Modules

Grades

Inbox

Calc

Pre-Processing

Pre Processing Assessment
Assignment

Model Building

Pre Processing Assessment

The due date for submitting this assignment has passed.
Due on 2024-03-10, 23:59 IST.

You may submit any number of times before the due date. The final submission will be considered for grading.

You have last submitted on: 2024-03-09, 18:27 IST

Following is description of different columns in the dataset.

CRIM: per capita crime rate in the vicinity

ZN: amount of residential land reserved in the vicinity.

INDUS: proportion of industrial land reserved nearby (in square kilometers)

RIVERSIDE: If the boundary faces river side (= 1 if tract bounds river; 0 otherwise)

POLINDEX: polution index

RM: number of rooms in the house.

AGE: Age of the property in years.

DIS: weighted distances to the major economic centres (in kilometers)

HIGHWAYCOUNT: Number of highways within 5 KM of distance.

TAX: full-value property-tax rate per 1 lac.

PTRATIO: student-teacher ratio in the vicinity.

IMM: Immigration index in the vicinity.

BPL: % of below poverty line population in the vicinity.

PRICE: Price of the home in lacs, this is the target column.

Note: For numerical type questions, always enter the answer correct upto 3 decimal places without rounding off, unless otherwise stated.

Click here to view the sklearn library reference

Click here to view the Colab File

1) Click here to view the dataset *1 point*

Which dataset are you using for this exam?

- V1
- V2
- V3
- V4
- V5

Yes, the answer is correct.
Score: 1

Accepted Answers:

V3

2) How many samples are there in the dataset? *2 points*

- 4000
- 5000
- 6000
- 7000

Yes, the answer is correct.
Score: 2

Accepted Answers:

4000

3) What is the average house price (in lacs)?

24.61622

24.61922

Yes, the answer is correct.
Score: 2

Accepted Answers:

(Type: Range) 24.611,24.621

2 points

4) How many houses have 5 or more rooms? **3 points**

- 3953
- 3921
- 3896
- 3932
- 3925
- None of these

Yes, the answer is correct.
Score: 3

Accepted Answers:

3896

5) What is the average price of the top 10 most expensive houses (in lacs)?

52.2007

Yes, the answer is correct.
Score: 3

Accepted Answers:

(Type: Range) 52.195,52.205

3 points

6) What is the total number of missing or unknown values in the number of rooms feature? **2 points**

(Hint: carefully look at the values the feature takes and find out implausible value.)

- 40
- 71
- 99
- 61
- 68
- None of these

Yes, the answer is correct.
Score: 2

Accepted Answers:

99

7) What is the total number of missing or unknown values in the age feature? **2 points**

(Hint: carefully look at the values the feature takes and find out implausible value.)

- 50
- 83
- 74
- 64
- 59
- None of these

Yes, the answer is correct.
Score: 2

Accepted Answers:

74

8) What is the total number of missing or unknown values in the RIVERSIDE feature? **2 points**

(Hint: carefully look at the values the feature takes and find out implausible value.)

- 88
- 101
- 56
- 62
- 80
- None of these

Yes, the answer is correct.
Score: 2

Accepted Answers:

56

9) How many houses are on riverside and were built within the last 50 years (i.e. a house 50 years old or younger)? For this question, ignore the rows that have missing values in either riverside feature or age feature. ***3 points***

- 44
- 42
- 52
- 43
- 37
- None of these

Yes, the answer is correct.
Score: 3

Accepted Answers:

52

10) How many houses are near to exactly 6, 7 or 8 highways (all three inclusive)? ***3 points***

- 1211
- 1174
- 1234
- 938
- 1209
- None of these

Yes, the answer is correct.
Score: 3

Accepted Answers:

1234

11) Create a column 'CATEGORY' and divide the houses in categories as following: ***4 points***

Category 1: house price <10 lacs
Category 2: 10 lacs <= house price <20 lacs
Category 3: 20 lacs <= house price <30 lacs
Category 4: 30 lacs <= house price <40 lacs
Category 5: house price >=40 lacs

Which category has the highest number of records?

- 1
- 2
- 3
- 4
- 5
- There is a tie between multiple categories

Yes, the answer is correct.
Score: 4

Accepted Answers:

3

12) **Apply Pre processing** *2 points*

Divide the data into training and test sets

1. Replace the respective missing or unknown values in features room count, riverside and age with np.nan.

2. Keep 30% of the data as test set.

3. Use random_state as 0

4. PRICE is the target, rest of the columns are the features.

5. Apply train test split.

Hint: look for the documentation of the usual function that divides the data into training and test datasets.

What is the number of samples in the training set?

- (•) 2800
- ( ) 2900
- ( ) 3000
- ( ) 3100

Yes, the answer is correct.
Score: 2

Accepted Answers:

2800

13) Apply following preprocessing steps: *5 points*

1. Drop CATEGORY column
2. CRIM: min max scaling
3. ZN: min max scaling
4. INDUS: standard scaling
5. POLINDEX: min max scaling
6. DIS: min max scaling
7. HIGHWAYCOUNT: min max scaling
8. TAX: min max scaling
9. PTRATIO: min max scaling
10. IMM: min max scaling
11. BPL: min max scaling
12. RM: impute with median then min max scaling
13. AGE: impute with mean then min max scaling
14. RIVERSIDE: Impute with most frequent value then one hot encode.

NOTE:
1. Make sure to preprocess the features in exactly above order. Answer of Q.16 depends upon correct order of featuring processing.
2. You may have to use multiple instances of a trasnformer for this question.

How many features are there after performing above transformation?

- (•) 14
- ( ) 15
- ( ) 16
- ( ) 17
- ( ) 18

Yes, the answer is correct.
Score: 5

Accepted Answers:

14

14) What is the mean of the transformed test data (features only)?
Note : Compute the mean of the whole feature matrix i.e. mean of all values in the transformed test feature matrix

No, the answer is incorrect.
Score: 0

Accepted Answers:

(Type: Range) 0.382,0.389

2 points

15) Apply PCA on processed training matrix only (do not transform the data). What is the value of largest eigen value?

No, the answer is incorrect.
Score: 0

Accepted Answers:

(Type: Range) 56.995,57.005

4 points

16) If you eliminate 1 feature with recursive feature elimination, which feature will be eliminated?

Type the index of the eliminated feature (index starts from 0).
Use LinearRegression model with default parameters as an estimator.
Use processed training data.

No, the answer is incorrect.
Score: 0

Accepted Answers:

(Type: Numeric) 3

5 points