**CASE STUDY 1**

1. a. Data frame dimensions, i.e. Number of rows and columns in data set: (506, 14).

b. **Original column titles: Python Output**

Index(['CRIME', 'ZONE', 'INDUST', 'CHAR RIV', 'NIT OXIDE', 'ROOMS', 'AGE','DISTANCE', 'RADIAL', 'TAX', 'ST RATIO', 'LOW STAT', 'MVALUE','C MVALUE'], dtype='object')

**Modified column titles with no space and one word for titles: Python Output**

Index(['CRIME', 'ZONE', 'INDUST', 'CHAR\_RIV', 'NIT\_OXIDE', 'ROOMS', 'AGE', 'DISTANCE', 'RADIAL', 'TAX', 'ST\_RATIO', 'LOW\_STAT', 'MVALUE', 'C\_MVALUE'], dtype='object')

c. **Modified list of column variables with dummy variables: Python Output**

Index(['CRIME', 'ZONE', 'INDUST', 'NIT\_OXIDE', 'ROOMS', 'AGE', 'DISTANCE', 'RADIAL', 'TAX', 'ST\_RATIO', 'LOW\_STAT', 'MVALUE', 'CHAR\_RIV\_Y', 'C\_MVALUE\_Yes'], dtype='object')

**d. Descriptive statistics for all the columns in the modified data frame:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **CRIME** | **ZONE** | **INDUST** | **NIT OXIDE** | **ROOMS** | **AGE** | **DISTANCE** | **RADIAL** | **TAX** | **ST RATIO** | **LOW STAT** | **MVALUE** | **CHAR RIV\_Y** | **C\_MVALUE\_Yes** |
| count | 506 | 506 | 506 | 506 | 506 | 506 | 506 | 506 | 506 | 506 | 506 | 506 | 506 | 506 |
| mean | 3.61 | 11.36 | 11.14 | 0.55 | 6.28 | 68.57 | 3.8 | 9.55 | 408.24 | 18.46 | 12.65 | 22.53 | 0.07 | 0.17 |
| std | 8.6 | 23.32 | 6.86 | 0.12 | 0.7 | 28.15 | 2.11 | 8.71 | 168.54 | 2.16 | 7.14 | 9.2 | 0.25 | 0.37 |
| min | 0.01 | 0 | 0.46 | 0.38 | 3.56 | 2.9 | 1.13 | 1 | 187 | 12.6 | 1.73 | 5 | 0 | 0 |
| 25% | 0.08 | 0 | 5.19 | 0.45 | 5.89 | 45.02 | 2.1 | 4 | 279 | 17.4 | 6.95 | 17.02 | 0 | 0 |
| 50% | 0.26 | 0 | 9.69 | 0.54 | 6.21 | 77.5 | 3.21 | 5 | 330 | 19.05 | 11.36 | 21.2 | 0 | 0 |
| 75% | 3.68 | 12.5 | 18.1 | 0.62 | 6.62 | 94.07 | 5.19 | 24 | 666 | 20.2 | 16.96 | 25 | 0 | 0 |
| max | 88.98 | 100 | 27.74 | 0.87 | 8.78 | 100 | 12.13 | 24 | 711 | 22 | 37.97 | 50 | 1 | 1 |

**Comment on Missing values**: The count of all the columns in the data set is 506 which is same as number of rows in the data set. Hence there are no missing values in the data set.

1. a. The intercept and regression coefficients of the Regression Model for Boston Housing Training Set are given below:

**Intercept:** 43.65

**Predictor Coefficient**

0 CRIME -0.14

1 ZONE 0.01

2 INDUST 0.12

3 NIT\_OXIDE -16.47

4 ROOMS 0.89

5 AGE -0.01

6 DISTANCE -0.72

7 RADIAL 0.20

8 TAX -0.01

9 ST\_RATIO -0.58

10 LOW\_STAT -0.45

11 CHAR\_RIV\_Y 2.11

12 C\_MVALUE\_Yes 10.99

**Mathematical equation for Linear regression Model:**

43.65 -0.14 \* CRIME + 0.01 \* ZONE + 0.12 \* INDUST -16.47 \* NIT\_OXIDE + 0.89 \* ROOMS -0.01 \* AGE -0.72 \* DISTANCE + 0.2 \* RADIAL -0.01 \* TAX -0.58 \* ST\_RATIO -0.45 \* LOW\_STAT + 2.11 \* CHAR\_RIV\_Y + 10.99 \* C\_MVALUE\_Yes

b. Prediction Performance Measures for Training Set

r2 : 0.839

Adjusted r2 : 0.832

Prediction Performance Measures for Validation Set

r2 : 0.834

Adjusted r2 : 0.822

The r2 and the adjusted r2 for this regression model are high (close to 1) for both training and validation sets which is an indicative of a good fit. The differences of r2 and adjusted r2 between training and validation sets are not high, hence there is no possibility of overfitting of the data. It indicates that the regression model that we built can be used for predicting.

c. Common Accuracy Measures for Training Set - All Variables

Regression statistics

Mean Error (ME) : 0.0000

Root Mean Squared Error (RMSE) : 3.5845

Mean Absolute Error (MAE) : 2.5961

Mean Percentage Error (MPE) : -2.7127

Mean Absolute Percentage Error (MAPE) : 13.1715

Common Accuracy Measures for Validation Set - All Variables

Regression statistics

Mean Error (ME) : 0.4347

Root Mean Squared Error (RMSE) : 3.8763

Mean Absolute Error (MAE) : 2.7696

Mean Percentage Error (MPE) : -2.2773

Mean Absolute Percentage Error (MAPE) : 13.3233

There is no much difference in the RMSE’s of training and validation sets, we can say there is no possibility of overfitting of the data. It indicates that the regression model that we built can be used for predicting. The MAPE’s of training and validation sets are less than 15% i.e. the margin of errors are less than 15% and are not much different (13.1715 and 13.3233), hence based on the MAPE’s, we can say this model has no possibility of overfitting of the data and can be considered for making predictions.

1. **a. Regression Model for Training Set Using Exhaustive Search**

Table

Description automatically generated

* Performance measures, *r2adj* and *AIC*, improve as predictors added up to *11* predictors, and then become worse again with more predictors
* Overall, the best number of predictors, *n=11,* is identified based on the *highest* *adjusted R2* and *lowest AIC.*

**Intercept** 43.89

**Predictor Coefficient**

0 CHAR\_RIV\_Y 2.13

1 CRIME -0.14

2 C\_MVALUE\_Yes 11.11

3 DISTANCE -0.63

4 INDUST 0.11

5 LOW\_STAT -0.46

6 NIT\_OXIDE -16.89

7 RADIAL 0.19

8 ROOMS 0.86

9 ST\_RATIO -0.61

10 TAX -0.01

**Mathematical equation for Linear regression Model based on Exhaustive Search:**

43.89 + 2.13 \* CHAR\_RIV\_Y -0.14 \* CRIME + 11.11 \* C\_MVALUE\_Yes -0.63 \* DISTANCE + 0.11 \* INDUST -0.46 \* LOW\_STAT -16.89 \* NIT\_OXIDE + 0.19 \* RADIAL + 0.86 \* ROOMS -0.61 \* ST\_RATIO -0.01 \* TAX

Accuracy Measures for Validation Set Using Exhaustive Search

Regression statistics

Mean Error (ME) : 0.4505

Root Mean Squared Error (RMSE) : 3.8674

Mean Absolute Error (MAE) : 2.7724

Mean Percentage Error (MPE) : -2.1963

Mean Absolute Percentage Error (MAPE) : 13.3441

**b. Regression Model for Training Set Using Forward Selection**

**Intercept:**  42.76

**Predictor Coefficient**

0 C\_MVALUE\_Yes 10.97

1 LOW\_STAT -0.45

2 CRIME -0.14

3 CHAR\_RIV\_Y 2.36

4 ST\_RATIO -0.60

5 ROOMS 0.87

6 DISTANCE -0.71

7 NIT\_OXIDE -15.95

8 RADIAL 0.11

**Mathematical equation for Linear regression Model based on Forward Selection:**

42.76 + 10.97 \* C\_MVALUE\_Yes -0.45 \* LOW\_STAT -0.14 \* CRIME + 2.36 \* CHAR\_RIV\_Y -0.6 \* ST\_RATIO + 0.87 \* ROOMS -0.71 \* DISTANCE -15.95 \* NIT\_OXIDE + 0.11 \* RADIAL

Accuracy Measures for Validation Set Using Forward Selection

Regression statistics

Mean Error (ME) : 0.4321

Root Mean Squared Error (RMSE) : 3.9314

Mean Absolute Error (MAE) : 2.8585

Mean Percentage Error (MPE) : -2.3792

Mean Absolute Percentage Error (MAPE) : 13.8040

**Model** **Number of predictors**

1) Model based on Exhaustive search 11

2) Model based on Forward Selection 9

Best Variables from Exhaustive Search Algorithm

['C\_MVALUE\_Yes', 'LOW\_STAT', 'CRIME', 'CHAR\_RIV\_Y', 'ST\_RATIO', 'ROOMS', 'DISTANCE', 'NIT\_OXIDE', 'RADIAL', ‘INDUST’, ‘TAX’]

Best Variables from Forward Selection Algorithm

['C\_MVALUE\_Yes', 'LOW\_STAT', 'CRIME', 'CHAR\_RIV\_Y', 'ST\_RATIO', 'ROOMS', 'DISTANCE', 'NIT\_OXIDE', 'RADIAL']

The forward selection algorithm gives only 9 best predictors for linear regression model whereas the exhaustive search algorithm gives 11 best predictors with “INDUST” and “TAX” predictors in addition to the predictors of forward selection algorithm. The RMSE’s are also not much different. Clearly, the model based on the Forward selection algorithm is considered to be the Parsimonious model and can be recommended for predictions.

**c.**

**Model** **Number of predictors RMSE**

1) Regression model with all predictors 13 3.8763

2) Model based on Exhaustive search 11 3.8674

3) Model based on Forward Selection 9 3.9314

The RMSE is lower for the model based on the Exhaustive search with 11 number of predictors than the RMSE’s of the regression model with all predictors and the model based on the Forward Selection with 9 number of predictors. The lower RMSE, the better the model. Based on RMSE, the multiple linear regression model based on the Exhaustive search is recommended for making predictions.

However, the RMSE’s are not much different for all the three models. The parsimonious model which is the model based on the Forward Selection with 9 number of predictors can be recommended for making predictions. The simplest model that performs better results.