Big Data Machine Learning WoW Portfolio Piece: Boston House Prices Regression

Group 16: Yiqun Hu, Junru He

BLUF

- Business Problem

This paper investigates several potential factors that may contribute to the value of owner-occupied homes. Purchasing a home is still one of the most significant expenditures that people make. Our study is to predict house selling values in Boston using a variety of residential property factors. Individuals may be able to use the information gathered to assist them in making better selections when purchasing a home. This allows consumers to make the most of their money while staying within their financial constraints. The findings could be used by a real estate agent to increase the likelihood of a sale by correctly marketing key features. We'd focus on descriptive and predictive analytics, as well as suggestions for obtaining more data so that additional tests could be run to optimize the purchase.

- Dataset

The dataset for this project was provided by Kaggle was come up with 14 variables in 506 records. The attribute information input features are as follows: 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) [parts/10M]; 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 centers; 9) RAD: index of accessibility to radial highways; 10) TAX: full-value property-tax rate per 10,000[/10k]; 11) PTRATIO: pupil-teacher ratio by town; 12) B: The result of the equation B=1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town; 13) LSTAT: % lower status of the population.

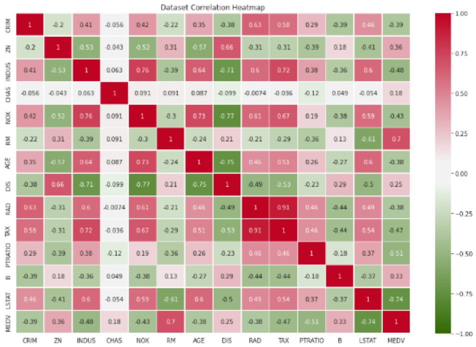
The output variable is MEDV: Median value of owner-occupied homes in 1000′s[k].

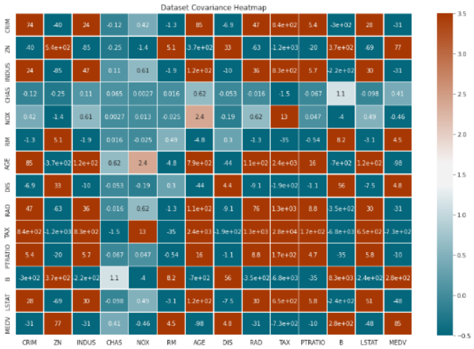
- Final Recommendation

  Using two regressors (the Random Forest Regressor and XGB Regressor), we find lower status of the population (LSTAT) is a significant driver of real estate value in Boston. Number of rooms and Crime Rate are also important, with price decreasing, respectively increasing with Nitric Oxides Concentration, respectively CHAS. The results of the two models are very similar, XGB regressor has a better performance. But there are some discrepancies in the explain ability analysis of them.

Analysis

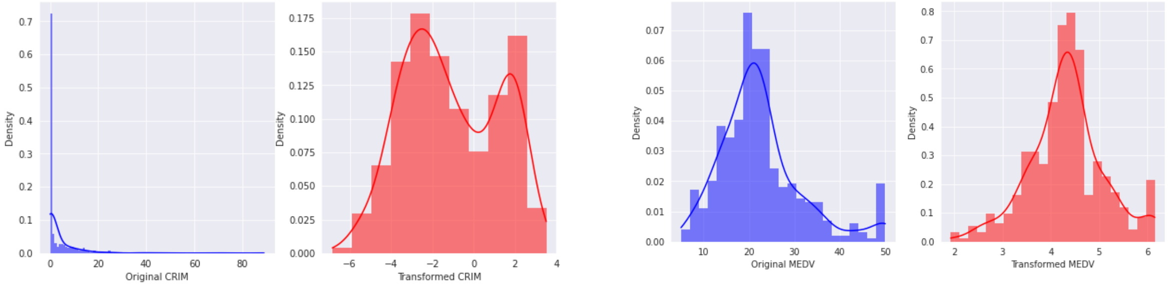
- EDA
  Firstly, we run the EDA to check the shape and summary of the dataset. Then check the skewness and kurtosis. We find out that CRIM has both high skewness and kurtosis, further skewness correction needed in the data preprocessing stage. Correlation and covariance matrix can tell the relationship between each variable.



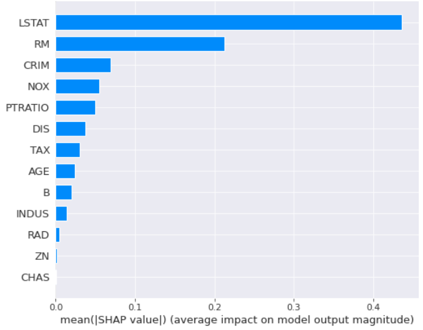


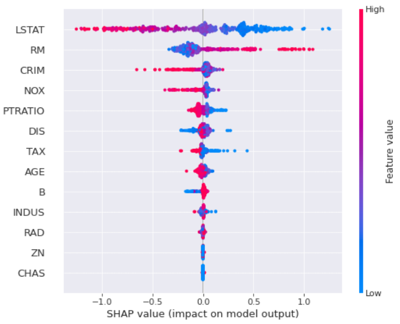
There is no missing value and duplicates in this dataset, so imputation or drop data are not required.

- Data Preprocess
Based on the EDA results, all variables in this Boston house price dataset are numerical variables, thus, dummy encoding is also no need to implement. Then, we use the Skewness Auto Transform function to do the skewness correction for all the variables. Here is an example variable CRIM contrast before and after skewness correction (CRIM skewness: 5.22). Meanwhile, Tukey rule is used to deal with outliers.

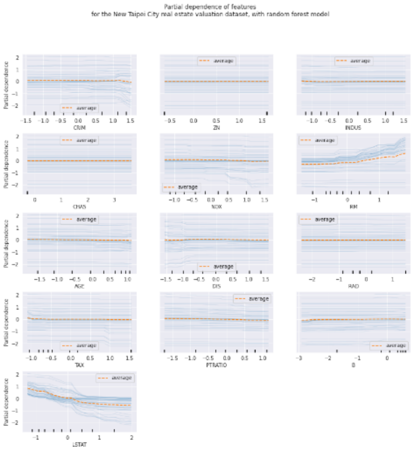


- Regression
The next step is scaling & holdout out the sample splitting the dataset into a training set and testing set on an 80-20 proportion. The shape of X and y are (506, 13), (506, 1); the shape of training set and test set are (404, 14), (102, 14) respectively. After split the dataset, we do the standardization, Grid Search a Random Forest, and fit Random Forest Regressor on holdout sample. Find out that LSTAT is the most important variable.



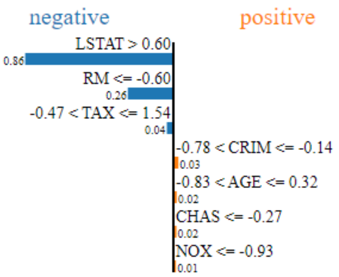


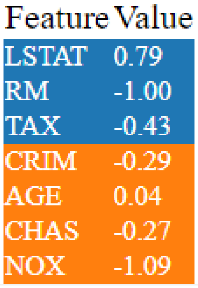
- Random Forest



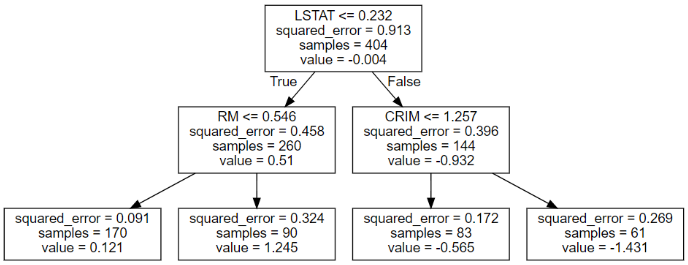
We plot the Partial Dependence, PDP & ICE plots with Random Forest Regressor Model. From the plots, there are no significant change in the partial dependence plot.

Additionally, home values appear to generally decline in value with LSTAT and increase above a standardized latitude of roughly -1. Home values appear to generally increase in value with RM and increase above a standardized latitude of roughly 1. The latitude story is not compatible with the partial dependence plot for ZN, CHAS, RAD and PTTRATIO. The ICE plots do not show much separation between the different instances.



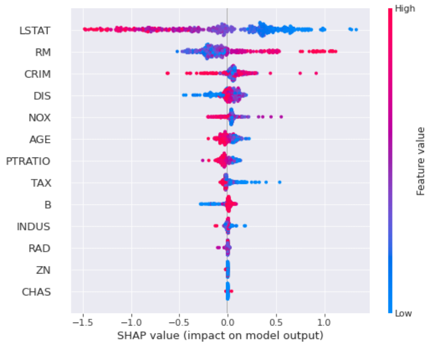


Feature	Value
LSTAT	0.79
RM	-1.00
TAX	-0.43
CRIM	-0.29
AGE	0.04
CHAS	-0.27
NOX	-1.09



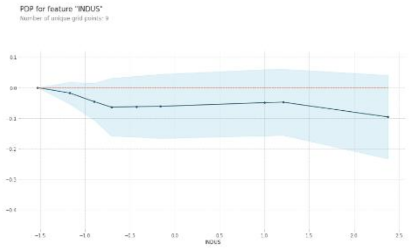
The LIME indicates LSTAT, RM and TAX have negative relationship with house price, and rest of them are positive.

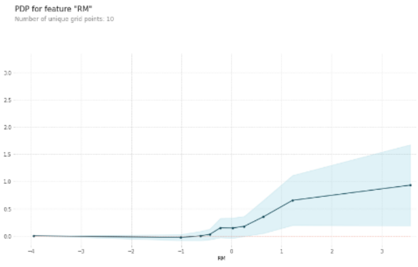
- XGB

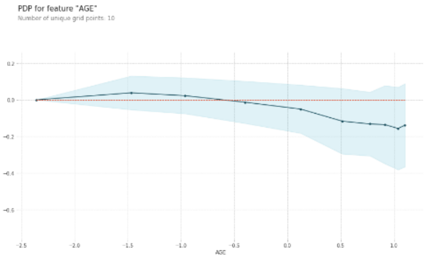


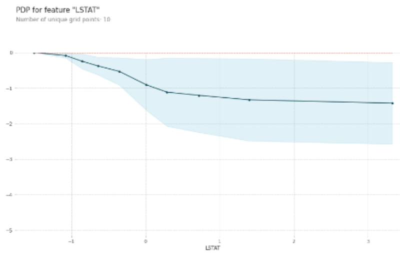
In XGB Regressor, LSTAT and RM also are two of the most important features.

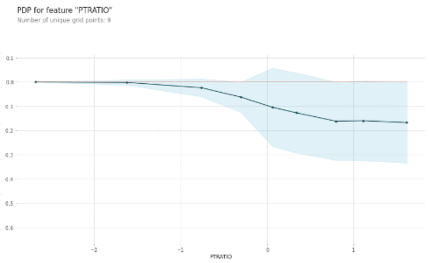
We plot the Partial Dependence plots, there are significant change in RM between standardized values of roughly -0.5 and 1. Additionally, home values appear to generally decline in value with INDUS, NOX, AGE, TAX and LSTAT.

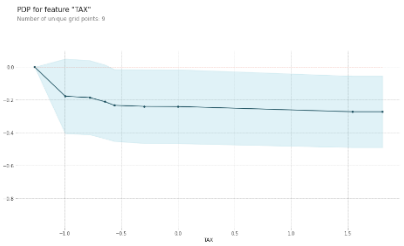




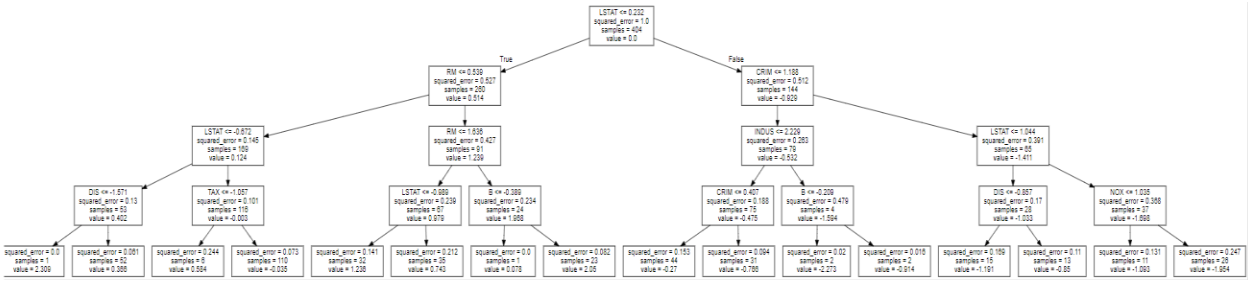




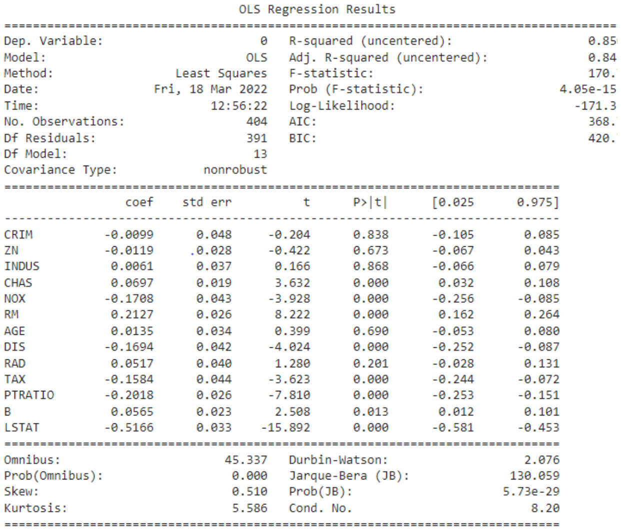




Then we plot the decision tree surrogate models for XGB Regressor.

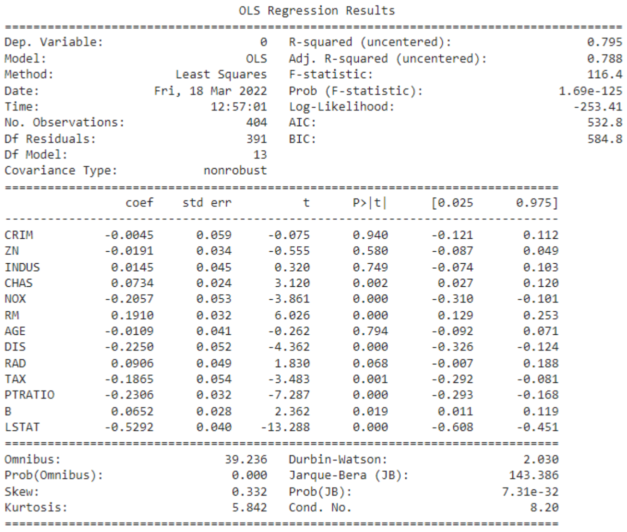


Conclusion



```
                            OLS Regression Results
=================================================================================
Dep. Variable:                      0   R-squared (uncentered):              0.85
Model:                            OLS   Adj. R-squared (uncentered):         0.84
Method:                 Least Squares   F-statistic:                         170.
Date:                Fri, 18 Mar 2022   Prob (F-statistic):              4.05e-15
Time:                        12:56:22   Log-Likelihood:                    -171.3
No. Observations:                 404   AIC:                                 368.
Df Residuals:                     391   BIC:                                 420.
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
CRIM          -0.0099      0.048     -0.204      0.838      -0.105       0.085
ZN            -0.0119     .0.028     -0.422      0.673      -0.067       0.043
INDUS          0.0061      0.037      0.166      0.868      -0.066       0.079
CHAS           0.0697      0.019      3.632      0.000       0.032       0.108
NOX           -0.1708      0.043     -3.928      0.000      -0.256      -0.085
RM             0.2127      0.026      8.222      0.000       0.162       0.264
AGE            0.0135      0.034      0.399      0.690      -0.053       0.080
DIS           -0.1694      0.042     -4.024      0.000      -0.252      -0.087
RAD            0.0517      0.040      1.280      0.201      -0.028       0.131
TAX           -0.1584      0.044     -3.623      0.000      -0.244      -0.072
PTRATIO       -0.2018      0.026     -7.810      0.000      -0.253      -0.151
B              0.0565      0.023      2.508      0.013       0.012       0.101
LSTAT         -0.5166      0.033    -15.892      0.000      -0.581      -0.453
==============================================================================
Omnibus:                       45.337   Durbin-Watson:                   2.076
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              130.059
Skew:                           0.510   Prob(JB):                     5.73e-29
Kurtosis:                       5.586   Cond. No.                         8.20
==============================================================================
```

Table 2 Random Forest Output



```
                            OLS Regression Results
=======================================================================================
Dep. Variable:                      0   R-squared (uncentered):                   0.795
Model:                            OLS   Adj. R-squared (uncentered):              0.788
Method:                 Least Squares   F-statistic:                              116.4
Date:                Fri, 18 Mar 2022   Prob (F-statistic):                   1.69e-125
Time:                        12:57:01   Log-Likelihood:                         -253.41
No. Observations:                 404   AIC:                                      532.8
Df Residuals:                     391   BIC:                                      584.8
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
CRIM          -0.0045      0.059     -0.075      0.940      -0.121       0.112
ZN            -0.0191      0.034     -0.555      0.580      -0.087       0.049
INDUS          0.0145      0.045      0.320      0.749      -0.074       0.103
CHAS           0.0734      0.024      3.120      0.002       0.027       0.120
NOX           -0.2057      0.053     -3.861      0.000      -0.310      -0.101
RM             0.1910      0.032      6.026      0.000       0.129       0.253
AGE           -0.0109      0.041     -0.262      0.794      -0.092       0.071
DIS           -0.2250      0.052     -4.362      0.000      -0.326      -0.124
RAD            0.0906      0.049      1.830      0.068      -0.007       0.188
TAX           -0.1865      0.054     -3.483      0.001      -0.292      -0.081
PTRATIO       -0.2306      0.032     -7.287      0.000      -0.293      -0.168
B              0.0652      0.028      2.362      0.019       0.011       0.119
LSTAT         -0.5292      0.040    -13.288      0.000      -0.608      -0.451
==============================================================================
Omnibus:                       39.236   Durbin-Watson:                   2.030
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              143.386
Skew:                           0.332   Prob(JB):                     7.31e-32
Kurtosis:                       5.842   Cond. No.                         8.20
==============================================================================
```

Table 1 XGB Regressor Output

Using two regressors (the Random Forest Regressor and XGB Regressor), we find lower status of the population (LSTAT) is a significant driver of real estate value in Boston. Number of rooms and Crime Rate are also important, with price decreasing, respectively increasing with Nitric Oxides Concentration, respectively CHAS.

Overall, the results of the two models are very similar, XGB regressor has a better performance. But there are some discrepancies in the explain ability analysis of them.

References

- *Harrison, David & Rubinfeld, Daniel. (1978). Hedonic housing prices and the demand for clean air. Journal of Environmental Economics and Management. 5. 81-102. 10.1016/0095-0696(78)90006-2. LINK*
- *Belsley, David A. & Kuh, Edwin. & Welsch, Roy E. (1980). Regression diagnostics: identifying influential data and sources of collinearity. New York: Wiley LINK*
- *The Boston house-price data of Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.*