Texas A&M University

ISEN-613 / Engineering Data Analysis

Boston Data Analysis Report

To: President X

From: Data Consultant (Keyla Gonzalez)

Subject: Analysis of Boston Suburban Housing Values, Executive Summary

Executive Summary

Over the last few years, data science has become one of the major tools for housing market predictions. The use of different regression algorithms has enhanced the forecasting of house prices by using particular sets of attributes. The main objective of this analysis is to obtain predictions of Boston suburban housing values based on structural properties, accessibility, and neighborhood features of the house.

Data analysis was performed to display key characteristics of the data, such as:

- A mean crime rate of ~4 with a high crime rate (90th percentile) of ~11 and a low (10th percentile) of 0.04.
- A number of houses nearby the Charles River of 35 out of 506.
- An average number of rooms per house of ~6 with a maximum value of ~9 and a minimum of ~4.
- An average proportion of houses over 25,000 sq.ft of ~11 with a minimum proportion of 0 and a maximum of 100.
- An average proportion of houses built prior to 1940 (age) of ~69 with a high age (90th percentile) of ~99 and low age of ~ 27 (10th percentile).
- A mean distance to major employment centers of ~3.8 with a maximum value of ~1 and a minimum of ~12
- And average accessibility to radial highways of around 9.6 with a maximum value of ~24 and a minimum of ~1.

Moreover, we can analyze the effect and interaction of certain predictors. For instance, towns with a high crime rate display an average age of ~94, distance employment centers of ~1.7 and a zero number of houses over 25,000 sq.ft. Hence, a certain condition may help us to understand the correlation of other factors.

On the other hand, several baseline scenarios were established in order to observe the impact of predictors and response. For example, for a large “distance to employment centers” the prediction was about $21,395; however, for a high “crime rate” the house value was around $12,700; and, for a high number of rooms (~8) the value was approximately $36,286. Thus, the impact of a house price is highly conditioned by the specific attribute values.

Furthermore, each of these attributes may represent an increase or decrease in the house value. For instance, for every one-unit increase in crime rate, the house value decreases by ~1.45%; contrarily, for every one-unit increase in the number of rooms, house value increases by about ~32%.

Finally, predictions were performed on a randomize new data with the top 5 predictions on a range of $21,650 to $33,520. These predictions were mainly influenced by the number of rooms, age of the house, houses nearby the Charles River, and the proportion of houses over 25,000 sq.ft.

Technical Summary

Workflow

The workflow of this analysis was divided into seven main parts: 1) data analysis and visualization, 2) multilinear regression, 3) predictions on baseline cases, 4) analysis of percentage increase change of predictors, 5) prediction on a randomized data set, 6) the diagnostic plots of the model and 7) multicollinearity analysis.

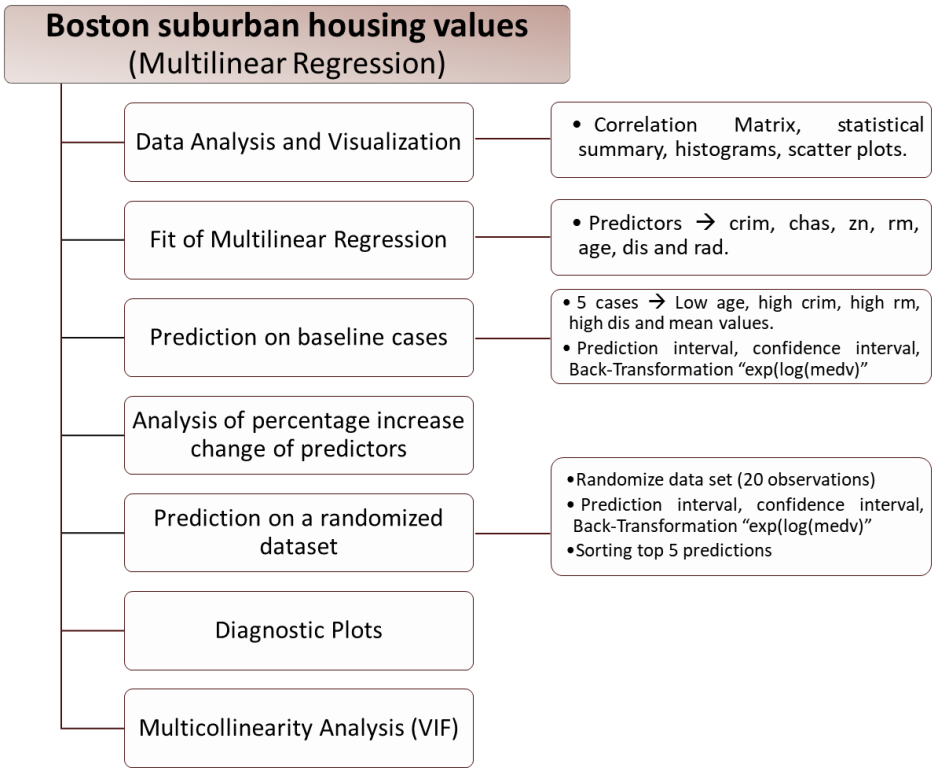

Fig. 1. Workflow of Boston suburban housing values data set

1) Data analysis and visualization

On this stage, a correlation matrix, statistical summary, histograms and scatter plots were obtained. The correlation matrix (Fig. 2) and scatter plots helped to understand the relationship between response and features, and features and features (multicollinearity). On the other hand, the histogram (Fig. 3) and statistical summary helped to understand the features/response distributions.

The attributes were picked based on three main characteristics:

1. Neighborhood
    - Crim: per capita crime rate by town.
    - Chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

2. Structural properties of a house:
    - Zn: proportion of residential land zoned for lots over 25,000 sq.ft.
    - Rm: average number of rooms per house.
    - Age: proportion of owner-occupied units built prior to 1940.
3. Accessibility.
    - Dis: weighted mean of distances to five Boston employment centers.
    - Rad: index of accessibility to radial highways.

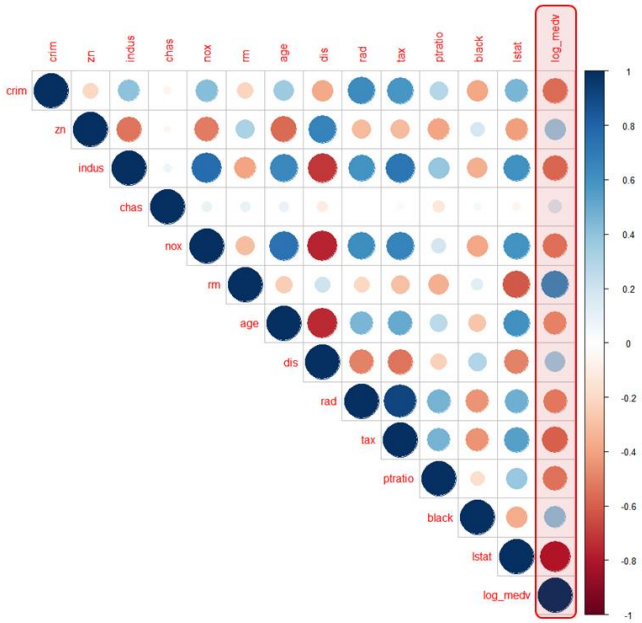

Fig. 2. Correlation matrix plot of response and predictors

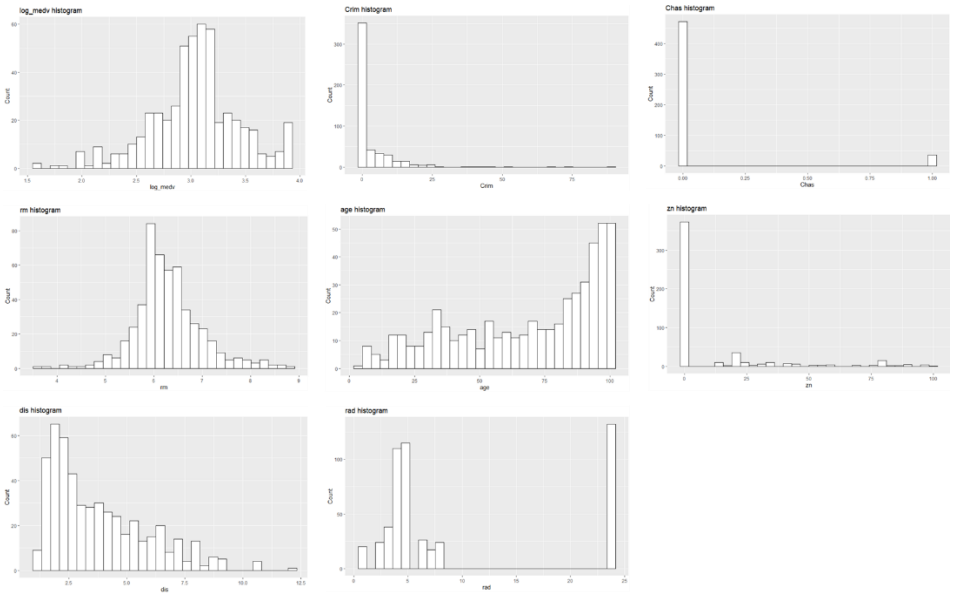

Fig. 3. Histograms of response (log(medv)) and predictors (crim, chas, zn, rm, age, dis, rad)

Finally, it is vital to mention that no data points were removed since there is no information regarding this matter; therefore, it is impossible to remove them if they are not classified as errors in the data set.

2) Fit of Multilinear Regression and 3) Predictions on baseline cases

A multilinear regression was performed on the seven features (crim, chas, zn, rm, age, dis, rad) and predictions were implemented on five baseline cases.

- Low age (10th quantile)
- High crim (90th quantile)
- High rm (90th quantile)
- High dis (90th quantile)
- Mean values

Moreover, the prediction and confidence intervals (PI and CI) were calculated to observe the range of predictions. The highest prediction was obtained for the case of a high number of rooms ($ 36,286).

			PI		CI		Quantile-based non-parametric PI (stdres)	
	Clients	Predictions	*lower*	*upper*	*lower*	*upper*	*lower (2.5%)*	*upper (97.5%)*
Low age	*1*	$ 18,462.95	$ 11,267.24	$ 30,254.15	$ 18,012.90	$ 18,924.25	$ 2,223.30	$ 160,630.71
High crim	*2*	$ 12,699.07	$ 7,731.04	$ 20,859.58	$ 12,021.88	$ 13,414.39	$ 1,529.22	$ 110,483.94
High rm	*3*	$ 36,286.14	$ 22,076.83	$ 59,640.98	$ 34,167.02	$ 38,536.69	$ 4,369.56	$ 315,695.31
High dis	*4*	$ 21,394.64	$ 13,022.23	$ 35,149.95	$ 20,218.07	$ 22,639.68	$ 2,576.33	$ 186,136.87
Mean	*5*	$ 16,229.60	$ 9,847.43	$ 26,748.10	$ 14,988.27	$ 17,573.73	$ 1,954.36	$ 141,200.16

Table. 1. Predictions on baseline cases with their respective PI and CI. Back transformation was performed to obtain the house value predictions.

4) Analysis of the percentage increase change of predictors

The analysis of the percentage increase was obtained based on the back-transformation of our coefficients and predictions. The percentage increase can be defined as:

$$(exp(coefficient) - 1) * 100$$

And it is going to allow us to interpret the effect of predictors on response. The conclusions were:

- For every one-unit increase in crim, medv decreases by about 1.45%
- For every one-unit increase in chas, medv increases by about 18%
- For every one-unit increase in rm, medv increases by about 32%
- For every one-unit increase in age, medv decreases by about 0.4%

- For every one-unit increase in zn, medv increases by about 0.15%
- For every one-unit increase in dis, medv decreases by about 3.5%
- For every one-unit increase in rad, medv decreases by about 0.7%

5) Predictions on a randomized dataset.

Predictions were determined on a randomized dataset that contains twenty observations in the dataset. Back transformation (exp(log(medv)), prediction and confidence intervals were as well estimated.

The top five predictions are listed in table 2 where we can notice a range of $21,650 to $33,520. Moreover, we can acknowledge:

- Top 1 a prediction of $33,520; with a low value of $19,250 and high value of $58,372.
- Top 2 a prediction of $31,302; with a low value of $18,180 and high value of $53,900.

Random Dataset		PI		Attributes						
Top 5	Predictions	*lower*	*upper*	Crim_test	Zn_test	Chas_test	Rm_test	Age_test	Dis_test	Rad_test
1	$ 33,521.13	$ 19,250.21	$ 58,371.63	60	71	1	9	42	1	4
2	$ 31,302.91	$ 18,179.83	$ 53,898.84	56	9	1	9	10	2	22
3	$ 28,577.30	$ 16,963.22	$ 48,143.11	28	97	0	8	90	1	3
4	$ 25,657.47	$ 15,404.38	$ 42,734.98	5	49	0	8	67	10	21
5	$ 21,648.01	$ 12,340.65	$ 37,975.01	71	82	0	9	63	3	1

Table. 2. Predictions on a randomize dataset with their respective PI and attributes. Back transformation was performed to obtain the house value predictions.

6) Diagnostic Plots

The diagnostic plots can mainly help us to assess our model based on five potential issues (non-linearity, a non-constant variance of error terms, outliers, high-leverage points, and collinearity). Based on figure 4, we can conclude that:

- Residuals (errors) vs fitted: There is linear behavior between the response and inputs since our residuals are close to the zero residuals line. However, it seems there is still a slight non-linear behavior in the model.

  On the other hand, this plot also helps us to distinguish a relative constant variance in the error terms (part of linear regression assumption); nevertheless, it seems there are some data points far from our main error terms distribution.

- Normal Q-Q (Quantile-Quantile): This plot can help us to observe if the errors are normally distributed. It seems the errors are not normally distributed at the beginning and end of the plot.

- Scale-location (unusual yi for given xi): Outlier can be detected through this diagnostic plot. The are several data points that may be classified as potential outliers; however, the absolute value is not greater than 3; and there is no clear information that those points are associated with errors in the dataset.

- Residuals vs Leverage (unusual value for xi): High leverage points are not been detected by this specific plot since all the values are far from Cook's distance line; hence, there are not unusual values for xi.

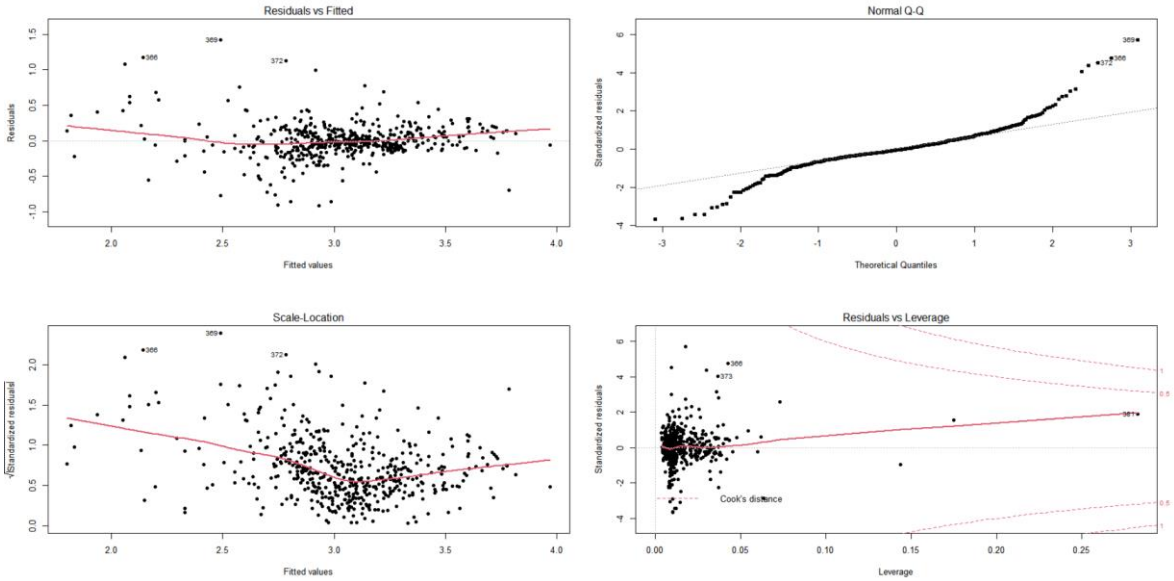

Fig. 4. Diagnostic Plots

7) Multicollinearity analysis

A multicollinearity analysis was carry out by applying the "Variance inflation factor" (VIF) and using the scatter plots and correlation matrix already generated.

The result of all of the predictors was for a VIF<5; thus, there is no collinearity in the predictor of the model.

- Crim: VIF=1.7, Chas: VIF=1, Rm: VIF= 1.17, Age: VIF=2.4, Zn: VIF=1.96, Dis: VIF=3.07, Rad: VIF=1.9.