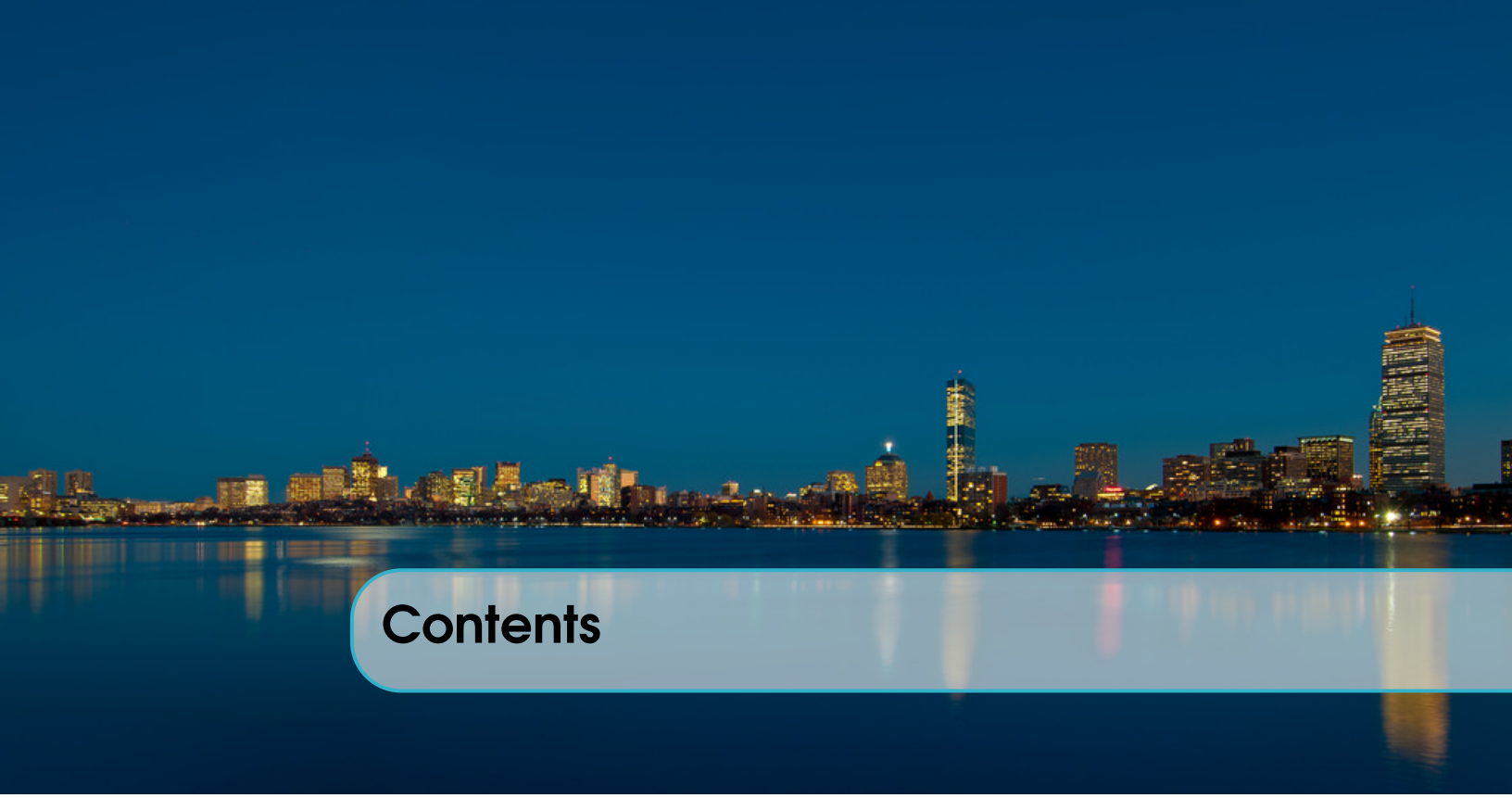


Predicting Boston Housing Prices

Machine Learning Study

Ihab Sultan



Contents

1 Introduction 2
1.1 Project Overview 2
1.2 Software and Libraries 2

2 Boston Housing Data Set 3
2.1 Date Set Source 3
2.2 Available Attributes 3

3 Project Report 4
3.1 Statistical Analysis and Data Exploration 4
3.2 Evaluating Model Performance 4
3.3 Analyzing Model Performance 6
3.4 Model Prediction 8

4 Further Explorations with R 9
4.1 Prices Histogram 9
4.2 Linear Regression 9
4.3 Correlation Visualizations 11

1. Introduction

1.1 Project Overview

The goal of this project is to apply basic machine learning concepts on data collected for housing prices in the Boston, Massachusetts area to predict the selling price of a new home by performing the following:

- Explore the data to obtain important features and descriptive statistics about the dataset.
- Split the data into testing and training subsets.
- Determine a suitable performance metric for this problem.
- Analyze performance graphs for a learning algorithm with varying parameters and training set sizes.
- Pick the optimal model that best generalizes for unseen data.
- Finally, test this optimal model on a new sample and compare the predicted selling price to the statistics.

In section 3 we assume the main features affecting house prices are its accessibility to radial highways, number of rooms per dwelling, and finally, the % lower status of the population.

Section 4 is the core study, where we describe the results obtained from the Decision Tree Regressor, and use grid search and cross-validation to optimize the model parameters.

We conclude with section 5 on further explorations using R statistical package, and we come to know that both the weighted distance to main employment centers and pupil-teacher ratio are actually more statistically significant than the distance to radial highways.

1.2 Software and Libraries

The following SW was used in the first part of the project:

- Python 2.7
- NumPy
- scikit-learn

In the last part of this project, R was used as an EDA tool:

- R 3.2.3 using corrplot and GGally

---

2. Boston Housing Data Set

2.1 Date Set Source

Data set can be found at: `https://archive.ics.uci.edu/ml/datasets/Housing`

2.2 Available Attributes

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 $10,000
11. **PTRATIO**: pupil-teacher ratio by town
12. **B**: $1000(B_k - 0.63)^2$ where B_k is the proportion of blacks by town
13. **LSTAT**: % lower status of the population
14. **MEDV**: Median value of owner-occupied homes in $1000's

It is worth noting that the last attribute (*MEDV*) is separated as an output label in sklearn dataset.

3. Project Report

3.1 Statistical Analysis and Data Exploration

- **How many data points (houses) were collected?**
  506
- **How many features are present for each house?**
  13 features
- **What is the minimum housing price? The maximum?**
  minimum = 5, maximum = 50 (in thousands)
- **What is the mean housing price? The median?**
  mean = 22.533, median = 21.2 (in thousands)
- **What is the standard deviation of all housing prices?**
  9.188

1) Of the available features for a given home, choose three you feel are significant and give a brief description for each of what they measure.

- *RAD*: index of accessibility to radial highways. It is expected that areas with better highway accessibility have higher prices on average.
- *RM*: average number of rooms per dwelling. With more rooms, the average price is expected to be higher.
- *LSTAT*: % lower status of the population. It is expected that areas with higher percentage of lower status population would have lower median prices.

2) Using your client's feature set "CLIENT FEATURES" in the template code, which values correspond to the chosen features?

- *RAD*: 24
- *RM*: 5.609
- *LSTAT*: 12.13

3.2 Evaluating Model Performance

3) Why do we split the data into training and testing subsets?

The data is split into training and testing sets to allow the trained model to be tested on different samples than the one used for training.
Such process would be very useful in detecting overfitting of our model which cannot be deduced from model performance on training data by itself. In general, any over-commitment to training data

would fail to generalize to the new data set.
Moving some of our collected samples to the test set obviously happens at the cost of decreasing the number of samples available for training, and that is the price we pay for this procedure.

4) Which performance metric below is most appropriate for predicting housing prices and analyzing error? Why?

- **Accuracy**
- **Precision**
- **Recall**
- **F1 score**
- **Mean Squared Error (MSE)**
- **Mean Absolute Error (MAE)**

As the task at hand involves regression (rather than classification), we'll have to use a regression performance metric. Two common regression performance metrics are:

- *Mean Squared Error*: This is the most useful metric for the case at hand, it magnifies large errors (through squaring) which makes sense for a price prediction model.
  The fact that this turns out to be a differentiable function as well is pretty useful for error (or loss) analysis, albeit this fact won't be used for the task at hand in this project.
- *Mean Absolute Error*: This is another valid metric for regression, though it won't be picked in this project due to the points mentioned above.

5) What is the grid search algorithm and when is it applicable?

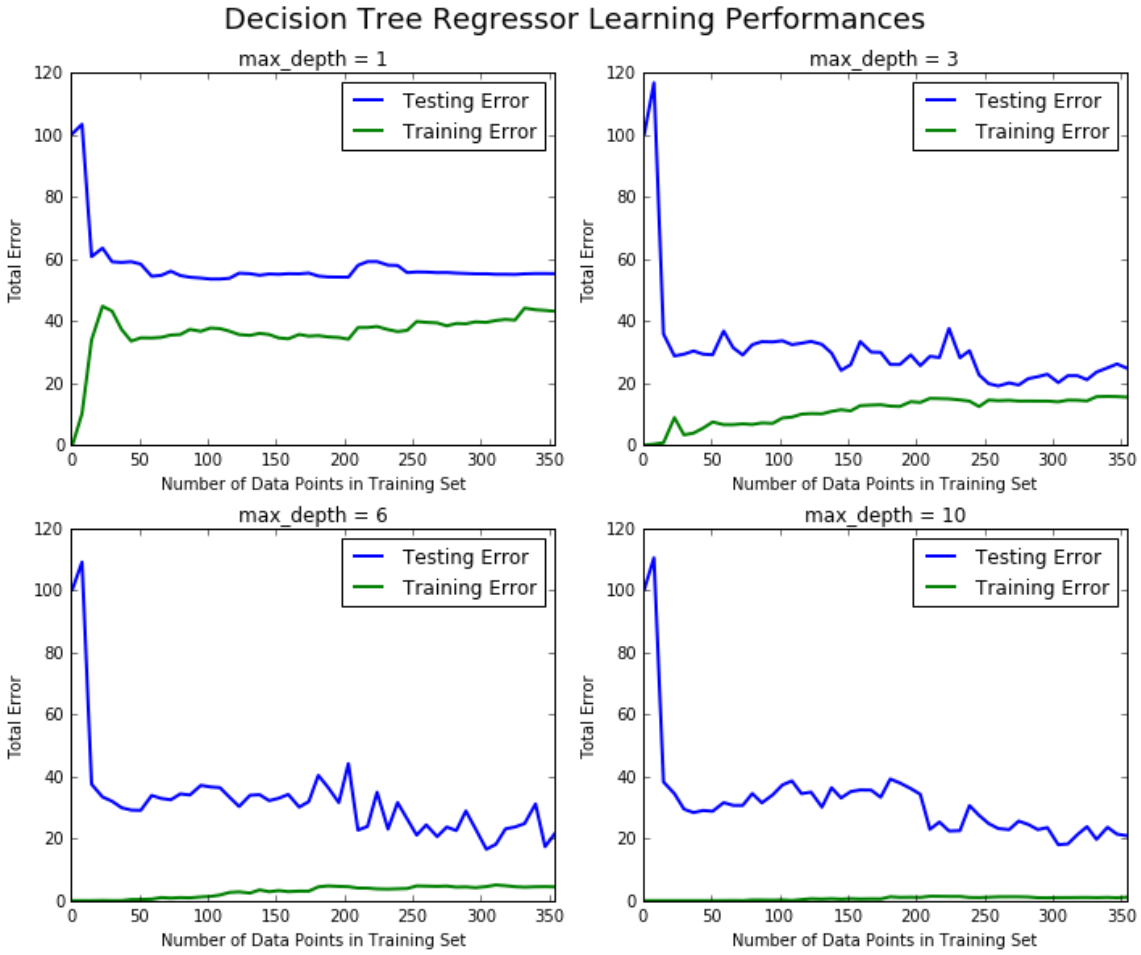
Grid Search is an algorithm for finding optimal model parameters (also known as hyperparameters) through trying multiple values from the parameter space and comparing performance metric obtained from different trials. In grid search, all provided combinations will be tested.
The grid search algorithm is applicable to any hyperparameter which can take one of multiple values in the parameter space.

6) What is cross-validation and how is it performed on a model? Why would cross-validation be helpful when using grid search?

Cross-Validation is the process of splitting samples into a fixed number of groups (called folds), and serially choosing one of the groups for validation, while using other groups for training.
The process would increase the training time by k-fold (where k is the number of groups) but would increase the model robustness to the choice of the training and test sets. It is recommended to shuffle the input samples prior to cross-validation in order to randomize any localized patterns found in the input samples.
Using cross-validation with grid search means that the hyper-parameters are optimized properly without being tied to a specific training or test sets.

3.3 Analyzing Model Performance

In the learning curves below, notice that maximum data set size is 354, due to the fact that we retain only 70% of the original data points for training and leave the rest for the test set.



7) Choose one of the learning curve graphs your code creates. What is the max depth for the model? As the size of the training set increases, what happens to the training error? Describe what happens to the testing error.

At max_depth = 6, the training error starts at zero and gradually increases with more training points, this is expected as the model won't be able to minimize the error to all data points when the number of training points increase.

Regrading testing error, we notice that error starts very high, which is expected given that we are asking the model to generalize using only few training samples. Error starts decreasing after that, and as we add more samples to the training set, we observe that MSE generally keeps decreasing except for few error bumps due to the nature of the data.

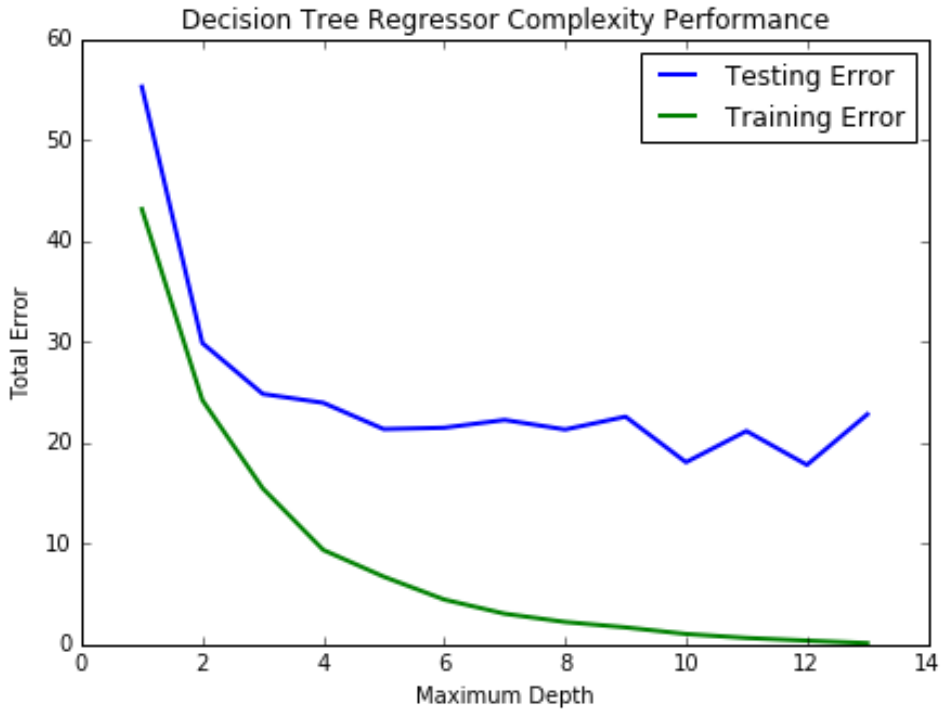
8) Look at the learning curve graphs for the model with a max depth of 1 and a max depth of 10. When the model is using the full training set, does it suffer from high bias or high

variance when the max depth is 1? What about when the max depth is 10?

When max_depth = 1, the model suffers from high bias. This is clear from the observation that both training and testing errors have converged to a high error (MSE), as follows:

- The fact that training and testing errors quickly converged to nearby values indicates that the model has extracted all the useful information from the training data and is not able to improve.
- Since the error level at which this happened is very high, this means the model itself is not capable of properly representing the underlying relationship between independent variables and regression output, indicating model's high bias.

At max_depth = 10 however, the testing error appears to be trending in the increasing direction which is a sign of high variance. In addition, one can notice that the delta between testing and training errors is very large at this depth, which is another indication of overfitting.



9) From the model complexity graph, describe the training and testing errors as the max depth increases. Based on your interpretation of the graph, which max depth results in a model that best generalizes the dataset? Why?

The training above happens on the full data set with increasing model complexity. We can clearly see that as complexity increases, the training error is decreasing. With testing error, there is a different story, we have some signs of overfitting beyond max_depth = 10.
By visual inspection, I would choose max_depth of 10 as a good balance between model complexity and model generalization.

3.4 Model Prediction

10) Using grid search, what is the optimal max depth for your model? How does this result compare to your initial intuition?

Optimal max_depth from grid search optimization is 6 (average of 4 and 8), which does not seem to be the optimal point from the complexity curve above.
One theory is that cross validation is a better way of verifying the generalization ability of a model compared to fixed training and testing sets.

11) With your parameter-tuned model, what is the best selling price for your client's home? How does this selling price compare to the statistics you calculated on the dataset?

Selling price for client's home is 21.63 thousand dollars per optimized model above. This price lies in between median and average prices of 21.2 and 22.533 respectively.

12) In a few sentences, discuss whether you would use this model or not to predict the selling price of future clients' homes in the Boston area.

This model is definitely better than using 'no model' at all. It gives a reasonable accuracy and is pretty fast to train and test.
With that being said, there is a good space for improvement in terms of model accuracy, and other models can be compared to DecisionTrees regression used in this project.
Another point to consider, is that the number of features used in the model might demand a larger data set. Given that this is the full data set provided to us, we can try removing irrelevant features and allow the model to utilize more prominent ones.
Finally, due to the small size of the data set, we had to use all of it in the model optimization step, whereas ideally we would retain a portion of the samples for final testing.

In summary, I would definitely prefer this model to not using any model at all, but I would do more work in order to optimize it further before using it in real life projects.

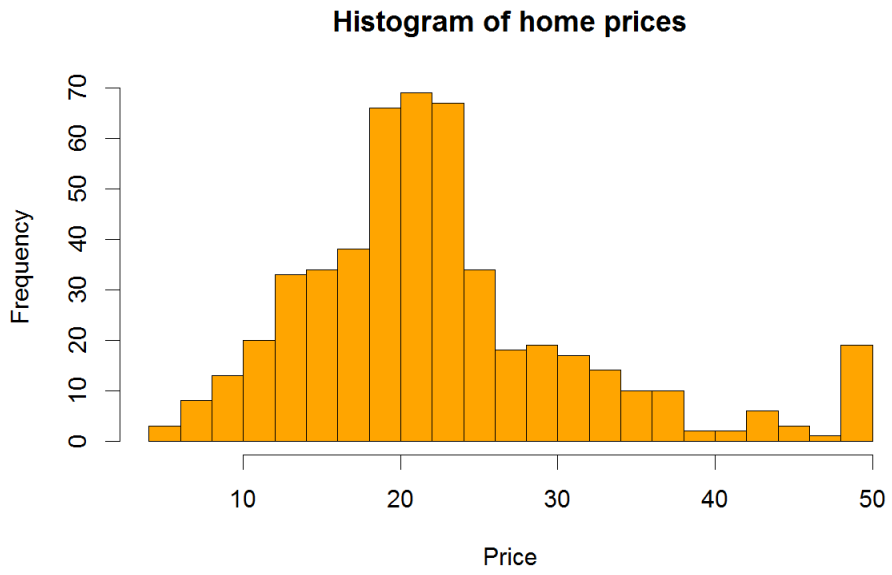
4. Further Explorations with R

This is not part of the project, but an addendum for plots generated using R statistical package from the housing dataset.

Notice that specifically in this section, MEDV (the median home value) is treated as another attribute instead of as a label.

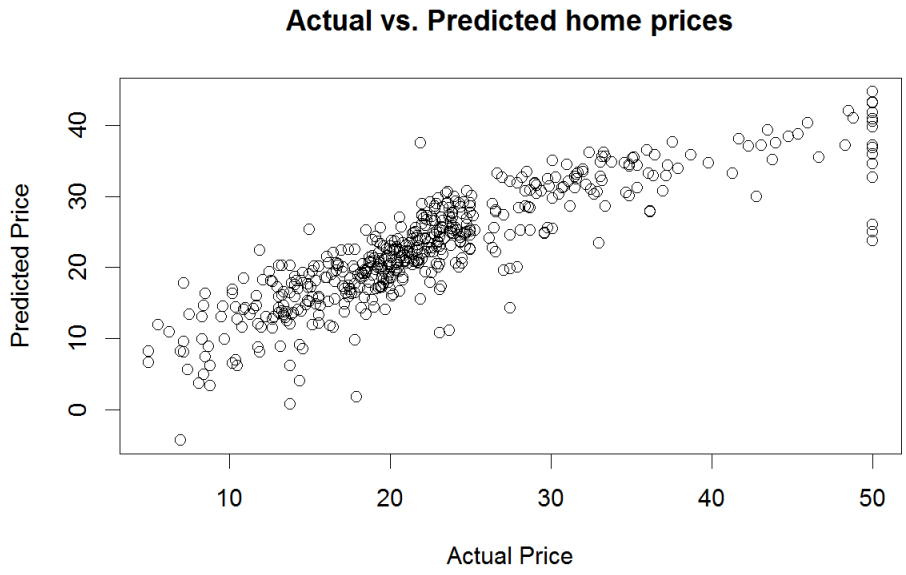
4.1 Prices Histogram

The following histogram shows some unexpected frequency at $50,000. This probably is due to clamping max value to 50K in the data set.



4.2 Linear Regression

A linear regression model was fitted to the data, following plot shows predicted against correct home prices for all samples.



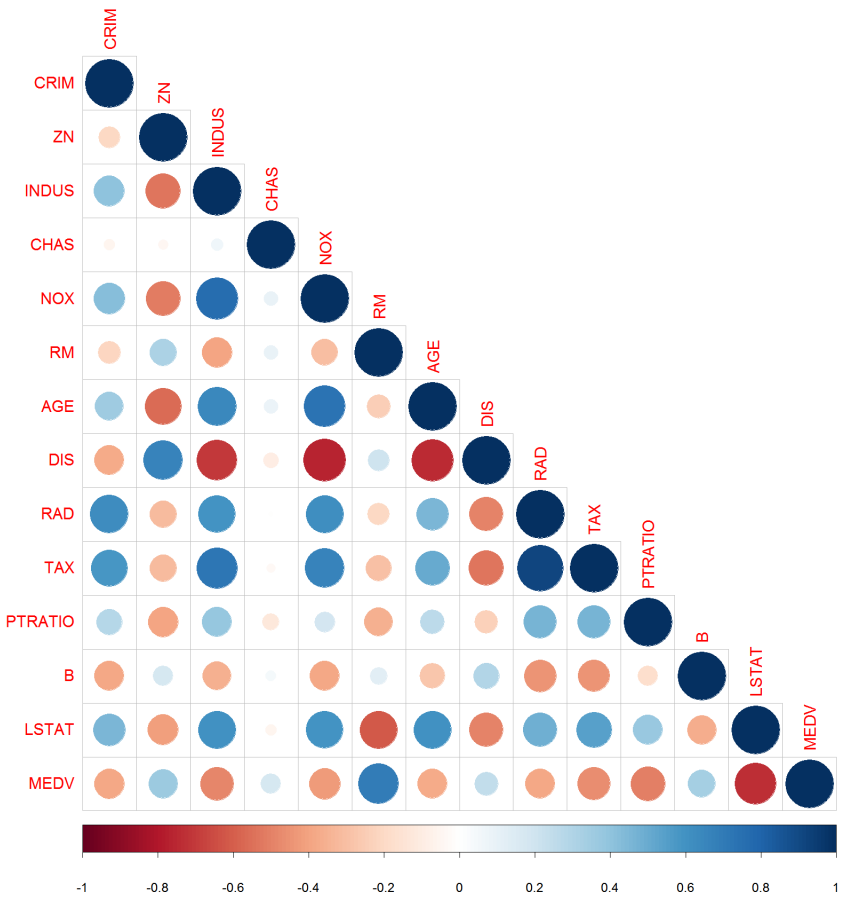
According to the generate linear regression model, following Parameters are the most significant:

1. **RM**: average number of rooms per dwelling - positive correlation.
2. **DIS**: weighted distances to five Boston employment centres - negative correlation.
3. **PTRATIO**: pupil-teacher ratio by town - negative correlation.
4. **LSTAT**: % lower status of the population - negative correlation.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
CRIM        -1.080e-01  3.286e-02  -3.287 0.001087 **
ZN           4.642e-02  1.373e-02   3.382 0.000778 ***
INDUS        2.056e-02  6.150e-02   0.334 0.738288
CHAS         2.687e+00  8.616e-01   3.118 0.001925 **
NOX         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
RM           3.810e+00  4.179e-01   9.116  < 2e-16 ***
AGE          6.922e-04  1.321e-02   0.052 0.958229
DIS         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
RAD          3.060e-01  6.635e-02   4.613 5.07e-06 ***
TAX         -1.233e-02  3.760e-03  -3.280 0.001112 **
PTRATIO     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
B            9.312e-03  2.686e-03   3.467 0.000573 ***
LSTAT       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

4.3 Correlation Visualizations



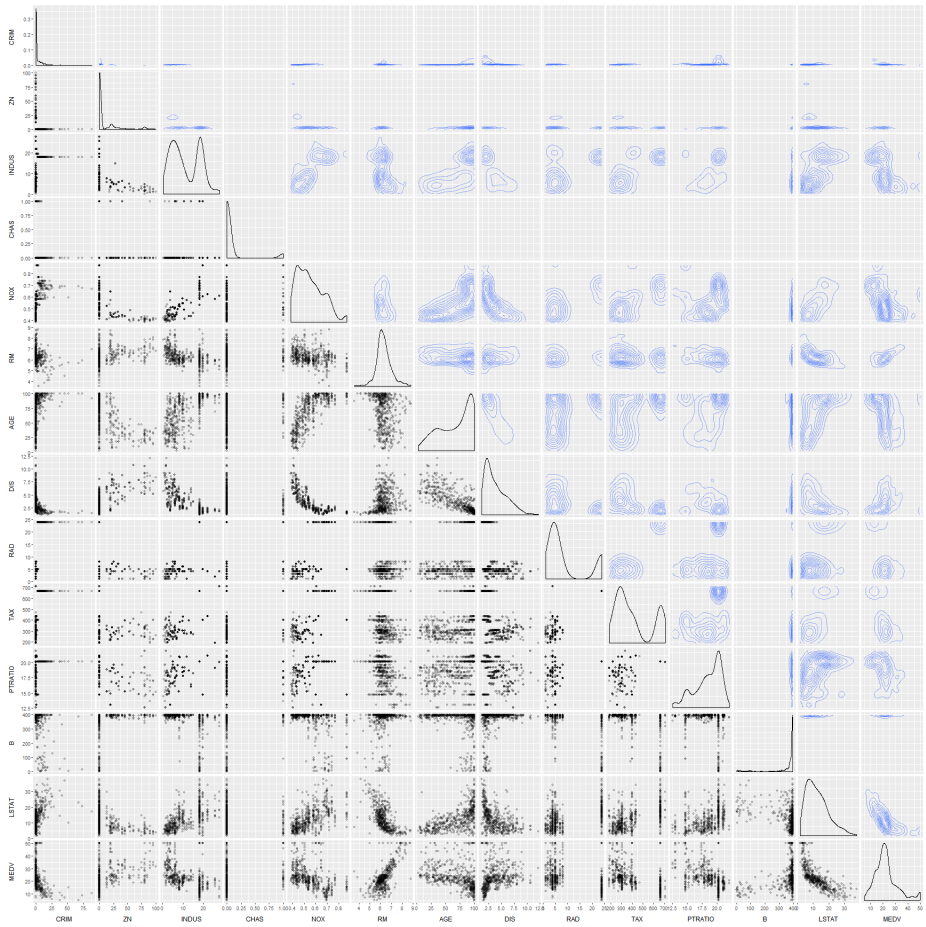
The *corrplot* above shows some interesting relations that can be further studied, such as strong positive correlation between:

- **RAD and TAX**: A pretty strong correlation between accessibility to radial highways and property taxes.
- **INDUS and NOX**: Relation between industrial acres relation to increase in nitric oxides.
- **INDUS and TAX**: Relation between business acres and property-tax.
- **NOX and TAX**: Relation between nitric oxide and taxes, most probably due to the two items above.
- **RM and MEDV**: Releation between number of rooms and median home price

The plot also shows strong negative correlations between:

- **LSTAT and MEDV**: negative correlation between percentage of lower status and median value of homes.
- **NOX and DIS**: As distance to employment centers increases, level of nitric oxides decreases.

Following plot displays more details about pairwise relations, and is generated using *ggpairs*:



Notice that ggpairs plot above also helps in finding non-linear relations which won't be captured by correlation values in corrplot