MA691: Advanced Statistical Algorithms

Sristy Sharma(180121043)

Prathapani Sravya(180123033)

Yokesh S(180121051)

Kritika Raj(180123024)

Dhoolam Sai Chandan(180123011)

November 20, 2021

1 Assignment (Regression)

1. Import the following datasets

- Boston Housing dataset(https://www.cs.toronto.edu/ delve/data/boston/ bostonDetail.html),
- Diabetes dataset (https://archive.ics.uci.edu/ml/datasets/diabetes)

2. Use dierent regression strategy [MLR, Ridge, Lasso, (k-NN, kernel Cobra) etc] and compare the accuracy based on suitable benchmark.

3. Use cross-validation, cross check if it aects accuracy.

4. Determine Outlier if possible.

Solution 1)

Figure 1.1: Importing Boston dataset

```
from sklearn.datasets import load_boston
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import Ridge, Lasso
from sklearn.metrics import r2_score
from sklearn.model_selection import cross_val_score, GridSearchCV
import warnings

warnings.filterwarnings('ignore', category = DeprecationWarning)
from warnings import filterwarnings
filterwarnings('ignore')
```

```
df_b = load_boston()
data_b = pd.DataFrame(df_b.data,columns = df_b.feature_names)
data_b['PRICE'] = df_b.target
X_b = data_b.drop('PRICE',axis = 1)
y_b = data_b['PRICE']
```

Figure 1.2: Diabetes dataset from UCI machine learning repository

```
: df = pd.read_table('diabetes.txt', names = ['id', 'date', 'time', 'code', 'value'])
```

Diabetes Dataset

Diabetes patient records were obtained from two sources: an automatic electronic recording device and paper records. The automatic device had an internal clock to timestamp events, whereas the paper records only provided "logical time" slots (breakfast, lunch, dinner, bedtime). For paper records, fixed times were assigned to breakfast (08:00), lunch (12:00), dinner (18:00), and bedtime (22:00). Thus paper records have fictitious uniform recording times whereas electronic records have more realistic time stamps.

Diabetes files consist of four fields per record. Each field is separated by a tab and each record is separated by a newline.

File Names and format:
(1) Date in MM-DD-YYYY format
(2) Time in XX:YY format
(3) Code
(4) Value

The Code field is deciphered as follows:

33 = Regular insulin dose
34 = NPH insulin dose
35 = UltraLente insulin dose
48 = Unspecified blood glucose measurement
57 = Unspecified blood glucose measurement
58 = Pre-breakfast blood glucose measurement
59 = Post-breakfast blood glucose measurement
60 = Pre-lunch blood glucose measurement
61 = Post-lunch blood glucose measurement
62 = Pre-supper blood glucose measurement
63 = Post-supper blood glucose measurement
64 = Pre-snack blood glucose measurement
65 = Hypoglycemic symptoms
66 = Typical meal ingestion
67 = More-than-usual meal ingestion
68 = Less-than-usual meal ingestion
69 = Typical exercise activity
70 = More-than-usual exercise activity
71 = Less-than-usual exercise activity
72 = Unspecified special event

Attribute Information:

Diabetes files consist of four fields per record. Each field is separated by a tab and each

record is separated by a newline.
File Names and format:
(1) Date in MM-DD-YYYY format
(2) Time in XX:YY format
(3) Code
(4) Value

Boston Housing Dataset

The Boston Housing Dataset is a derived from information collected by the U.S. Census Service concerning housing in the area of Boston MA. The following describes the dataset columns:

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

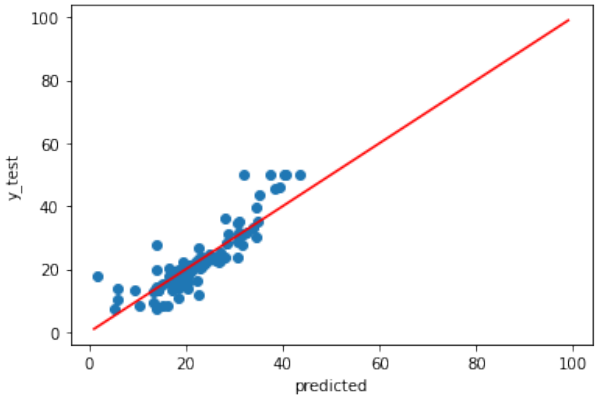
Solution 2)

We used $r2$ square for finding accuracy of models on both datsets i.e. Boston Housing and Diabetes dataset.

a)For Boston dataset:

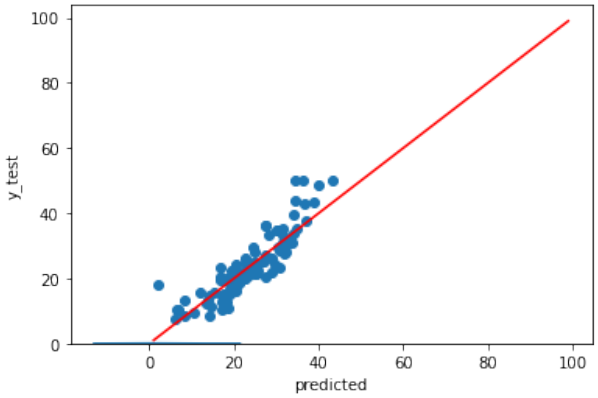
1)MLR (Multiple Linear regression)
Accuracy:
Training accuracy : 0.7309
Testing accuracy : 0.76601

Figure 1.3: Boston dataset(MLR)



2)Ridge regression
Accuracy:
Training accuracy : 0.72699
Testing accuracy : 0.77799

Figure 1.4: Boston dataset(Ridge)



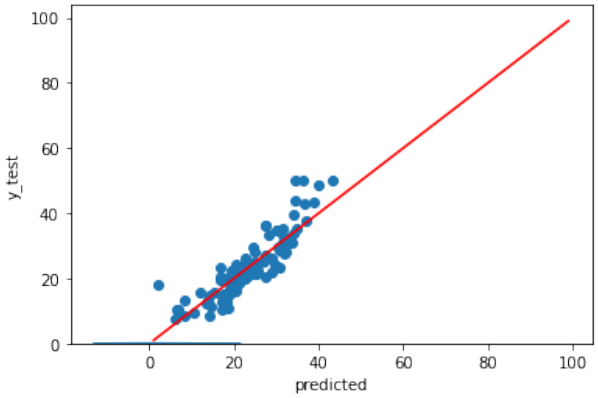
3)Lasso regression

Accuracy:

Training accuracy : 0.72699

Testing accuracy : 0.77799

Figure 1.5: Boston dataset(Lasso)



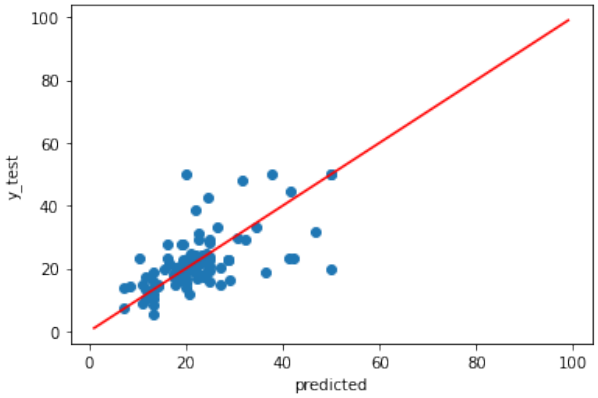
4)KNN(k-nearest neighbors)

Accuracy:

Training accuracy : 1.0

Testing accuracy : 0.27368

Figure 1.6: Boston dataset(KNN)



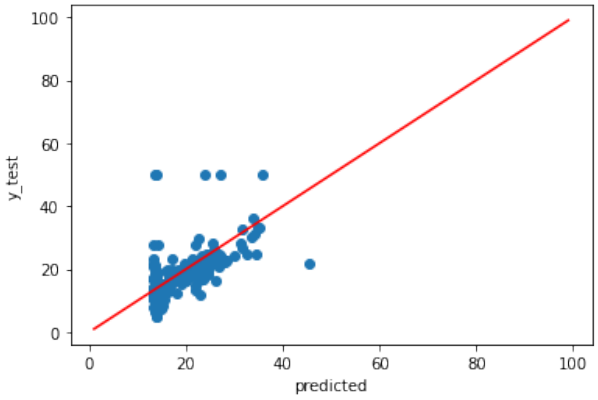
5)Kernel Cobra

Accuracy:

Training accuracy : 1.0

Testing accuracy : 0.27368

Figure 1.7: Boston dataset(Kernel cobra)



From the training and testing accuracy we observed that MLR, Ridge and Lasso regressions are performing better than Kernel Cobra and KNN. The KNN and Kernel Cobra models are overfitted on the Boston Dataset.

b)For Diabetes dataset:

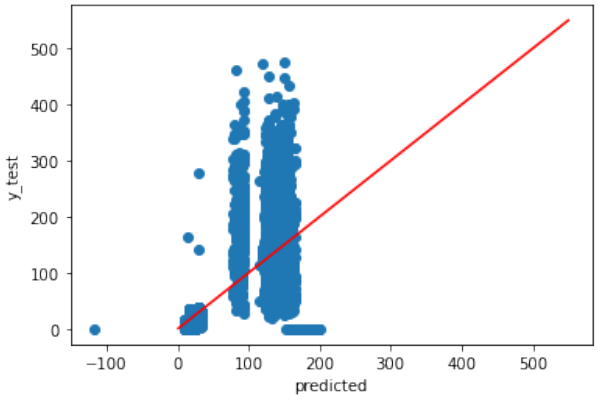
1)MLR (Multiple Linear regression)

Accuracy:

Training accuracy : 0.42356

Testing accuracy : 0.42865

Figure 1.8: Diabetes dataset(MLR)



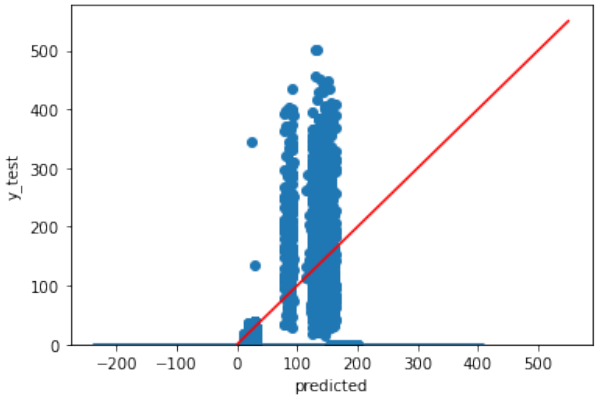
2)Ridge regression

Accuracy:

Training accuracy : 0.42495

Testing accuracy : 0.42308

Figure 1.9: Diabetes dataset(Ridge)



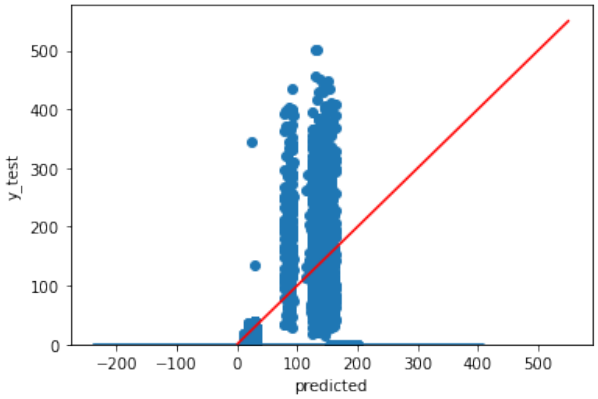
3)Lasso regression

Accuracy:

Training accuracy : 0.42495

Testing accuracy : 0.42308

Figure 1.10: Diabetes dataset(Lasso)



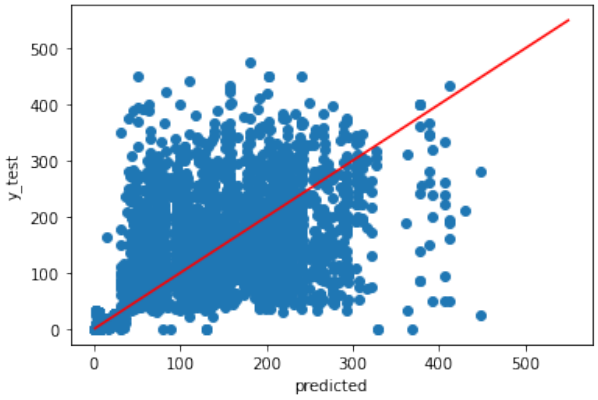
4)KNN(k-nearest neighbors)

Accuracy:

Training accuracy : 0.48877

Testing accuracy : 0.49478

Figure 1.11: Diabetes dataset(KNN)



From the training and testing accuracy we observed that all regressions perform almost same on the diabetes dataset among which KNN performed better than other models.

Solution 3)

Cross Validation:
We have done cross validation using different number of splits

a)For Boston dataset:

1)MLR (Multiple Linear regression)

Accuracy:

```
training accuracy using cross validation for  2  splits: 0.7129894723884295
testing accuracy using cross validation for  2  splits: 1.0
training accuracy using cross validation for  6  splits: 0.7023684017019939
testing accuracy using cross validation for  6  splits: 1.0
training accuracy using cross validation for  10  splits: 0.7123850563561016
testing accuracy using cross validation for  10  splits: 1.0
training accuracy using cross validation for  14  splits: 0.7003296378676457
testing accuracy using cross validation for  14  splits: 1.0
training accuracy using cross validation for  18  splits: 0.6688717685384468
testing accuracy using cross validation for  18  splits: 1.0
training accuracy using cross validation for  22  splits: 0.6928310738662232
testing accuracy using cross validation for  22  splits: 1.0
training accuracy using cross validation for  26  splits: 0.6560156499921372
testing accuracy using cross validation for  26  splits: 1.0
```

We observed that MLR model is performing with better accuracy after cross validation.

2)Ridge regression

Accuracy:

```
training accuracy using cross validation for  2  splits: 0.7049447854395877
testing accuracy using cross validation for  2  splits: 0.5983717384163795
training accuracy using cross validation for  6  splits: 0.6956456932616755
testing accuracy using cross validation for  6  splits: 0.5989153960333504
training accuracy using cross validation for  10  splits: 0.695442192157836
testing accuracy using cross validation for  10  splits: 0.5920222563471058
training accuracy using cross validation for  14  splits: 0.6763778704796498
testing accuracy using cross validation for  14  splits: 0.5984489132843006
training accuracy using cross validation for  18  splits: 0.6598204956656817
testing accuracy using cross validation for  18  splits: 0.49091283269523955
training accuracy using cross validation for  22  splits: 0.6185485535178177
testing accuracy using cross validation for  22  splits: 0.2517924360039833
training accuracy using cross validation for  26  splits: 0.6080789816985452
testing accuracy using cross validation for  26  splits: 0.4695632051127543
```

We observed that Ridge model is performing with lower accuracy after cross validation.

3)Lasso regression

Accuracy:

```
training accuracy using cross validation for  2  splits: 0.7087423023717209
testing accuracy using cross validation for  2  splits: 0.5822297921057609
training accuracy using cross validation for  6  splits: 0.6997638267027524
testing accuracy using cross validation for  6  splits: 0.6046109597865461
training accuracy using cross validation for  10  splits: 0.7001126893733407
testing accuracy using cross validation for  10  splits: 0.6077358096538751
training accuracy using cross validation for  14  splits: 0.6780760320166407
testing accuracy using cross validation for  14  splits: 0.6114019142501023
training accuracy using cross validation for  18  splits: 0.6640278172145176
testing accuracy using cross validation for  18  splits: 0.4527402375133214
training accuracy using cross validation for  22  splits: 0.6224980225882725
testing accuracy using cross validation for  22  splits: 0.20454854128384745
training accuracy using cross validation for  26  splits: 0.6108663141531001
testing accuracy using cross validation for  26  splits: 0.4799423330748081
```

We observed that Lasso model is performing with lower accuracy after cross validation.

4)KNN (k-nearest neighbors)

Accuracy:

```
training accuracy using cross validation for  2  splits: 0.33678983189943784
testing accuracy using cross validation for  2  splits: -0.6337787448976931
training accuracy using cross validation for  6  splits: 0.3647581341893605
testing accuracy using cross validation for  6  splits: -1.0007056339681306
training accuracy using cross validation for  10  splits: 0.30626412025022287
testing accuracy using cross validation for  10  splits: -2.1925631115506627
training accuracy using cross validation for  14  splits: 0.27455848276625044
testing accuracy using cross validation for  14  splits: -1.6832128805193083
training accuracy using cross validation for  18  splits: 0.2976777701360751
testing accuracy using cross validation for  18  splits: -1.2546479072521701
training accuracy using cross validation for  22  splits: 0.29452832968989656
testing accuracy using cross validation for  22  splits: -1.3480803399235153
training accuracy using cross validation for  26  splits: 0.2380331657978558
testing accuracy using cross validation for  26  splits: -2.051320635127184
```

We observed that KNN model is giving negative cross validation.That means that the fitted model is worse than the null hypothesis.

5)Kernal Cobra

Accuracy:

```
training accuracy using cross validation for  2  splits: 0.80384375805472
testing accuracy using cross validation for  2  splits: 0.34852367606542567
training accuracy using cross validation for  6  splits: 0.8110774140239312
testing accuracy using cross validation for  6  splits: 0.3953149589870732
training accuracy using cross validation for  10  splits: 0.8027034495790961
testing accuracy using cross validation for  10  splits: 0.3444816846720861
training accuracy using cross validation for  14  splits: 0.8275078920987872
testing accuracy using cross validation for  14  splits: 0.4425689542482911
training accuracy using cross validation for  18  splits: 0.7706131438217451
testing accuracy using cross validation for  18  splits: 0.4683181989356499
training accuracy using cross validation for  22  splits: 0.7295165182471028
testing accuracy using cross validation for  22  splits: 0.32064484496189044
training accuracy using cross validation for  26  splits: 0.7393092404633332
testing accuracy using cross validation for  26  splits: 0.3803410330075142
```

We observed that kernel cobra is performing with better accuracy after cross validation.

b)For Diabetes dataset:

1)MLR (Multiple Linear regression)

Accuracy:

```
training accuracy using cross validation for  2  splits: 0.42341755687621424
testing accuracy using cross validation for  2  splits: 0.4271190203674475
training accuracy using cross validation for  6  splits: 0.42336710269704275
testing accuracy using cross validation for  6  splits: 0.42782767751163675
training accuracy using cross validation for  10  splits: 0.4233082215060483
testing accuracy using cross validation for  10  splits: 0.4265192012030588
training accuracy using cross validation for  14  splits: 0.42298640637893087
testing accuracy using cross validation for  14  splits: 0.42673265096470514
training accuracy using cross validation for  18  splits: 0.42320855047288536
testing accuracy using cross validation for  18  splits: 0.42648053014414744
training accuracy using cross validation for  22  splits: 0.42236666240854215
testing accuracy using cross validation for  22  splits: 0.42714633207279884
training accuracy using cross validation for  26  splits: 0.42286454259943795
testing accuracy using cross validation for  26  splits: 0.42663893114581675
```

We observed that MLR model is performing with similar accuracy after cross validation.

2)Ridge regression

Accuracy:

```
training accuracy using cross validation for  2  splits: 0.424720773536081
testing accuracy using cross validation for  2  splits: 0.41984567774689613
training accuracy using cross validation for  6  splits: 0.42488858236170013
testing accuracy using cross validation for  6  splits: 0.42142183204508177
training accuracy using cross validation for  10  splits: 0.42487440050734293
testing accuracy using cross validation for  10  splits: 0.4217869691652525
training accuracy using cross validation for  14  splits: 0.42475329023028696
testing accuracy using cross validation for  14  splits: 0.4209371388627866
training accuracy using cross validation for  18  splits: 0.4247195963726245
testing accuracy using cross validation for  18  splits: 0.42091682917813045
training accuracy using cross validation for  22  splits: 0.4242140230931615
testing accuracy using cross validation for  22  splits: 0.42032494789212743
training accuracy using cross validation for  26  splits: 0.42448676124084095
testing accuracy using cross validation for  26  splits: 0.42038166484295064
```

We observed that Ridge model is performing with similar accuracy after cross validation.

3)Lasso regression

Accuracy:

```
training accuracy using cross validation for  2  splits: 0.4247208496076768
testing accuracy using cross validation for  2  splits: 0.41979102098979115
training accuracy using cross validation for  6  splits: 0.4248885918518748
testing accuracy using cross validation for  6  splits: 0.4214229015132081
training accuracy using cross validation for  10  splits: 0.4248744242489382
testing accuracy using cross validation for  10  splits: 0.4217877438637786
training accuracy using cross validation for  14  splits: 0.4247533279431437
testing accuracy using cross validation for  14  splits: 0.42093930033655536
training accuracy using cross validation for  18  splits: 0.4247196005978011
testing accuracy using cross validation for  18  splits: 0.42091927081907676
training accuracy using cross validation for  22  splits: 0.42421402629631366
testing accuracy using cross validation for  22  splits: 0.420327722155238
training accuracy using cross validation for  26  splits: 0.4244867502983303
testing accuracy using cross validation for  26  splits: 0.4203833648186383
```

We observed that Lasso model is performing with similar accuracy after cross validation.

4)KNN (k-nearest neighbors)

Accuracy:

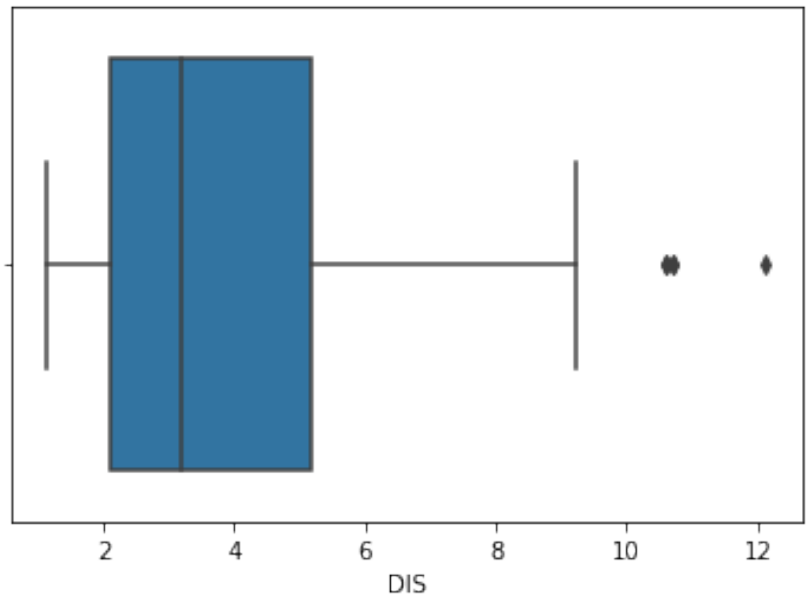
```
training accuracy using cross validation for  2  splits: 0.44771978709121096
testing accuracy using cross validation for  2  splits: 0.45267232569371996
training accuracy using cross validation for  6  splits: 0.4287942316939816
testing accuracy using cross validation for  6  splits: 0.4442550963388306
training accuracy using cross validation for  10  splits: 0.44488643150852514
testing accuracy using cross validation for  10  splits: 0.4110235244645065
training accuracy using cross validation for  14  splits: 0.47015865405142326
testing accuracy using cross validation for  14  splits: 0.4355167705041212
training accuracy using cross validation for  18  splits: 0.44152727870599207
testing accuracy using cross validation for  18  splits: 0.45393511788373797
training accuracy using cross validation for  22  splits: 0.4460170681148622
testing accuracy using cross validation for  22  splits: 0.4691399502644879
training accuracy using cross validation for  26  splits: 0.4418735514845843
testing accuracy using cross validation for  26  splits: 0.43528682640025596
```

We observed that KNN model is performing with similar accuracy even after cross validation.

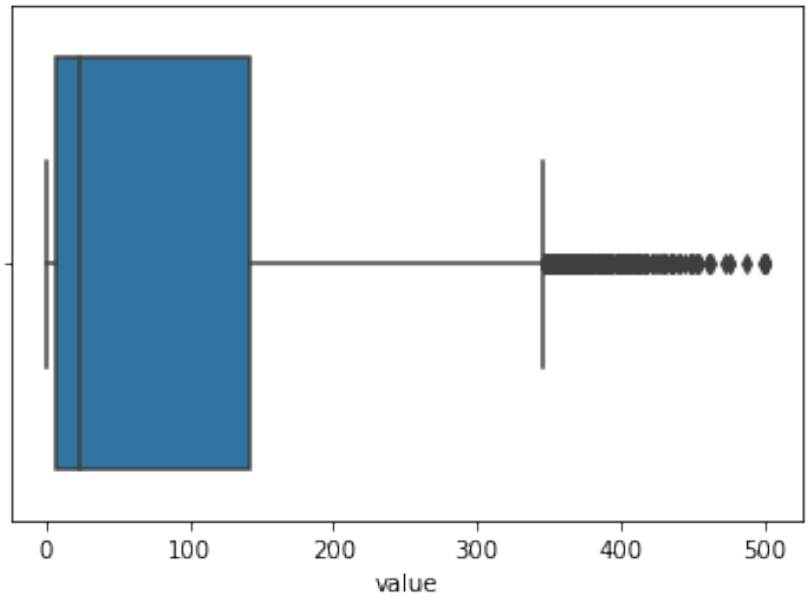
Solution 4)

Outliers:

a)For Boston dataset:



a)For Diabetes dataset:



We observed two outliers in the Boston dataset, whereas multiple outliers in the Diabetes

dataset resulted in better accuracy on the Boston dataset compared to Diabetes dataset.