Q1: Boston House Prices

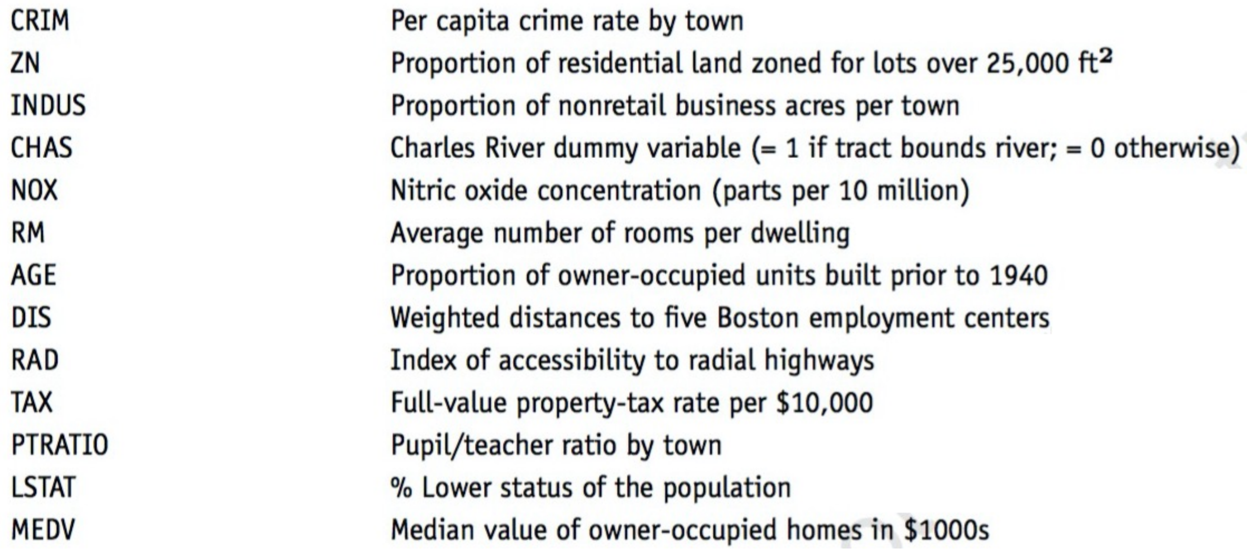
Fall, 2022

G Park

2022-12-04

QUESTION ONE

Question 1: Predicting Boston Housing Prices. (25 points) The file BostonHousing.csv contains information by the US Bureau of the Census concerning housing in the area of Boston, Massachusetts. The dataset includes information on 506 census-housing tracts in the Boston area. The goal is to predict the median house price in new tracts based on information such as crime rate, pollution, and number of rooms. The dataset contains 13 predictors, and the response is the median house price (MEDV). Table 1 describes each of the predictors and the response.



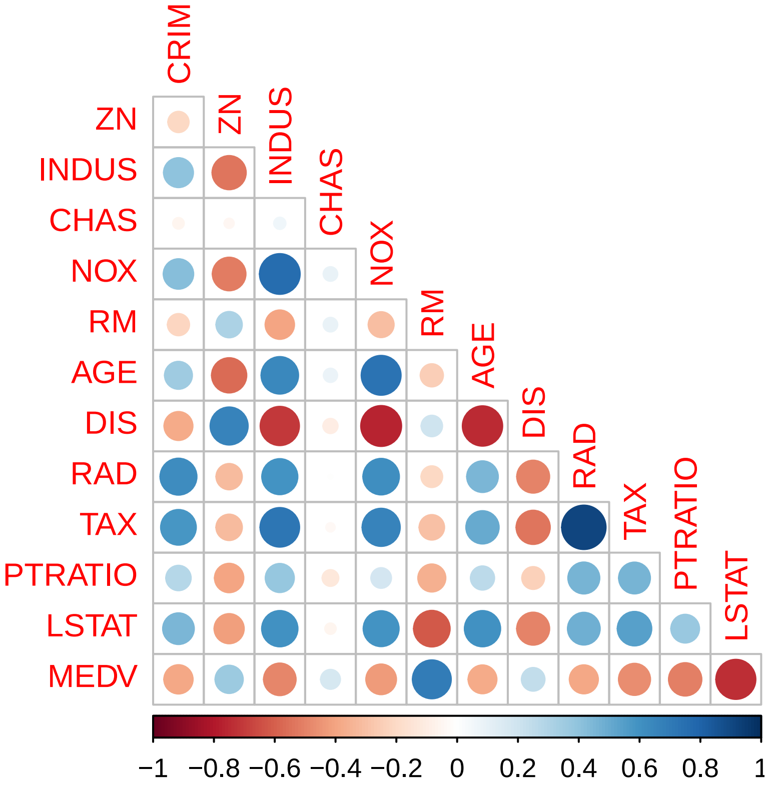
CRIM	Per capita crime rate by town
ZN	Proportion of residential land zoned for lots over 25,000 ft^2
INDUS	Proportion of nonretail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; = 0 otherwise)
NOX	Nitric oxide concentration (parts per 10 million)
RM	Average number of rooms per dwelling
AGE	Proportion of owner-occupied units built prior to 1940
DIS	Weighted distances to five Boston employment centers
RAD	Index of accessibility to radial highways
TAX	Full-value property-tax rate per $10,000
PTRATIO	Pupil/teacher ratio by town
LSTAT	% Lower status of the population
MEDV	Median value of owner-occupied homes in $1000s

Figure 1: variables Description

Compute the correlation table for all variables and show which variables has the strongest positive and which variable has the strongest positive and which variable has the strongest negative correlations with the median house price (MEDV). (1 points)

```
##          CRIM    ZN INDUS  CHAS   NOX    RM   AGE   DIS   RAD   TAX PTRATIO
## CRIM        1
## ZN       -0.2     1
## INDUS    0.41 -0.53     1
## CHAS    -0.06 -0.04  0.06     1
## NOX      0.42 -0.52  0.76  0.09     1
## RM      -0.22  0.31 -0.39  0.09  -0.3     1
## AGE      0.35 -0.57  0.64  0.09  0.73 -0.24     1
## DIS     -0.38  0.66 -0.71  -0.1 -0.77  0.21 -0.75     1
## RAD      0.63 -0.31   0.6 -0.01  0.61 -0.21  0.46 -0.49     1
## TAX      0.58 -0.31  0.72 -0.04  0.67 -0.29  0.51 -0.53  0.91     1
## PTRATIO  0.29 -0.39  0.38 -0.12  0.19 -0.36  0.26 -0.23  0.46  0.46       1
## LSTAT    0.46 -0.41   0.6 -0.05  0.59 -0.61   0.6  -0.5  0.49  0.54    0.37
## MEDV    -0.39  0.36 -0.48  0.18 -0.43   0.7 -0.38  0.25 -0.38 -0.47   -0.51
##         LSTAT MEDV
## CRIM
## ZN
## INDUS
## CHAS
## NOX
## RM
## AGE
## DIS
## RAD
## TAX
## PTRATIO
## LSTAT       1
## MEDV    -0.74     1
```

- The average number of rooms per dwelling has the highest positive correlation with median house price (0.7).
- Percent lower status of the population has the highest negative correlation with median house price (-0.74).



Why should the data be partitioned into training and validation sets? What will the training set be used for? What will the validation set be used for? (3 points)

We split the data into training and validation sets to allow us to train and tune a model on one set (the training set) and evaluate the model on a completely new set (the validation set) that has not been used in model development and selection. This technique allows us to select a model that would do best in the real world by limiting `overfitting`.

The training set is useful for training and selecting appropriate machine learning model or models.

We use the validation set to approximate our model’s unbiased accuracy in new settings. In other words the validation set allows us to evaluate how the model would perform given new data.

Partition records into 60% for training and 40% for validation sets. Then fit a multiple linear regression to MEDV as a function of CRIM, CHAS, and RM for training sets and show the summarized regression results. Based on your regression results, make a prediction on your validation sets. Be sure to write the equation for predicting the median house price from the predictors in the model and interpret your regression results. (8 points).

```
##
## Call:
## lm(formula = MEDV ~ CRIM + CHAS + RM, data = training_set)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24.09  -2.78  -0.20   2.23  39.19
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -30.7248     3.4044   -9.03  < 2e-16 ***
## CRIM         -0.2371     0.0364   -6.51  3.1e-10 ***
## CHAS          2.1048     1.3466    1.56     0.12
## RM            8.5918     0.5335   16.10  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.2 on 299 degrees of freedom
## Multiple R-squared:  0.547,  Adjusted R-squared:  0.543
## F-statistic:  121 on 3 and 299 DF,  p-value: <2e-16
```

The estimated equation is as follows:

$MEDV = -30.7248 - 0.2371CRIM + 2.1048CHAS + 8.5918RM + error_term$

There is an inverse relationship between `crime` and `median house prices`. Both Charles River dummy (CHAS) and number of rooms (RM) have a positive relationship with median house prices, ceteris paribus.

Specifically, a unit increase in crime reduces median house prices by 237.10 holding all other variables constant. A house located in the Charles river is, on average US$ 2104.80 more expensive than a house not located there, all else remaining the same. An extra room raises prices by US$ 8591.80 all else remaining the same.

Fit another multiple linear regression model to the median house price (MEDV) as a function of LSTAT, INDUS, and NOX for training sets and show the summarized regression results. Then, make a prediction based on this model for your validation sets. (5 points)

```
##
## Call:
## lm(formula = MEDV ~ LSTAT + INDUS + NOX, data = training_set)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9.21  -4.29  -1.11   2.15  25.02
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  32.3136     2.0248   15.96   <2e-16 ***
## LSTAT        -0.9395     0.0640  -14.69   <2e-16 ***
## INDUS        -0.1464     0.0884   -1.66    0.099 .
## NOX           6.7692     4.8808    1.39    0.167
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.2 on 299 degrees of freedom
## Multiple R-squared:  0.547,  Adjusted R-squared:  0.542
## F-statistic:  120 on 3 and 299 DF,  p-value: <2e-16
```

The associated equation is as follows:

$MEDV = 32.31360 - 0.93954LSTAT - 0.14641INDUS + 6.76920NOX + error_term$

Both the proportion of lower status population`LSTAT` and proportion of non-retail businesses `INDUS` have an inverse relationship with `median house prices`. Nitrogen Oxide (`NOX`) concentration has a positive relationship with median house prices, ceteris paribus.

Ceteris paribus, a unit increase in low status population corresponds to a reduction in median house prices by US$ 939.54 on average. A unit increase in the proportion of non-retail businesses is associated with a rise in median house prices by an average of US$ 146.41. Finally, a unit rise in Nitrogen Oxide corresponds to an average increase in house prices by US$ 6769.20.

Now get the accuracy metrics for these two predicted models and show their accuracy metrics respectively. Based on their accuracy metrics, select which model is the best model in terms of the accuracy metrics, and why. (8 points)

The root mean squared error for the first and second regression are as follows.

```
## [1] 6.2
```

```
## [1] 6.2
```

The mean absolute error is as follows.

```
## [1] 4.3
```

```
## [1] 4.4
```

In both cases, the first model that relates median house prices with CRIM, CHAS, and RM is better because it has lower RMSE and lower MSE. Also, the first model has higher R^2 and adjusted R^2.

```
## # A tibble: 1 x 12
##   r.squared adj.r.squa~1 sigma stati~2  p.value    df logLik   AIC   BIC devia~3
##       <dbl>        <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.547        0.543  6.19    121. 3.47e-51     3  -980. 1970. 1989.  11445.
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squa~1 sigma stati~2  p.value    df logLik   AIC   BIC devia~3
##       <dbl>        <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.547        0.542  6.19    120. 4.55e-51     3  -980. 1971. 1989.  11466.
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```