DSEM 6150 ASSIGNMENT 6

XGBOOST ALGORITHM

PRESENTED BY: RUCHI KAPADIWALA

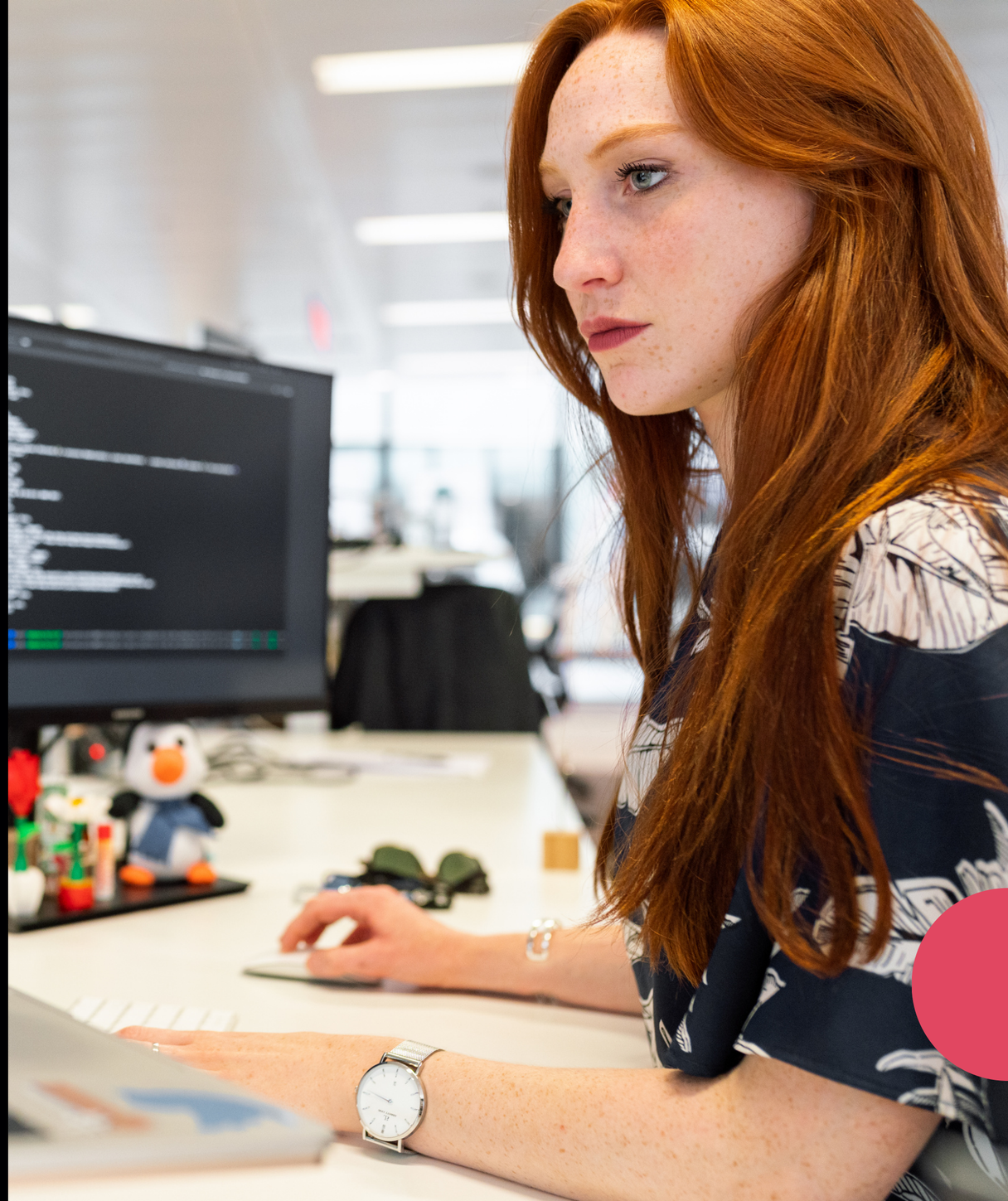
DSEM 6150 ASSIGNMENT 6

INTRODUCTION

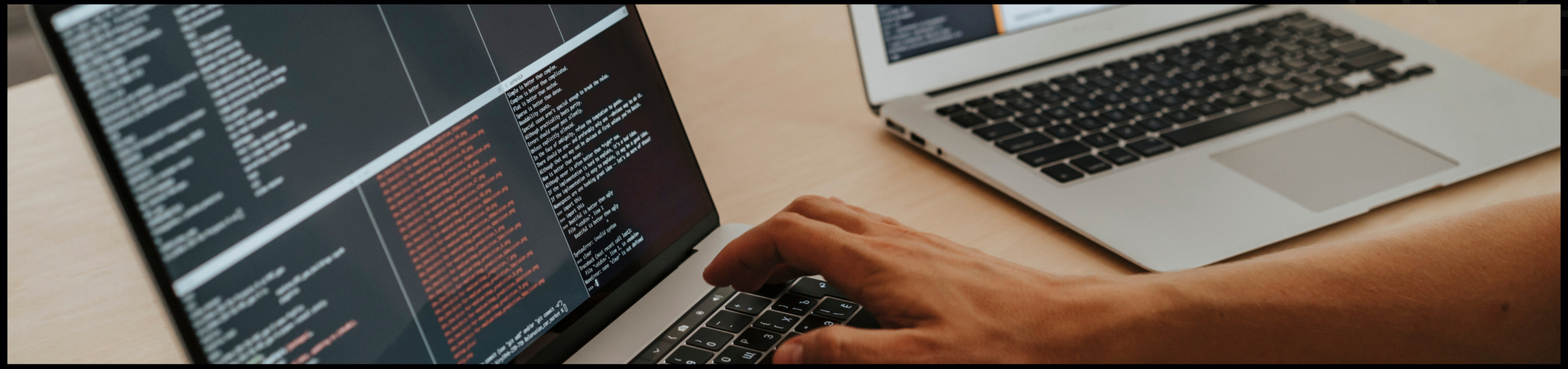
XGBoost(Extreme Gradient Boosting) is a popular machine learning algorithm that is widely used for solving regression and classification problems.

It is an implementation of the gradient boosting algorithm, which builds an ensemble of weak decision trees and combines them to make a strong prediction.

XGBoost is known for its scalability, speed, and high predictive accuracy, making it a popular choice for many data science applications.



DSEM 6150 ASSIGNMENT 6



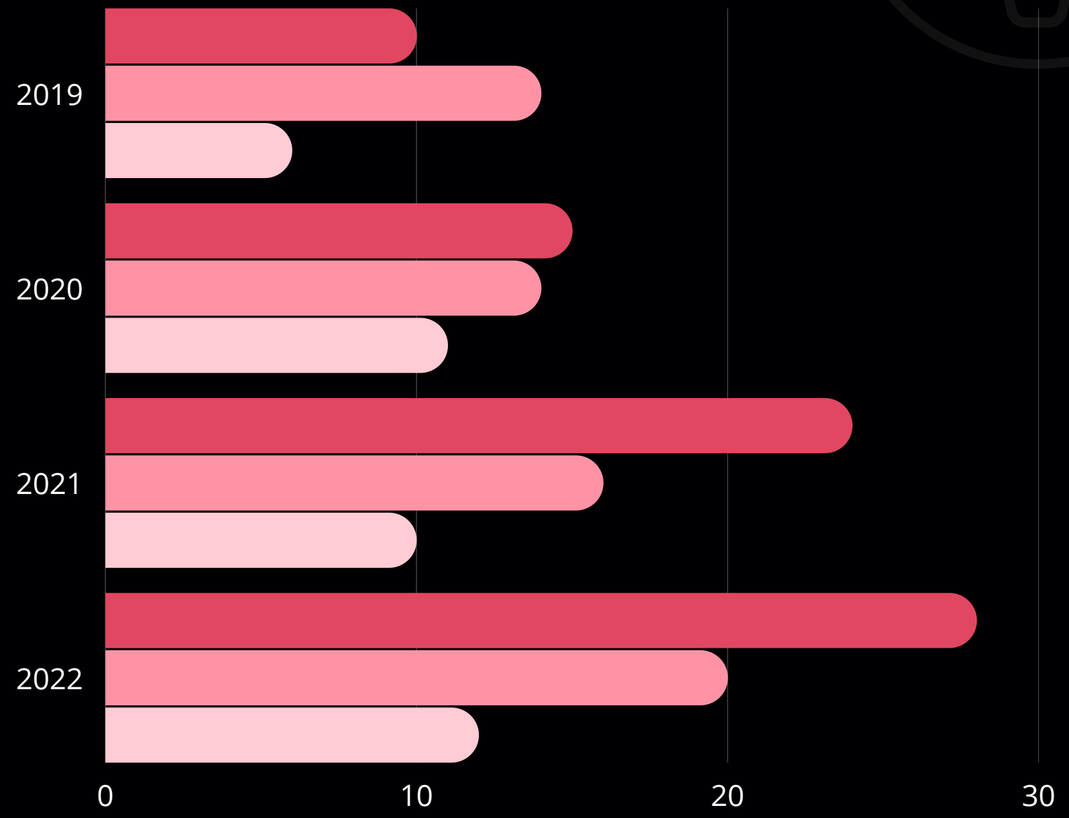
PROBLEM STATEMENT

Applying XGBoost algorithm on the Boston housing dataset to predict the median value. Do the usual train/test split and then compare the result with the result of applying linear regression. You can use a suitable metric like RMSE.

DATA DESCRIPTION

Description

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:



DSEM 6150 ASSIGNMENT 6

01

CRIM

per capita crime rate by town

02

ZN

proportion of residential land zoned for lots over 25,000 sq.ft.

03

INDUS

proportion of non-retail business acres per town.

04

CHAS

Charles River dummy variable (1 if tract bounds river; 0 otherwise)

05

NOX

nitric oxides concentration (parts per 10 million)

06

RM

average number of rooms per dwelling

DSEM 6150 ASSIGNMENT 6

07

AGE

proportion of owner-occupied units built prior to 1940

08

DIS

weighted distances to five Boston employment centres

09

RAD

index of accessibility to radial highways

10

PTRATIO

full-value property-tax rate per $10,000

11

B

1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

12

LSTAT

% lower status of the population

13

MV

Median value of owner-occupied homes in $1000's

DSEM 6150 ASSIGNMENT 6

TECHNICAL APPROACH

Data preprocessing: Cleaning and transforming data to prepare it for modeling

Splitting Target Variable and Independent Variable

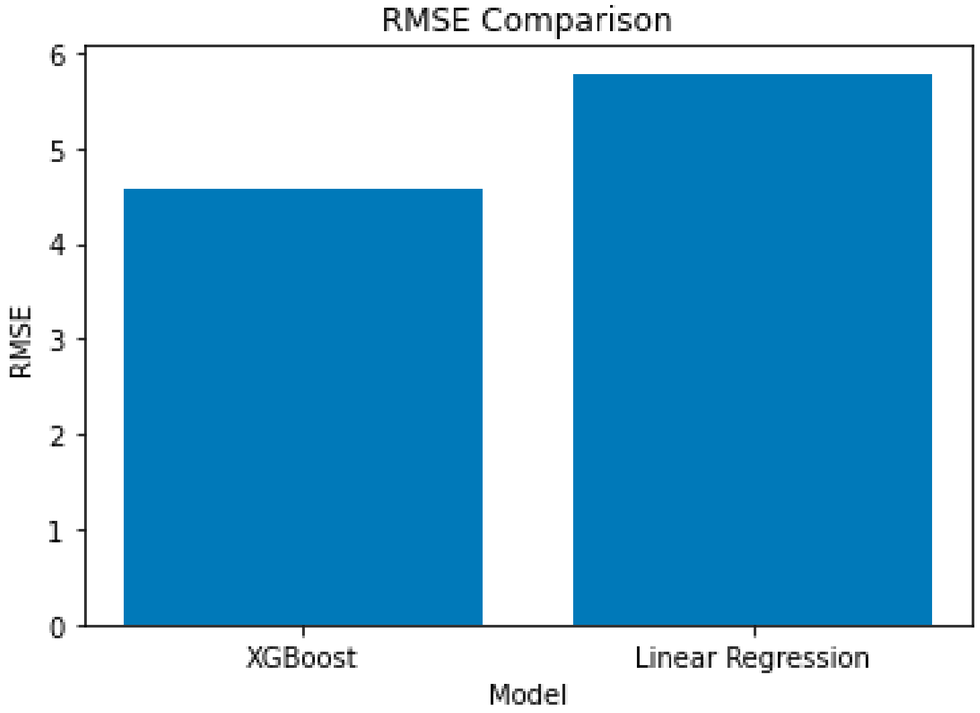
Splitting Training and Testing Data

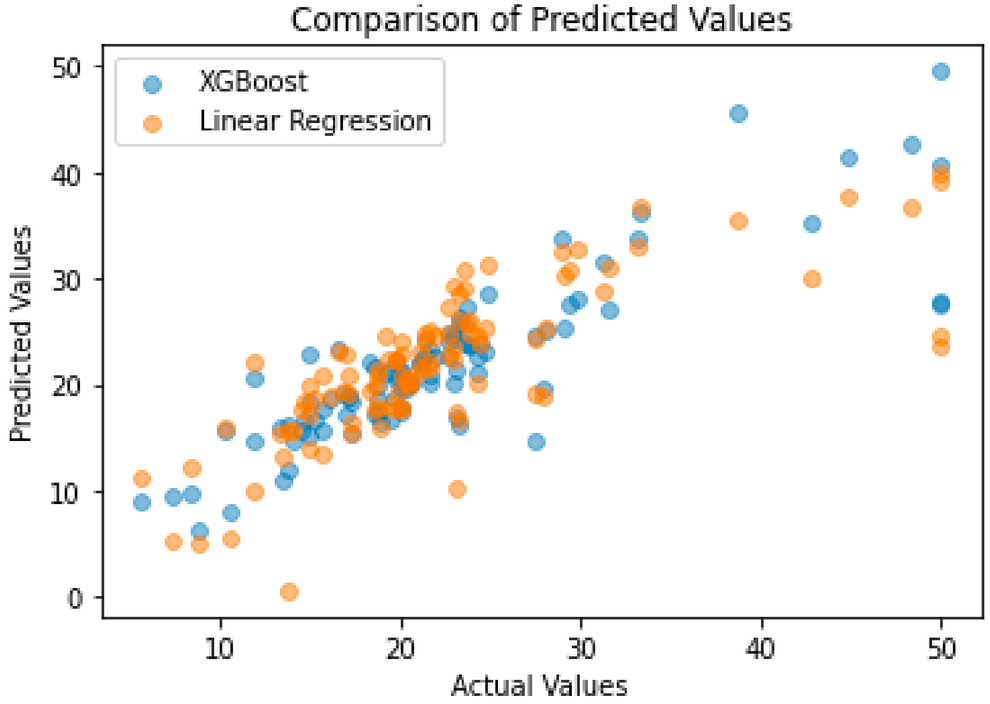
Fit the model to the data

Training the model on the preprocessed data

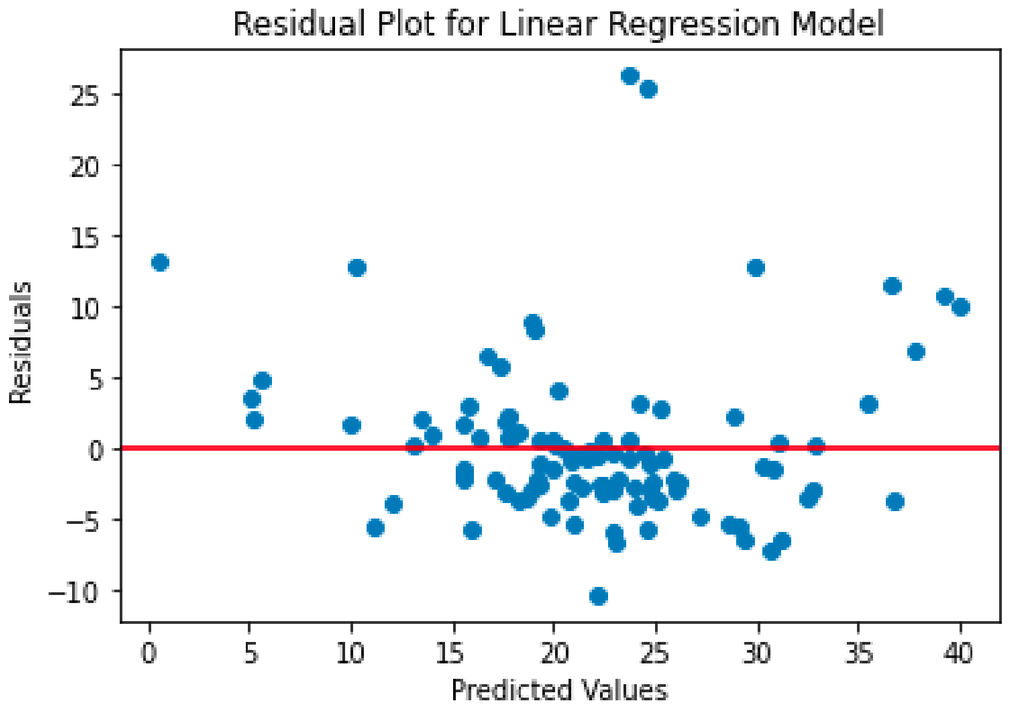
Deploying the final model to make predictions on new data

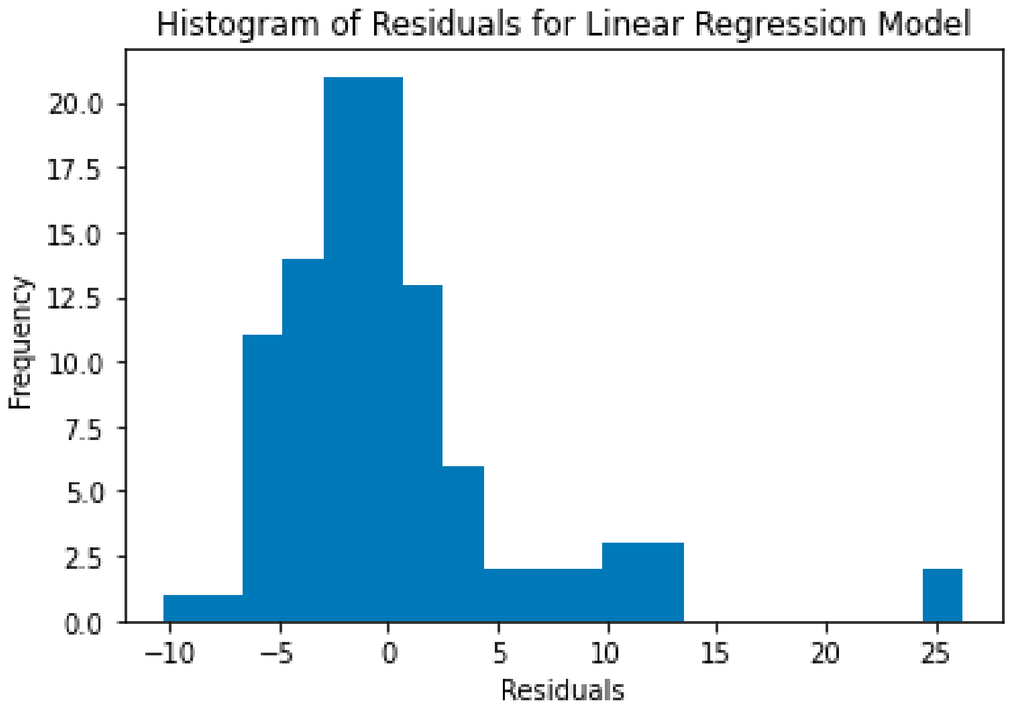
Results


RMSE Comparison
RMSE
0
1
2
3
4
5
6
XGBoost
Linear Regression
Model

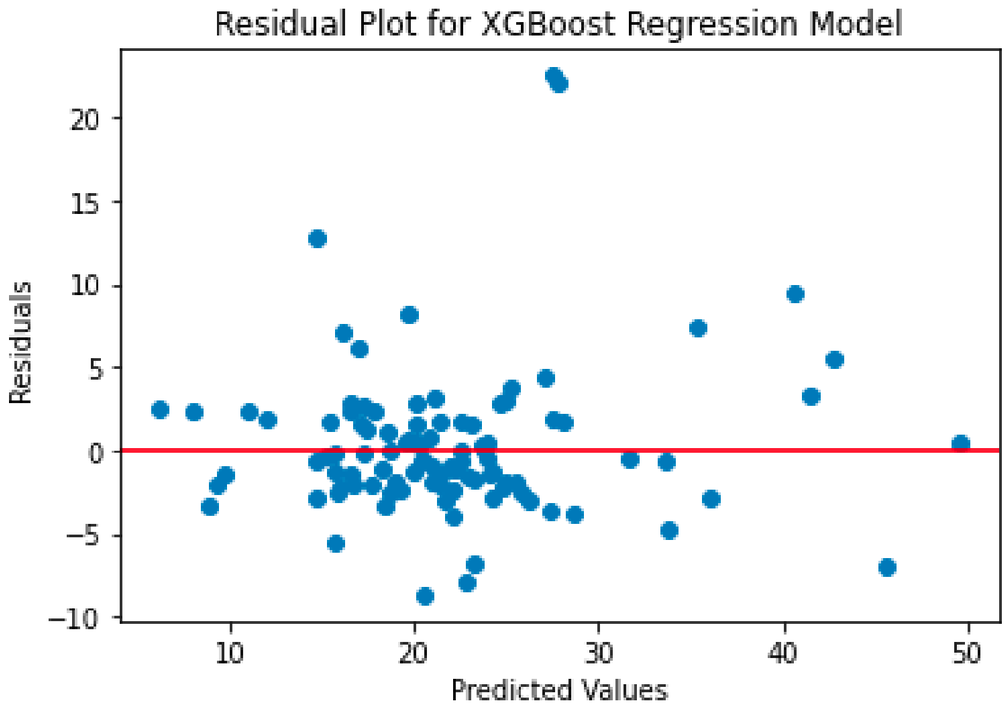

Comparison of Predicted Values
XGBoost
Linear Regression
Predicted Values
0
10
20
30
40
50
10
20
30
40
50
Actual Values

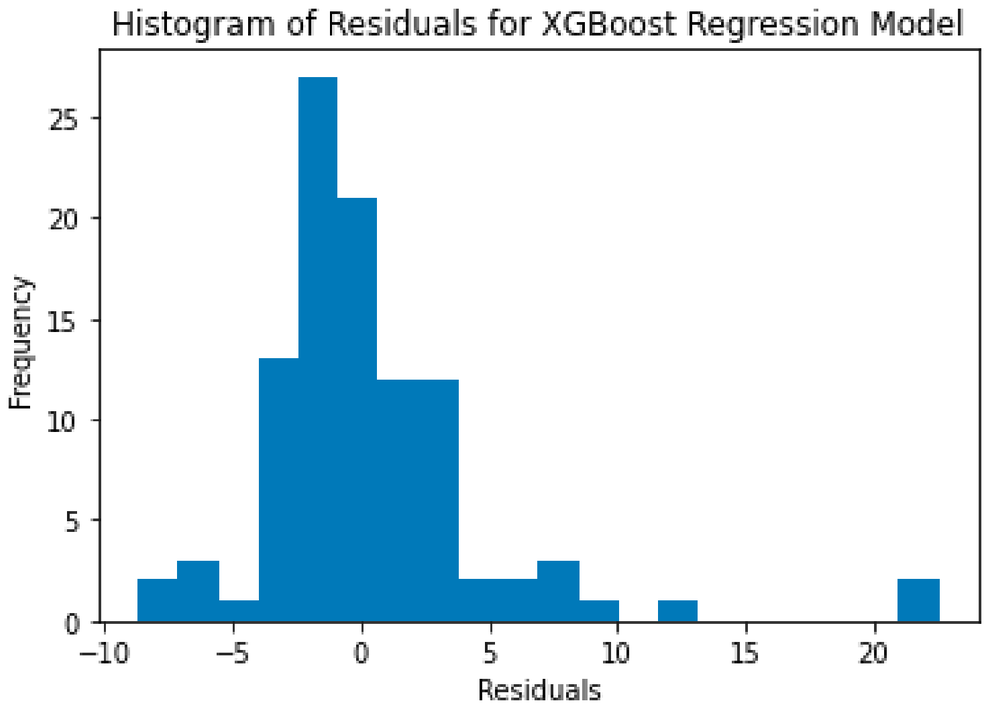
Results


Residual Plot for Linear Regression Model
Residuals
Predicted Values
25
20
15
10
5
0
−5
−10
0
5
10
15
20
25
30
35
40

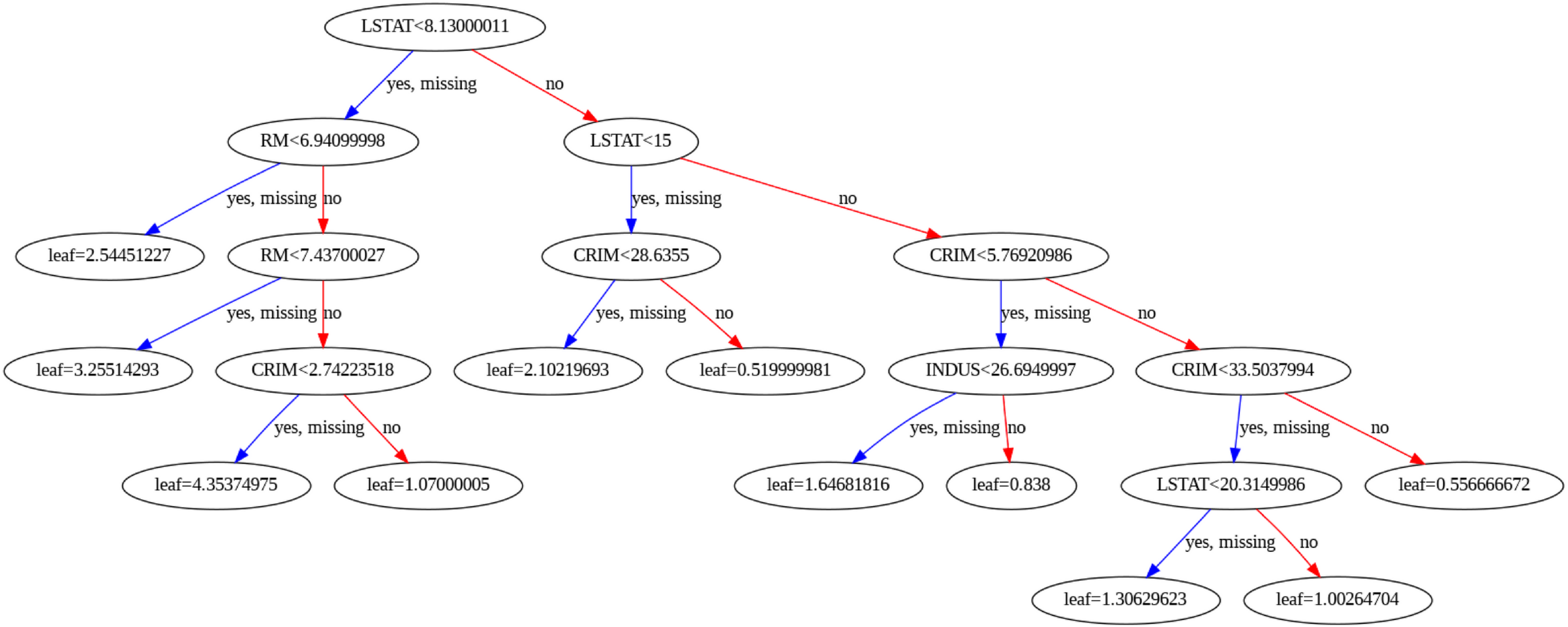

Histogram of Residuals for Linear Regression Model
Frequency
Residuals
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−10
−5
0
5
10
15
20
25

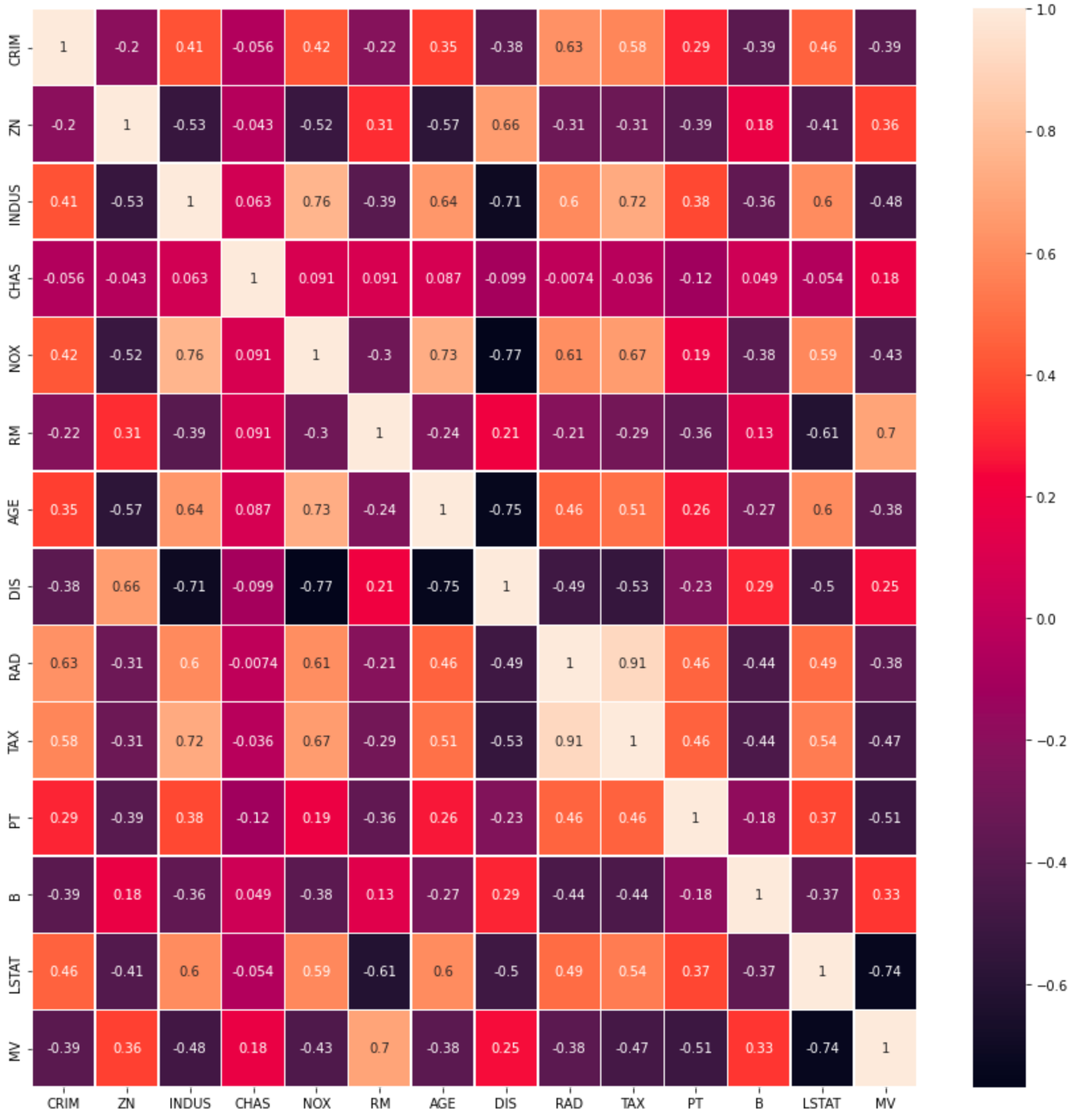
Results


Residual Plot for XGBoost Regression Model
Residuals
Predicted Values
20
15
10
5
0
−5
−10
10
20
30
40
50


Histogram of Residuals for XGBoost Regression Model
Frequency
Residuals
25
20
15
10
5
0
−10
−5
0
5
10
15
20

Results


LSTAT<8.13000011
yes, missing
no
RM<6.94099998
LSTAT<15
yes, missing
no
yes, missing
no
leaf=2.54451227
RM<7.43700027
CRIM<28.6355
CRIM<5.76920986
yes, missing
no
yes, missing
no
yes, missing
no
leaf=3.25514293
CRIM<2.74223518
leaf=2.10219693
leaf=0.519999981
INDUS<26.6949997
CRIM<33.5037994
yes, missing
no
yes, missing
no
yes, missing
no
leaf=4.35374975
leaf=1.07000005
leaf=1.64681816
leaf=0.838
LSTAT<20.3149986
leaf=0.556666672
yes, missing
no
leaf=1.30629623
leaf=1.00264704


CRIM 1 -0.2 0.41 -0.056 0.42 -0.22 0.35 -0.38 0.63 0.58 0.29 -0.39 0.46 -0.39
ZN -0.2 1 -0.53 -0.043 -0.52 0.31 -0.57 0.66 -0.31 -0.31 -0.39 0.18 -0.41 0.36
INDUS 0.41 -0.53 1 0.063 0.76 -0.39 0.64 -0.71 0.6 0.72 0.38 -0.36 0.6 -0.48
CHAS -0.056 -0.043 0.063 1 0.091 0.091 0.087 -0.099 -0.0074 -0.036 -0.12 0.049 -0.054 0.18
NOX 0.42 -0.52 0.76 0.091 1 -0.3 0.73 -0.77 0.61 0.67 0.19 -0.38 0.59 -0.43
RM -0.22 0.31 -0.39 0.091 -0.3 1 -0.24 0.21 -0.21 -0.29 -0.36 0.13 -0.61 0.7
AGE 0.35 -0.57 0.64 0.087 0.73 -0.24 1 -0.75 0.46 0.51 0.26 -0.27 0.6 -0.38
DIS -0.38 0.66 -0.71 -0.099 -0.77 0.21 -0.75 1 -0.49 -0.53 -0.23 0.29 -0.5 0.25
RAD 0.63 -0.31 0.6 -0.0074 0.61 -0.21 0.46 -0.49 1 0.91 0.46 -0.44 0.49 -0.38
TAX 0.58 -0.31 0.72 -0.036 0.67 -0.29 0.51 -0.53 0.91 1 0.46 -0.44 0.54 -0.47
PT 0.29 -0.39 0.38 -0.12 0.19 -0.36 0.26 -0.23 0.46 0.46 1 -0.18 0.37 -0.51
B -0.39 0.18 -0.36 0.049 -0.38 0.13 -0.27 0.29 -0.44 -0.44 -0.18 1 -0.37 0.33
LSTAT 0.46 -0.41 0.6 -0.054 0.59 -0.61 0.6 -0.5 0.49 0.54 0.37 -0.37 1 -0.74
MV -0.39 0.36 -0.48 0.18 -0.43 0.7 -0.38 0.25 -0.38 -0.47 -0.51 0.33 -0.74 1
CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX PT B LSTAT MV
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4 −0.6