Locational and Characteristic Effects on Median Housing Value

Jeremy L. Smoljan

University of Nevada Las Vegas

IS 471-1001: Big Data

Professor Michael Lee

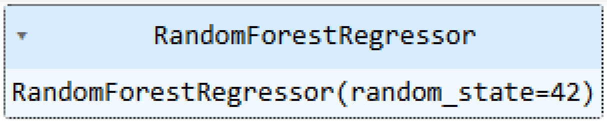
06 December 2024

Locational and Characteristic Effects on Median Housing Value

This report analyzes the effects of location on the price of occupied homes in comparison to the effects of the characteristics of the house itself. This is done through models created in python from a modified version of data published by Harrison, D. and Rubinfield, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.

Setting the Model

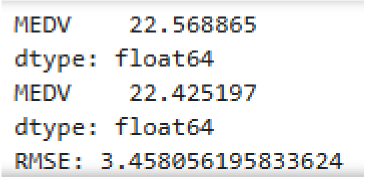
This analysis begins by setting the data into a random forest regression model. The random forest regression model assists this analysis due to the high number of independent variables affecting the dependent variable, the median housing value. It also helps due to the high correlation between the independent variables which greatly affects the overall analysis. In performing a multitude of separate tests utilizing different model analysis techniques, a random forest regression model continuously outperformed all its competitors. The training and testing was set at a test size of 25% of the total data with a random state of 42. The model itself was set to 100 estimators with a random state of 42 and the variable 'RIVER' was converted to equal 1 for 'yes' and 0 for 'no' for the specific case of determining the increase or decrease in home value if a home was situated near the river.



```
RandomForestRegressor
RandomForestRegressor(random_state=42)
```

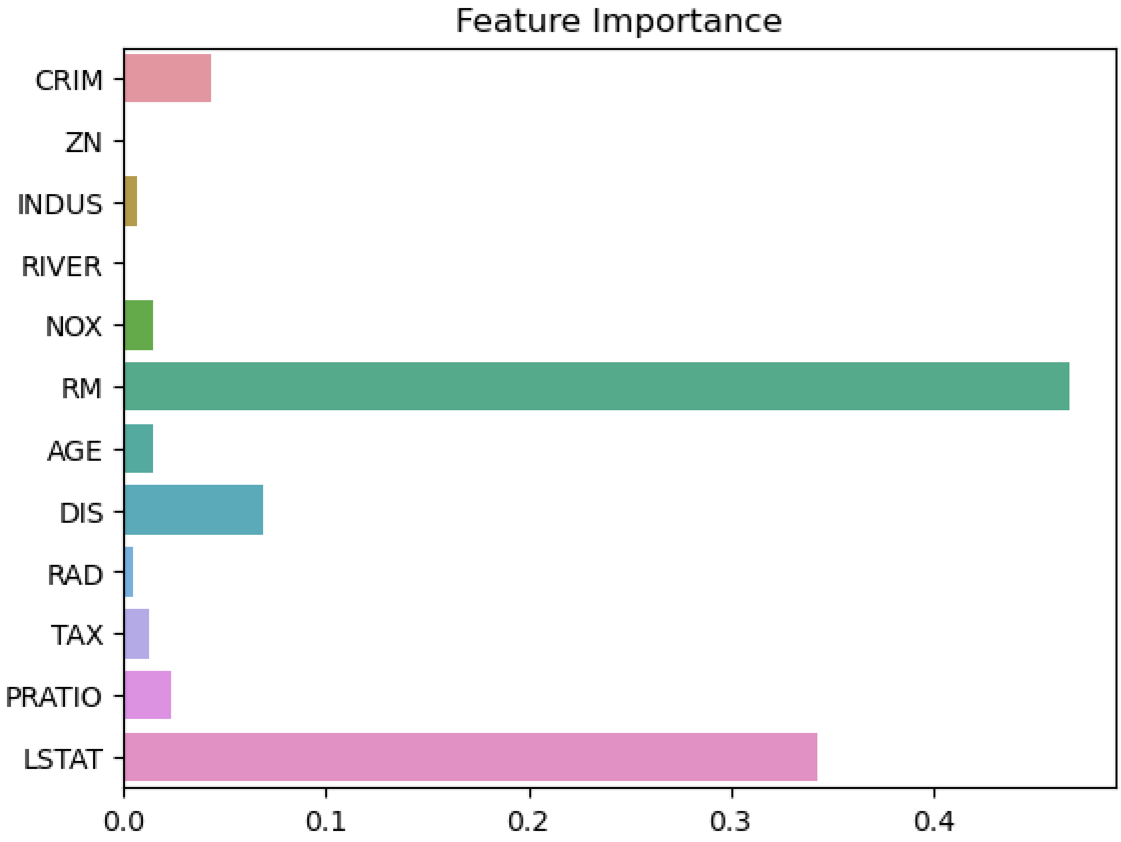
Evaluating the Model

Below are the listed mean values of 'MEDV', the dependent variable, from the training and testing models along with the RMSE, the root mean square error. These values show that at a mean value of $22.5k for the home price, the deviation is +-$3.45k by utilizing the current model. That totals to a 15.33% average distance to the mean values utilizing this model. Although the RMSE is relatively high, in this scenario it is only 5.33% greater than what might be acceptable for a difference in housing values for a well accepted model due to the larger ranging scale of values found in homes. Finding the RMSE and creating a model utilizing the lowest value was important because it helps in determining the accuracy of the model as a whole. In a range of tests, the RMSE of a random forest regression model outperformed all other models used by a minimum of one whole number.

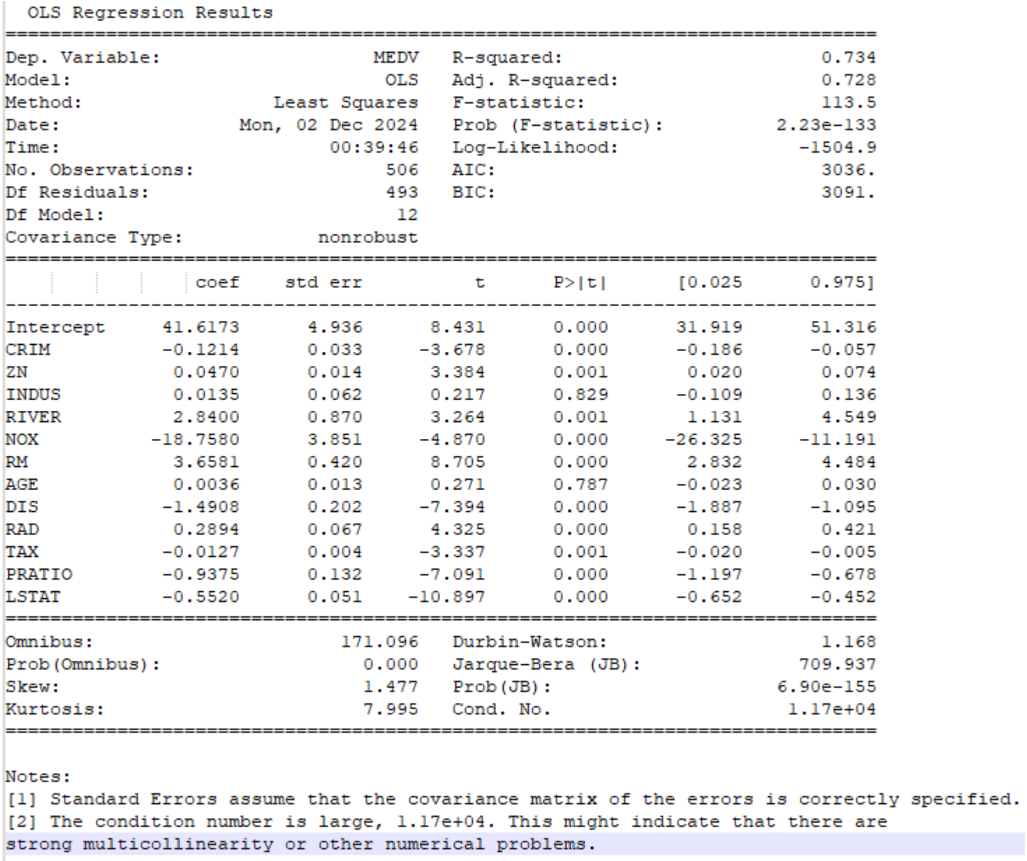


```
MEDV    22.568865
dtype: float64
MEDV    22.425197
dtype: float64
RMSE: 3.458056195833624
```

In an evaluation of the importance of each individual feature on its effect towards the median value, the number of rooms scaled highest at 0.48. This means that the number of rooms, or a characteristic of the house itself, was found to showcase the greatest impact towards the value of a home. The second greatest impacting variable was the 'LSTAT', or the percentage of the lower status of the population nearby which was rated at 0.34 and it is part of the location of the home. One major point to note is that the likelihood of a house having more rooms may be potentially related to the 'better' location around it as it is defined. The correlation between the characteristics of the house and the location around it is highly defined throughout these analyses. As each iteration of the model was completed with missing variables, such as removing the 'RM' variable or the 'LSTAT' variable, a large negative shift in the RMSE and adjusted R-Square was found, averaging a -0.1 decrease for each removed independent variable. At the same time, the condition number was consistently rated higher than 1.17e+04, which represents that there is a high likelihood of multicollinearity between the variables.

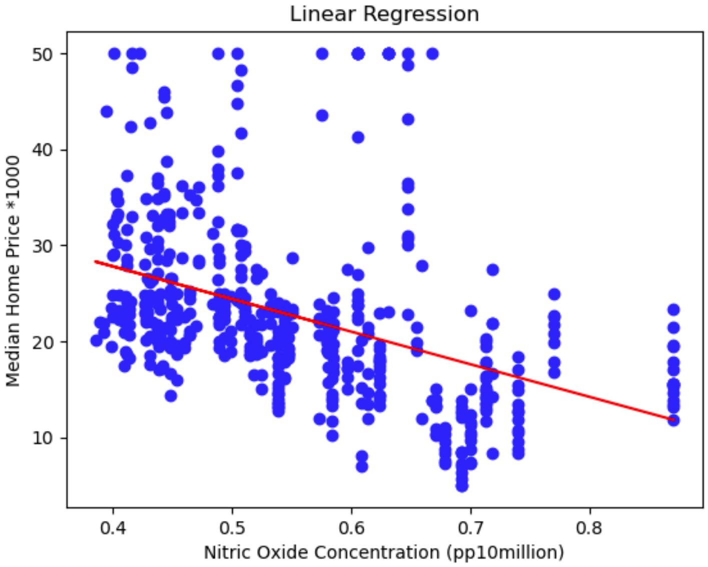


Continuing with the analysis of the effects of location and characteristics on the median value of a home, an OLS regression was performed to further determine the accuracy of the data in explaining the change in prices. The adjusted r-squared value was found to be 0.728, showing that this data has an accuracy rating of 72.8% in determining a home's median value. Furthermore, the probability(F-statistic) is 2.23e-133 which is exceptionally below a p-value of 0.05. As the probability(F-Statistic) shows how likely the data does not correlate to the dependent variable, it being significantly below 0.05 means the model is significant and the independent variables likely do affect the dependent variable at the current predicted accuracy. In further study of the regression model, we can see that the 'INDUS' and 'AGE' variables had a p-value greater than 0.05, showing that they are not likely to be correlated to determining the median value of a home. Also, in this regression analysis, we find that 'NOX' or the total concentration of nitric oxide for an area had the largest coefficient value on home prices at -18.758 plus/minus 3.85. This could potentially be explained by the idea that nitric oxide concentrations are likely to be extremely high in specific industrial areas where property value is generally lower than the average and homes are less likely to be built. Meaning that the majority of home values are unaffected by 'NOX' due to the high quantity of homes being built in areas unaffected by large concentrations of nitric oxide, however, those that are built in areas with high 'NOX' do see a large decrease in median home value as seen below in the linear regression plot. So, even though 'NOX' has a marginally high coefficient value and impact on price, its overall impact in determining home values is quite low due to the smaller amount of homes it can impact.

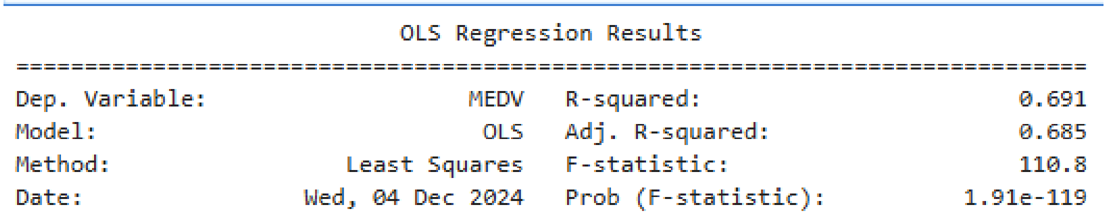


```
 OLS Regression Results
==============================================================================
Dep. Variable:                   MEDV   R-squared:                       0.734
Model:                            OLS   Adj. R-squared:                  0.728
Method:                 Least Squares   F-statistic:                     113.5
Date:                Mon, 02 Dec 2024   Prob (F-statistic):          2.23e-133
Time:                        00:39:46   Log-Likelihood:                -1504.9
No. Observations:                 506   AIC:                             3036.
Df Residuals:                     493   BIC:                             3091.
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     41.6173      4.936      8.431      0.000      31.919      51.316
CRIM          -0.1214      0.033     -3.678      0.000      -0.186      -0.057
ZN             0.0470      0.014      3.384      0.001       0.020       0.074
INDUS          0.0135      0.062      0.217      0.829      -0.109       0.136
RIVER          2.8400      0.870      3.264      0.001       1.131       4.549
NOX          -18.7580      3.851     -4.870      0.000     -26.325     -11.191
RM             3.6581      0.420      8.705      0.000       2.832       4.484
AGE            0.0036      0.013      0.271      0.787      -0.023       0.030
DIS           -1.4908      0.202     -7.394      0.000      -1.887      -1.095
RAD            0.2894      0.067      4.325      0.000       0.158       0.421
TAX           -0.0127      0.004     -3.337      0.001      -0.020      -0.005
PRATIO        -0.9375      0.132     -7.091      0.000      -1.197      -0.678
LSTAT         -0.5520      0.051    -10.897      0.000      -0.652      -0.452
==============================================================================
Omnibus:                      171.096   Durbin-Watson:                   1.168
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              709.937
Skew:                           1.477   Prob(JB):                    6.90e-155
Kurtosis:                       7.995   Cond. No.                     1.17e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.17e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

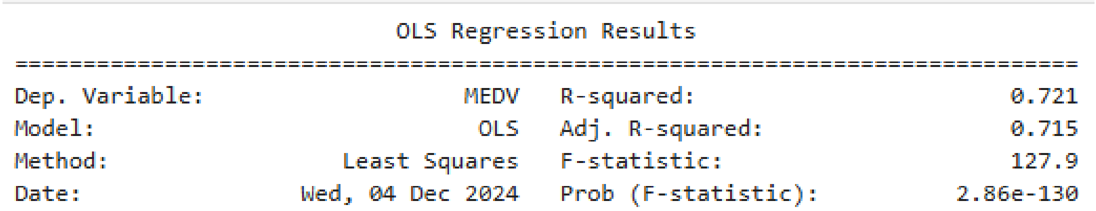


Differential Analysis. In the example below, two separate OLS regression analyses were performed both without ‘AGE’ and the first without the variable ‘RM’ and one without the variable ‘NOX’. This was done to determine the overall impact these two variables had in determining overall changes on a home’s median price. As seen below, the adjusted r-squared value dropped from 0.728 to 0.685, creating a gap of 4.3% in unexplained value to the change in median price in a regression done without ‘RM’. In comparison, a regression without ‘NOX’ only dropped to an adjusted r-squared value of 0.715 which showcases it’s statistically less impactful to the median value as a whole as considered in the previous paragraph. In both examples, a standard test was performed using the ‘AGE’ independent variable and it was found that it created no difference within the listed values. This shows that the ‘AGE’ value was not applicable to help explain any changes found within the dependent variable, the median house value.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   MEDV   R-squared:                       0.691
Model:                            OLS   Adj. R-squared:                  0.685
Method:                 Least Squares   F-statistic:                     110.8
Date:                Wed, 04 Dec 2024   Prob (F-statistic):          1.91e-119
```

(Without ‘RM’)



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   MEDV   R-squared:                       0.721
Model:                            OLS   Adj. R-squared:                  0.715
Method:                 Least Squares   F-statistic:                     127.9
Date:                Wed, 04 Dec 2024   Prob (F-statistic):          2.86e-130
```

(Without ‘NOX’)

Conclusion

The original aim of this analysis as written above was to determine the effects of location in comparison to characteristics of a home on its median value. What I found was the price of a home is likely to be largely affected by the location of the home which would include a higher build quality in comparison to the characteristics alone such as the number of bedrooms or the age of the house or even not all. Due to the high multicollinearity of the independent variables, it should be considered that the number of bedrooms is likely to be higher in areas that are deemed to be a better location in comparison to being independently related to the median value of a home. In a multitude of tests, the accuracy of the prediction of the median value decreased significantly when values were tested alone in comparison to as a whole. This was found to be true through both a lower adjusted r-squared value and a higher RMSE in independent tests. The final result is that the price of the home can be better determined by utilizing both the location and characteristics of the home with a forest regression model error of 15.33% in price prediction and an OLS regression accuracy of 72.8%.