Capstone 2 Project Report

Prince Zogli, PhD.

9/29/21

Title: Predicting housing prices based on Boston housing prices.

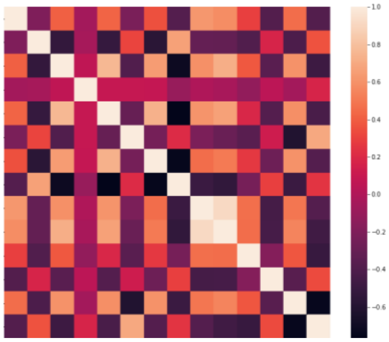
Mentors

Eleanor Thomas

Reza Farahani

Introduction:

Problem statement



Modern day Housing prices are known to be driven by factors such as Population growth, demographics, interest rates, government policies and the economy indicators (interest rate, tax, unemployment rate, income, etc.), home size and location[1].

However, it is not clear which of these factors have been historically, main determinants of housing prices.

Though historical reports suggest that, population growth, demographic change, employment rate, tax, income, commute and accessibility to commercial facilities were identified as influencing factors for housing price [2], it is not clear which amongst these were critical determinants. Therefore, the report will attempt to use a classic Boston housing data, to determine the factors that influenced housing prices in the 70's and how these factors might have changed over the past 40 years.

Objective: With this project I aim to investigate factors that influence housing prices in Boston for new homes.

The question to be answered is: What variables significantly determine the price of houses in Boston, how does the locality impact the value of an individual's home? Using publicly available information/statistics, we will attempt to infer historically, how these factors change over a period.

From a real estate business perspective knowledge of housing price determinants in a historical context can guide business owners on how to invest in housing projects and price their homes. Furthermore, such a knowledge can influence the kind of housing design and features developers could add to their homes to mitigate any negative effect of a factor, thereby enhancing the value of their homes irrespective of the area such homes are situated.

References:

1. Nguyen, J. Key Factors That Drive the Real Estate Market. *Key Factors That Drive the Real Estate Market*.

2. Case, K. E. The market for single-family homes in the Boston area. *New England Economic Review* (1986).

Data, Justification, and source:

The data is made of features such as structural quality, neighborhood, accessibility, and air pollution such as per capita crime rate by town, proportion of non-retail business acres per town, student-to-teacher ratio, proximity to water bodies, demographic and index of accessibility to radial highways. Such features offer the opportunity to make comparative assessment, given features like population, air quality, demographics amongst others have been identified as

The Boston housing data, which was originally published by Harrison, D. and Rubinfeld, D.L. `Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978. The data will be acquired from https://archive.ics.uci.edu/ml/machine-learning-databases/housing/.
The dataset contains information collected by the U.S Census Service concerning housing around Boston Massachusetts.

The data has 13 predictive features for Median value of owner-occupied homes in Boston(MEDV). The 13 features are described below:

- CRIM: per capita crime rate by town
- ZN :proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS: proportion of non-retail business acres per town
- CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX: nitric oxides concentration (parts per 10 million)
- RM: average number of rooms per dwelling
- AGE: proportion of owner-occupied units built prior to 1940
- DIS: weighted distances to five Boston employment centres
- RAD: index of accessibility to radial highways
- TAX: full-value property-tax rate per $10,000
- PTRATIO: pupil-teacher ratio by town
- B: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- LSTAT : % lower status of the population
- MEDV: Median value of owner-occupied homes in $1000's

Data Wrangling and cleaning:

Data is clean with no missing values. CHAS was not considered a categorical variable, pd.get_dummies program was used to generate catigorical variables for these feature and changed the data type to categorical.
A closer look at housing prices per age group suggest that there are outliers especially for Houses 70 years or older.

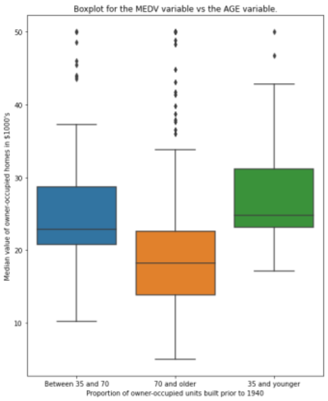


Figure 1: Proportion of owner-occupied units built prior to 1940 by age grouping. Though median price of senior occupied homes is less, there are a more senior homes valued at a higher price than average homes. Given their age, homes 35 and under have better value than homes 35 and above. As expected very old homes are valued less.

We found that TAX and RAD are highly correlated features. One of these two features may need to be dropped.

Given that the columns LSTAT, INDUS, RM, TAX, NOX, and PTRAIO have correlation scores about 0.5 and above with MEDV, they are good candidates for predictors(see Figure 2 below).

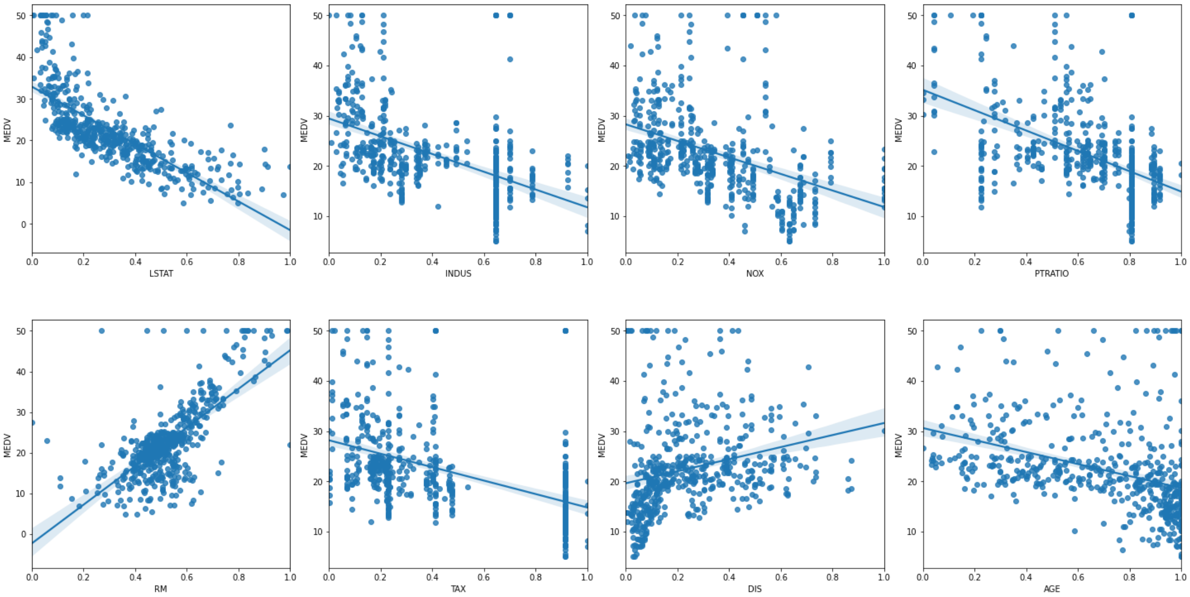


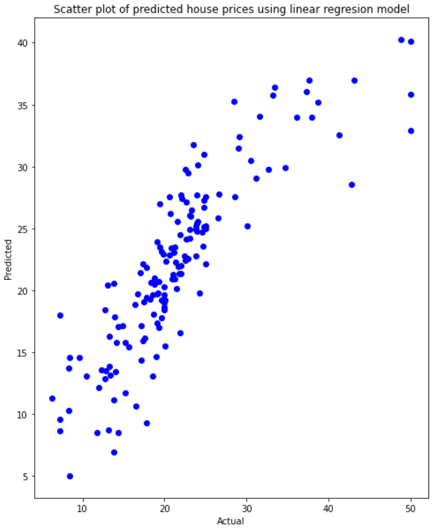
Figure 2: Correlation of 8 features with median price variable(MEDV)

Building models

The data was split into train and test using 70:30% ratio. Train data was used to train 3 models, Linear regression, Support Vector Machine(SVM), and k-nearest neighbors (KNN) and Gradient boosting(GB).

The Mean square error(MSE) of the models was used to select the best model.

See below the MSE of the models: KNN: -29.74 (+/- 26.20), GB: -30.90 (+/- 8.55), SVR: -48.84 (+/- 28.13), and LR: -31.05 (+/- 37.10).



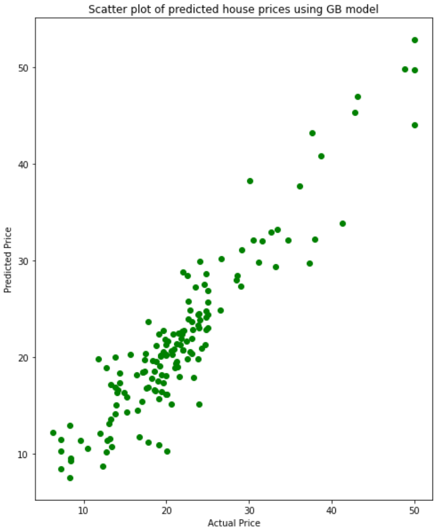


Figure3: Comparing predictive power of 2 best models LR and GB.
GB shows a better predictive power .

Conclusion:

GB models performs slightly better than LR model and should be used for predicting prices.

Among the features, RAD: proximity to highway, RM: Average number of room, LSAT: Demographic/% of lower(poor) in population, NOX: Nitrix oxide/air quality, TAX: tax, PTRatio: pupil-teacher ratio, and DIS: access to work/commute are among the key influencers Figure 4).

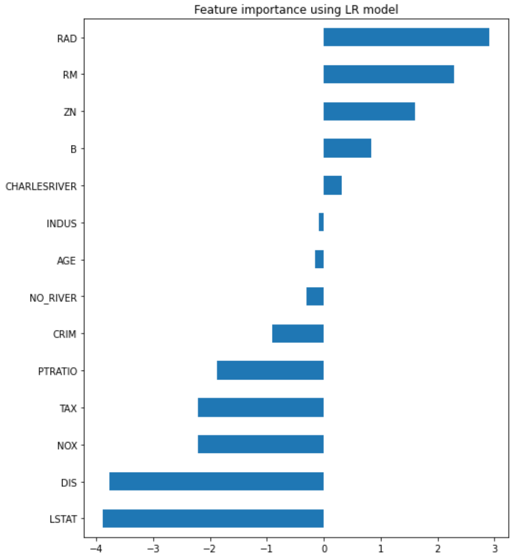


Figure 4: The important features selected using the LR model. RAD, RM, LSAT, NOX, TAX, PTRatio and, DIS are important for predicting housing prices. It is surprising CRIM(Crime) is not a strong feature.

As shown factors such as demographics, number of rooms, air quality and commute are factors that have always influence housing prices. As such investors and buyers should consider these factors when deciding to invest or purchase a home respectively.

Future direction:

- Remove features with less influence on our model and re-train the models.
- Compare current dataset with the 1978 data to make a better comparison.