**Boston Housing Data Analysis Report**

**보스턴 지역 집값 데이터 분석 보고서**

**Amirhossein Goudarzi**

**아미르**

**Executive Summary**

Intro Summary

Boston housing data included 14 sets of attributes each having 506 data. Main attribute which is the main point of our interest is the MEDV (median value of owner occupied homes in $1000). The main purpose of this paper is proposing the best model which predicts the movements of MEDV in different areas of Boston.

Result

Conclusion

**Data Exploration (데이터 속성 탐색)**

Boston housing data included 14 sets of attributes each having 506 data. Main attribute which is the main point of our interest is the MEDV (median value of owner occupied homes in $1000). The main purpose of this paper is proposing the best model which predicts the movements of MEDV in different areas of Boston.

Initially I analyzed the MEDV and then compared it with other 13 attributes to find the most useful attributes and analyze their effects on MEDV. As it is visible in the histogram below, MEDV has a median of $21,000 but it is scattered and there are large frequencies happening in higher values. MEDV has almost matching mean and median but it has a standard deviation of $9,000 which is comparably large. Whicker chart of the MEDV also shows a large deviation between the median, MIN and MAX values.

I have also compared the whicker chart of different attributes to find more about the behavior of each attribute regardless of the MEDV[[1]](#footnote-1). Among all of the attributes, TAX has one of the widest distribution while B, CRIM, ZN, AGE, DIS and LSTAT show a much deviated result from the median. A correlation factor can show a relation between these irregularities and the changes in MEDV.

Correlation chart shows a high correlation with MEDV in almost all of the attributes except DIS and CHAS which have lower correlation comparing to other attributes. This means MEDV is less correlated with being located near the Charles River comparing to other factors. Moreover, when checking near zero variance to find indifferent attributes to the changes of MEDV, we find that all of the attributes are non-zero variance therefore we cannot eliminate them through this process.

Also correlation factors between other attributes[[2]](#footnote-2) shows that while INDUS has a low correlation with MEDV but there is a high correlation between INDUS and other attributes such as NOX, AGE, DIS, TAX and LSTAT. This shows that some attributes may indirectly affect MEDV while they don’t show up in the correlation analysis. AGE and DIS, RAD and TAX are also highly correlated.

|  |  |
| --- | --- |
| **Attribute** | **Correlation** |
| **RM** | 0.695 |
| **ZN** | 0.36 |
| **B** | 0.333 |
| **DIS** | 0.25 |
| **CHAS** | 0.175 |
| **AGE** | -0.377 |
| **RAD** | -0.382 |
| **CRIM** | -0.388 |
| **NOX** | -0.427 |
| **TAX** | -0.469 |
| **INDUS** | -0.484 |
| **PTRATIO** | -0.508 |
| **LSTAT** | -0.738 |

Running a simple regression analysis[[3]](#footnote-3), we can see that there is a high T value for the CRIM, ZEN, INDUS, CHAS, AGE, TAX and B while the P value of INDUS and AGE are considerably high. These values and relations will be effected in the modeling stage and it will help creating a more relevant model with higher R square.

**Modelling**

Training step by step

**Performance Analysis**

K fold test

**Conclusion**

Applicability

What are the limitations and problems of this model

What should be added

**Appendices**

**Appendix I** <Whisker Chart Comparison All In One>

**Appendix II** <Whisker Chart Comparison of all Attributes>

**Appendix III** <Attributes Comparison Data>

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **CRIM** | **ZN** | **INDUS** | **CHAS[[4]](#footnote-4)** | **NOX** | **RM** | **AGE** | **DIS** | **RAD** | **TAX** | **PTRATIO** | **B** | **LSTAT** | **MEDV** |
| **MIN** | 0.006 | 0.000 | 0.460 | 0 | 0.385 | 3.561 | 2.900 | 1.130 | 1.000 | 187.000 | 12.600 | 0.320 | 1.730 | 5.000 |
| **First Quartile** | 0.082 | 0.000 | 5.190 | 0 | 0.449 | 5.886 | 45.025 | 2.100 | 4.000 | 279.000 | 17.400 | 375.378 | 6.950 | 17.025 |
| **Median** | 0.257 | 0.000 | 9.690 | 0 | 0.538 | 6.209 | 77.500 | 3.207 | 5.000 | 330.000 | 19.050 | 391.440 | 11.360 | 21.200 |
| **Third Quartile** | 3.677 | 12.500 | 18.100 | 0 | 0.624 | 6.624 | 94.075 | 5.188 | 24.000 | 666.000 | 20.200 | 396.225 | 16.955 | 25.000 |
| **Max** | 88.976 | 100.000 | 27.740 | 1 | 0.871 | 8.780 | 100.000 | 12.127 | 24.000 | 711.000 | 22.000 | 396.900 | 37.970 | 50.000 |
| **STD** | 8.602 | 23.322 | 6.860 | 0.254 | 0.116 | 0.703 | 28.149 | 2.106 | 8.707 | 168.537 | 2.165 | 91.295 | 7.141 | 9.197 |
| **Average** | 3.614 | 11.364 | 11.137 | 0.069 | 0.555 | 6.285 | 68.575 | 3.795 | 9.549 | 408.237 | 18.456 | 356.674 | 12.653 | 22.533 |

**Appendix IV** <Correlation Factor between Attributes>

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **CRIM** | **ZN** | **INDUS** | **CHAS** | **NOX** | **RM** | **AGE** | **DIS** | **RAD** | **TAX** | **PTRATIO** | **B** | **LSTAT** | **MEDV** |
| **CRIM** | 1.000 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **ZN** | -0.200 | 1.000 |  |  |  |  |  |  |  |  |  |  |  |  |
| **INDUS** | 0.407 | -0.534 | 1.000 |  |  |  |  |  |  |  |  |  |  |  |
| **CHAS** | -0.056 | -0.043 | 0.063 | 1.000 |  |  |  |  |  |  |  |  |  |  |
| **NOX** | 0.421 | -0.517 | 0.764 | 0.091 | 1.000 |  |  |  |  |  |  |  |  |  |
| **RM** | -0.219 | 0.312 | -0.392 | 0.091 | -0.302 | 1.000 |  |  |  |  |  |  |  |  |
| **AGE** | 0.353 | -0.570 | 0.645 | 0.087 | 0.731 | -0.240 | 1.000 |  |  |  |  |  |  |  |
| **DIS** | -0.380 | 0.664 | -0.708 | -0.099 | -0.769 | 0.205 | -0.748 | 1.000 |  |  |  |  |  |  |
| **RAD** | 0.626 | -0.312 | 0.595 | -0.007 | 0.611 | -0.210 | 0.456 | -0.495 | 1.000 |  |  |  |  |  |
| **TAX** | 0.583 | -0.315 | 0.721 | -0.036 | 0.668 | -0.292 | 0.506 | -0.534 | 0.910 | 1.000 |  |  |  |  |
| **PTRATIO** | 0.290 | -0.392 | 0.383 | -0.122 | 0.189 | -0.356 | 0.262 | -0.232 | 0.465 | 0.461 | 1.000 |  |  |  |
| **B** | -0.385 | 0.176 | -0.357 | 0.049 | -0.380 | 0.128 | -0.274 | 0.292 | -0.444 | -0.442 | -0.177 | 1.000 |  |  |
| **LSTAT** | 0.456 | -0.413 | 0.604 | -0.054 | 0.591 | -0.614 | 0.602 | -0.497 | 0.489 | 0.544 | 0.374 | -0.366 | 1.000 |  |
| **MEDV** | -0.388 | 0.360 | -0.484 | 0.175 | -0.427 | 0.695 | -0.377 | 0.250 | -0.382 | -0.469 | -0.508 | 0.333 | -0.738 | 1.000 |

**Appendix V** <Regression Analysis of All Attributes>

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Regression Statistics*** | |  |  |  |  |  |  |  |
| **Multiple R** | 0.860605987 |  |  |  |  |  |  |  |
| **R Square** | 0.740642664 |  |  |  |  |  |  |  |
| **Adjusted R Square** | 0.733789726 |  |  |  |  |  |  |  |
| **Standard Error** | 4.745298182 |  |  |  |  |  |  |  |
| **Observations** | 506 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | ***df*** | ***SS*** | ***MS*** | ***F*** | ***Significance F*** |  |  |  |
| **Regression** | 13 | 31637.51084 | 2433.65468 | 108.0766662 | 6.7222E-135 |  |  |  |
| **Residual** | 492 | 11078.78458 | 22.51785483 |  |  |  |  |  |
| **Total** | 505 | 42716.29542 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | ***Coefficients*** | ***Standard Error*** | ***t Stat*** | ***P-value*** | ***Lower 95%*** | ***Upper 95%*** | ***Lower 95.0%*** | ***Upper 95.0%*** |
| **Intercept** | 36.45948839 | 5.103458811 | 7.144074193 | 3.28344E-12 | 26.43222601 | 46.48675076 | 26.43222601 | 46.48675076 |
| **CRIM** | -0.108011358 | 0.032864994 | -3.286516871 | 0.00108681 | -0.172584412 | -0.043438304 | -0.172584412 | -0.043438304 |
| **ZN** | 0.046420458 | 0.013727462 | 3.381576282 | 0.00077811 | 0.019448778 | 0.073392139 | 0.019448778 | 0.073392139 |
| **INDUS** | 0.020558626 | 0.061495689 | 0.334310042 | 0.738288071 | -0.100267941 | 0.141385193 | -0.100267941 | 0.141385193 |
| **CHAS** | 2.686733819 | 0.861579756 | 3.118380858 | 0.00192503 | 0.993904193 | 4.379563446 | 0.993904193 | 4.379563446 |
| **NOX** | -17.76661123 | 3.819743707 | -4.651257411 | 4.24564E-06 | -25.27163356 | -10.26158889 | -25.27163356 | -10.26158889 |
| **RM** | 3.809865207 | 0.417925254 | 9.1161402 | 1.97944E-18 | 2.988726773 | 4.63100364 | 2.988726773 | 4.63100364 |
| **AGE** | 0.000692225 | 0.013209782 | 0.052402427 | 0.958229309 | -0.02526232 | 0.026646769 | -0.02526232 | 0.026646769 |
| **DIS** | -1.475566846 | 0.199454735 | -7.398003603 | 6.01349E-13 | -1.867454981 | -1.08367871 | -1.867454981 | -1.08367871 |
| **RAD** | 0.306049479 | 0.06634644 | 4.612899768 | 5.07053E-06 | 0.175692169 | 0.436406789 | 0.175692169 | 0.436406789 |
| **TAX** | -0.012334594 | 0.003760536 | -3.28000914 | 0.001111637 | -0.019723286 | -0.004945902 | -0.019723286 | -0.004945902 |
| **PTRATIO** | -0.952747232 | 0.130826756 | -7.282510564 | 1.30884E-12 | -1.209795296 | -0.695699168 | -1.209795296 | -0.695699168 |
| **B** | 0.009311683 | 0.002685965 | 3.466792558 | 0.000572859 | 0.004034306 | 0.01458906 | 0.004034306 | 0.01458906 |
| **LSTAT** | -0.524758378 | 0.050715278 | -10.3471458 | 7.77691E-23 | -0.624403622 | -0.425113133 | -0.624403622 | -0.425113133 |

**References**

1. Jhlkj
2. Hjjgi
3. <https://www.r-bloggers.com/near-zero-variance-predictors-should-we-remove-them/>

1. Appendix I,II,III [↑](#footnote-ref-1)
2. Appendix IV [↑](#footnote-ref-2)
3. Appendix V [↑](#footnote-ref-3)
4. CHAS has a binary value, therefore it was not suitable for a histogram demonstration. But its effects will be reviewed in the modelling part. [↑](#footnote-ref-4)