INTRODUCTION TO MACHINE LEARNING PROJECT

LECTURER: DR. MUHAMMED DAVUD

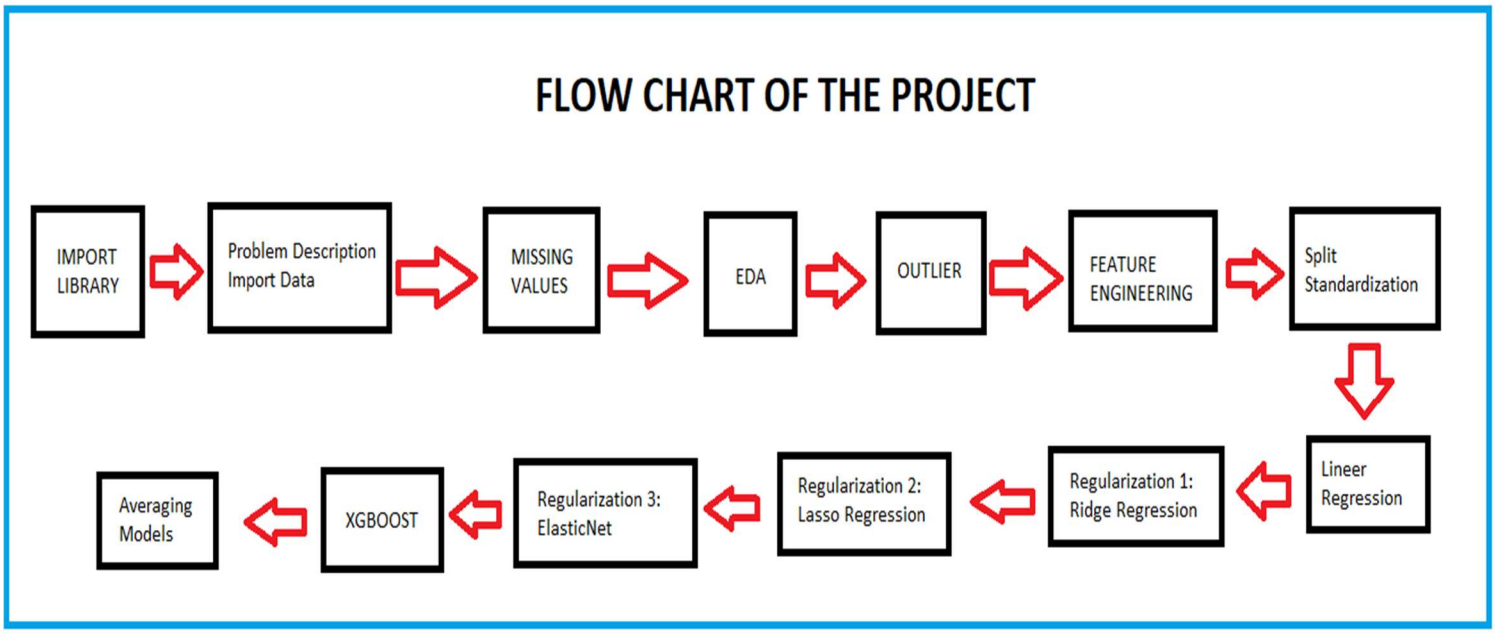
STUDENT: MEHMET ŞENER – 030717050

DEPARTMENT: SOFTWARE ENGINEERING

EXPLANATION OF THE PROJECT

I did a project in introduction to machine learning course , which is Fuel consumption prediction of the vehicles. In my Project, I wanted to use Python programming language because of fact that it gives us many advantages.

I did a lot of operation for my project as a respectively ;



DATASET INFORMATION

This dataset is a slightly modified version of the dataset provided in the StatLib library. In line with the use by Ross Quinlan (1993) in predicting the attribute "mpg", 8 of the original instances were removed because they had unknown values for the "mpg" attribute. The original dataset is available in the file "auto-mpg.data-original".

"The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes." (Quinlan, 1993)

Attribute Information of the Dataset

1 - mpg : continuous (MPG , which means mile per galon)

2 - cylinders : multi-valued discrete

3 - displacement : continuous

4 - horsepower : continuous

5 -weight : continuous

6 - acceleration : continuous

7 - model-year : multi-valued discrete

8 - origin : multi-valued discrete

9 - car name : string (unique for each instance)

DATASET LINK

https://archive.ics.uci.edu/ml/datasets/Auto+MPG

Libraries I Use in this Project;

- Numpy
- Seaborn
- Pandas
- Matplotlib
- Scipy
- Scipy.stats
- Sklearn
- Xgboost

STEPS OF THE PROJECTS

1 – Imputing Missing Value : In the dataset , Horsepower column has 6 Nan . I have been changed the missing values with mean.

```
RangeIndex: 398 entries, 0 to 397
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   target        398 non-null    float64
 1   Cylinders     398 non-null    int64
 2   Displacement  398 non-null    float64
 3   Horsepower    392 non-null    float64
 4   Weight        398 non-null    float64
 5   Acceleration  398 non-null    float64
 6   Model Year    398 non-null    int64
 7   Origin        398 non-null    int64
```

2 – EDA : Expolaratory Data Analysize

3- Outlier detection and removel.

4 – Feature Engineering : Skewness

5 – Feature Engineering : One hot encoding

6 – Prepprocess : Training / Testing Separation and Standardization

7 – Lineer Regression

8 – Regularization 1 : Ridge Regression

9 – Regularization 2 : Lasso Regression

10 – Regularization 3:

ElasticNet

Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve on the regularization of statistical models.

11 – XGBOOST Algorithm :

XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

12 – Averaging Models

GRAHPS OF THE OPERATIONS IN THE PROJECT

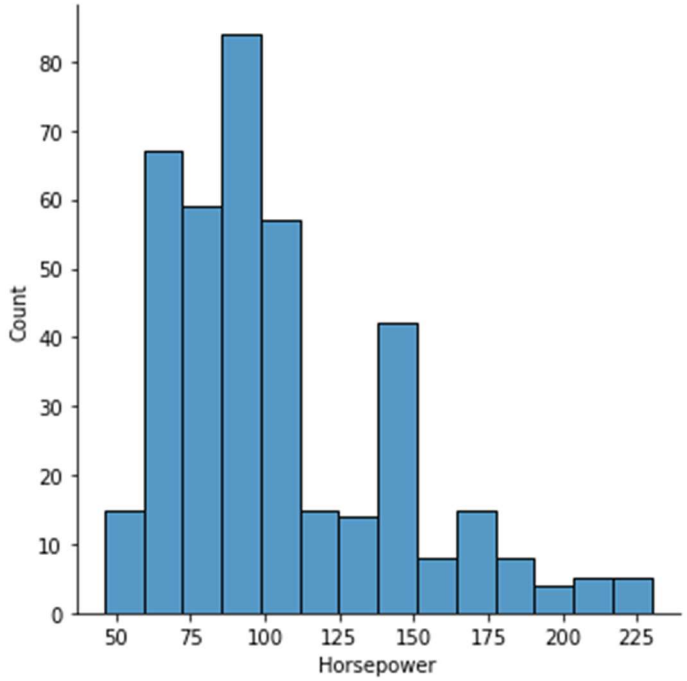


FIGURE 1

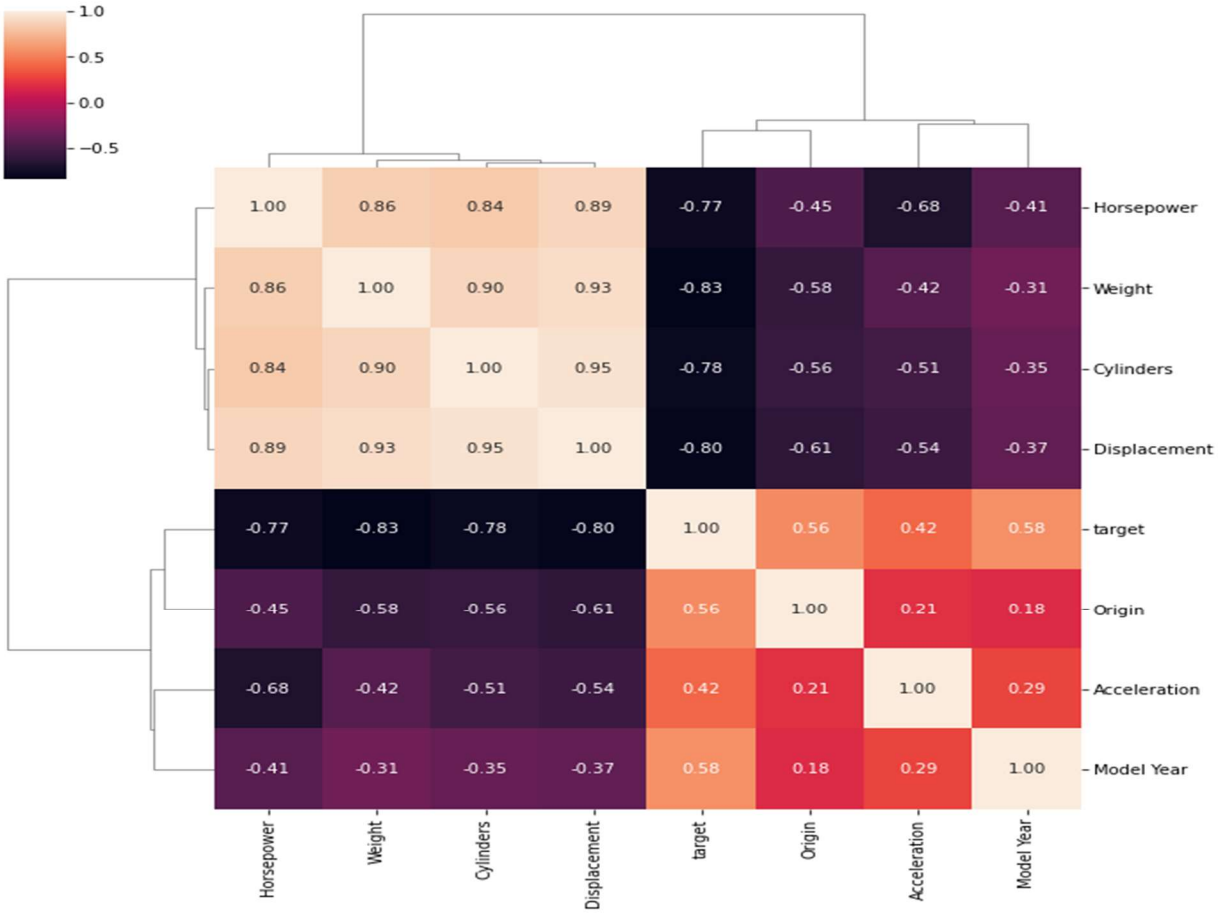


FIGURE 2

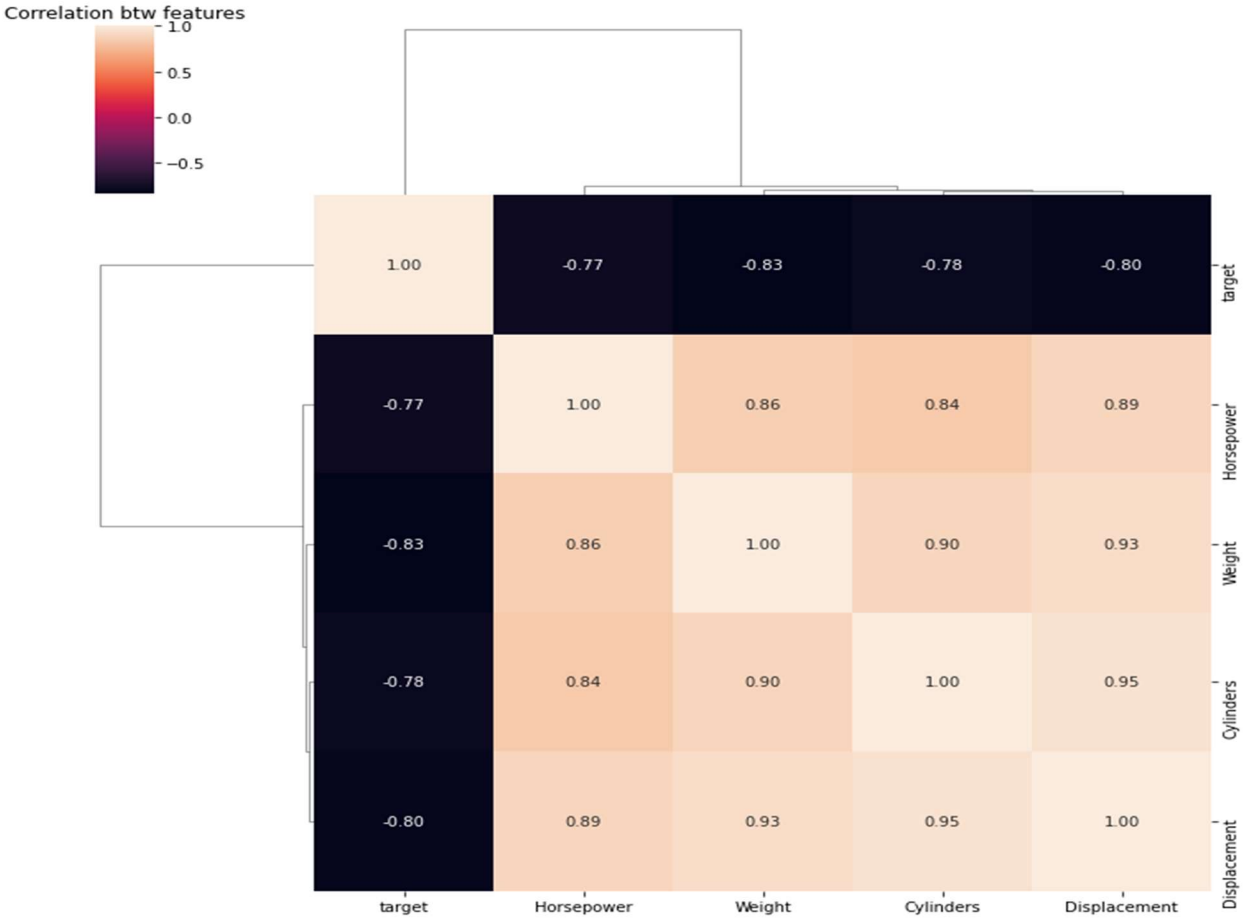

Correlation btw features
1.0
0.5
0.0
−0.5
1.00
-0.77
-0.83
-0.78
-0.80
-0.77
1.00
0.86
0.84
0.89
-0.83
0.86
1.00
0.90
0.93
-0.78
0.84
0.90
1.00
0.95
-0.80
0.89
0.93
0.95
1.00
target
Horsepower
Weight
Cylinders
Displacement
target
Horsepower
Weight
Cylinders
Displacement

FIGURE 3

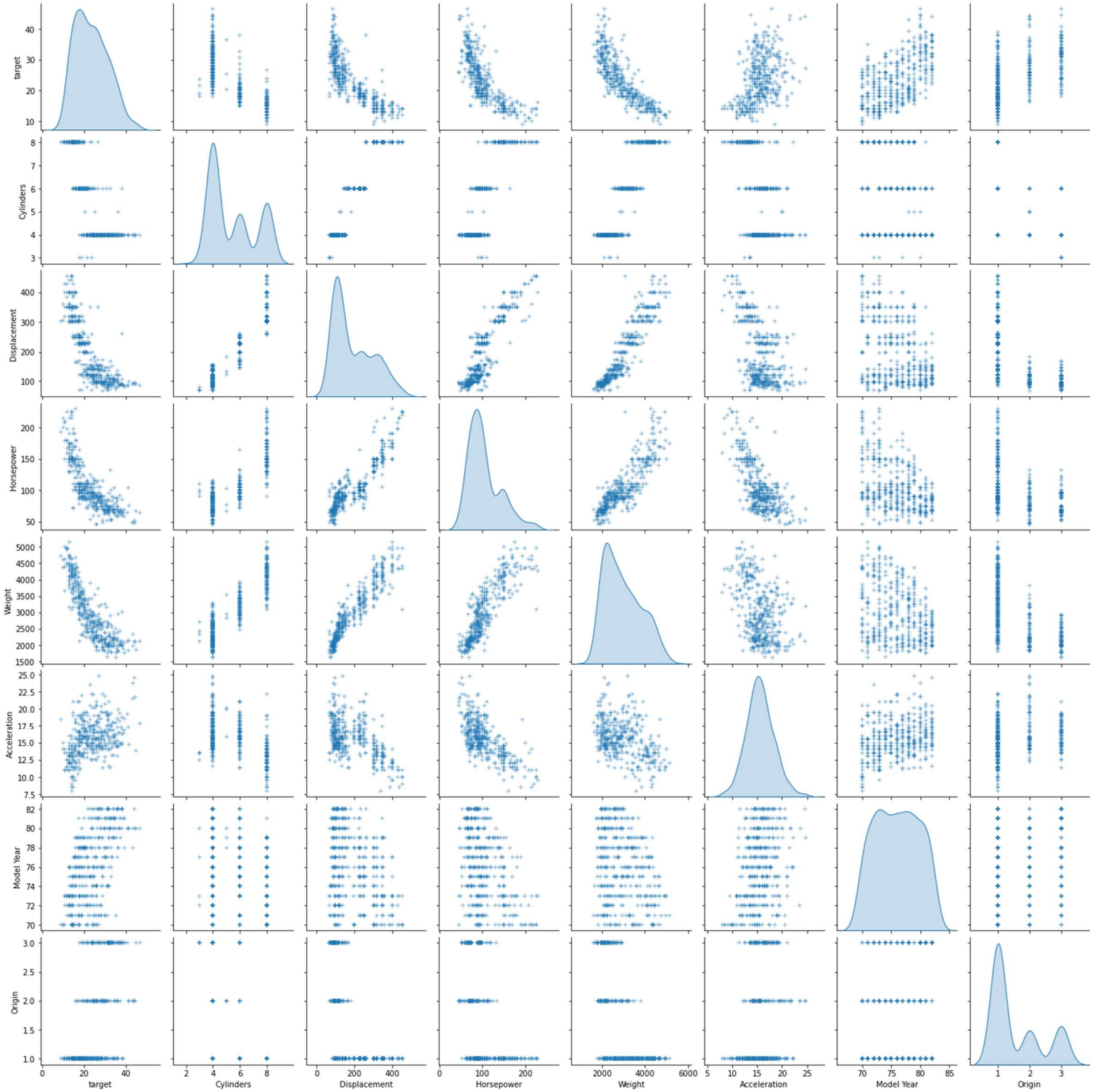

target
Cylinders
Displacement
Horsepower
Weight
Acceleration
Model Year
Origin

FIGURE 4

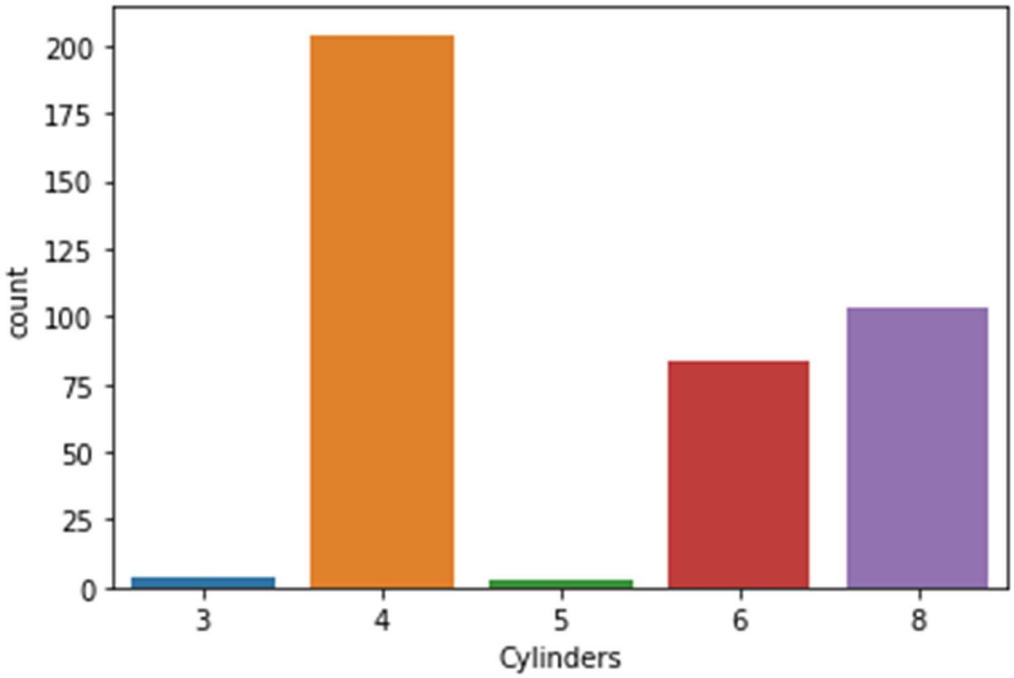


FIGURE 5

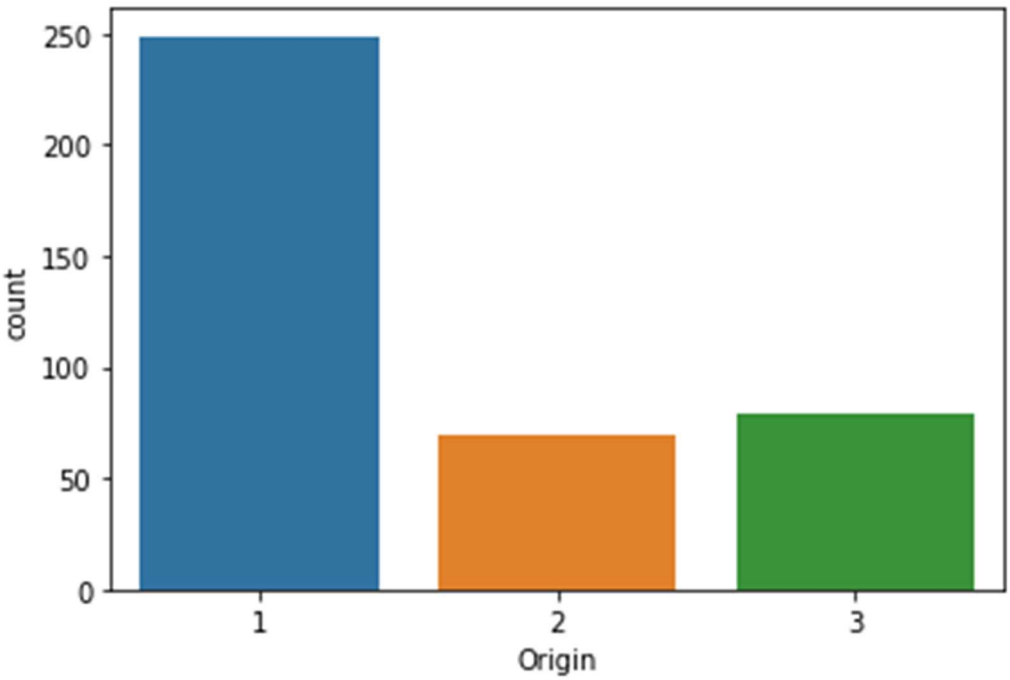


FIGURE 6

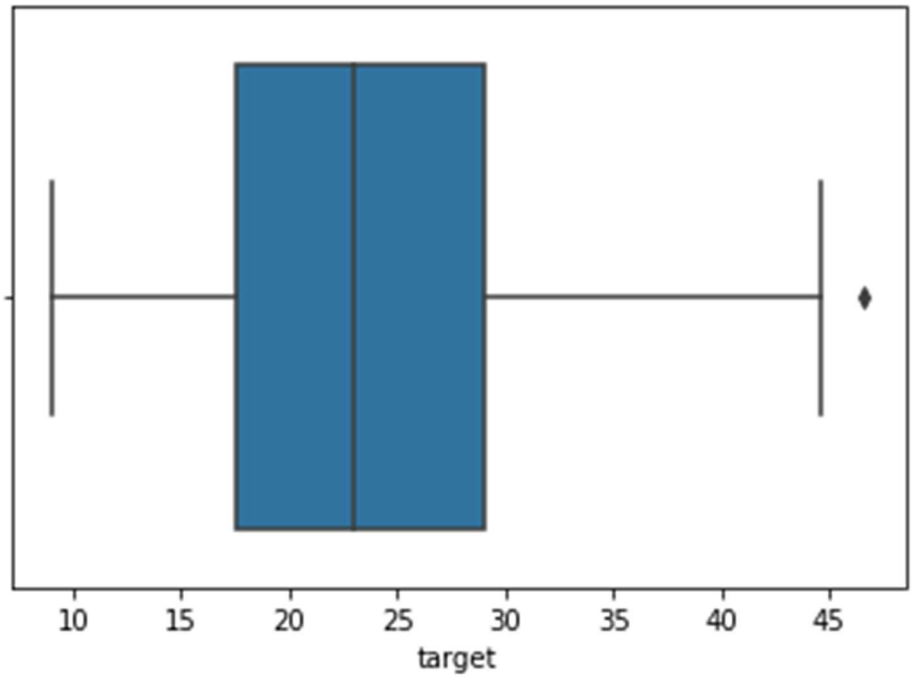


FIGURE 7

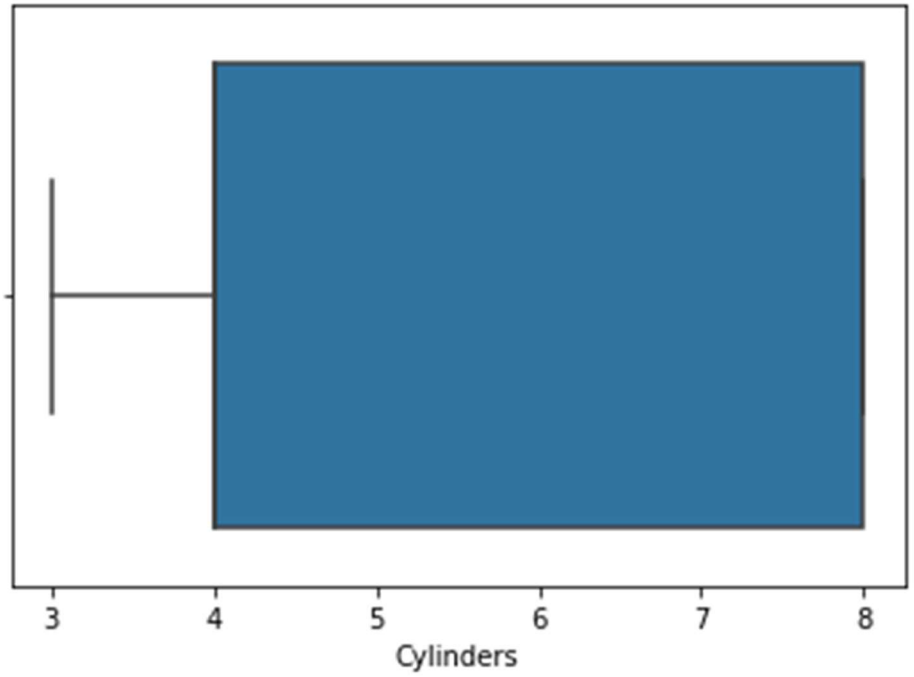


FIGURE 8

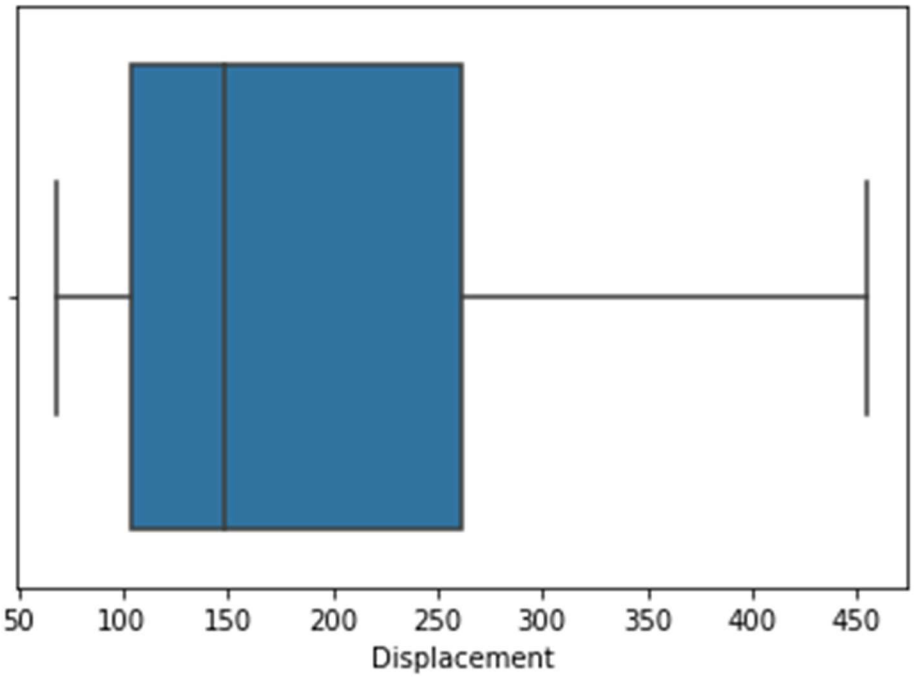


FIGURE 9

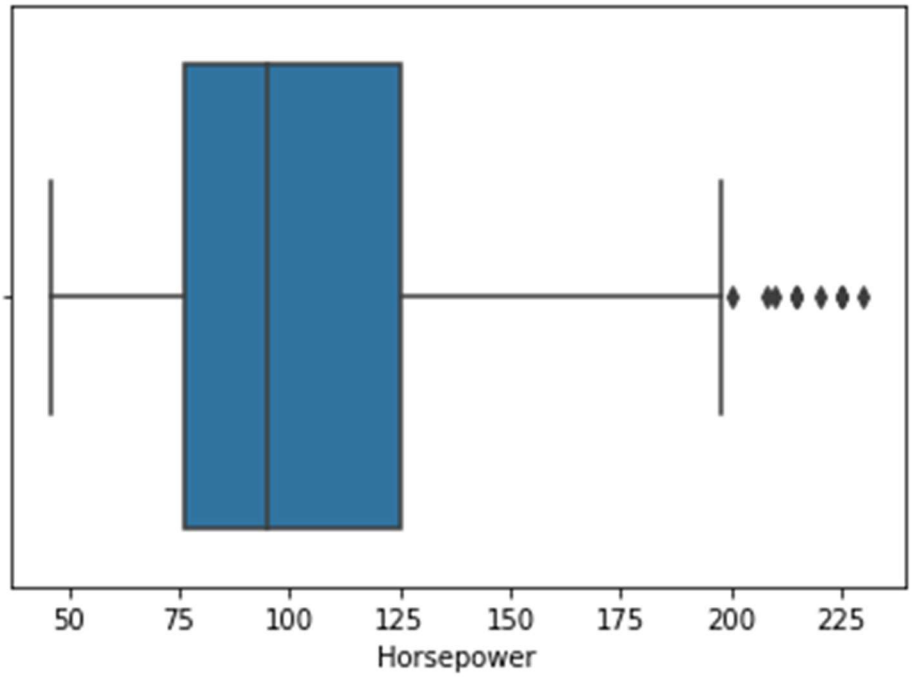


FIGURE 10

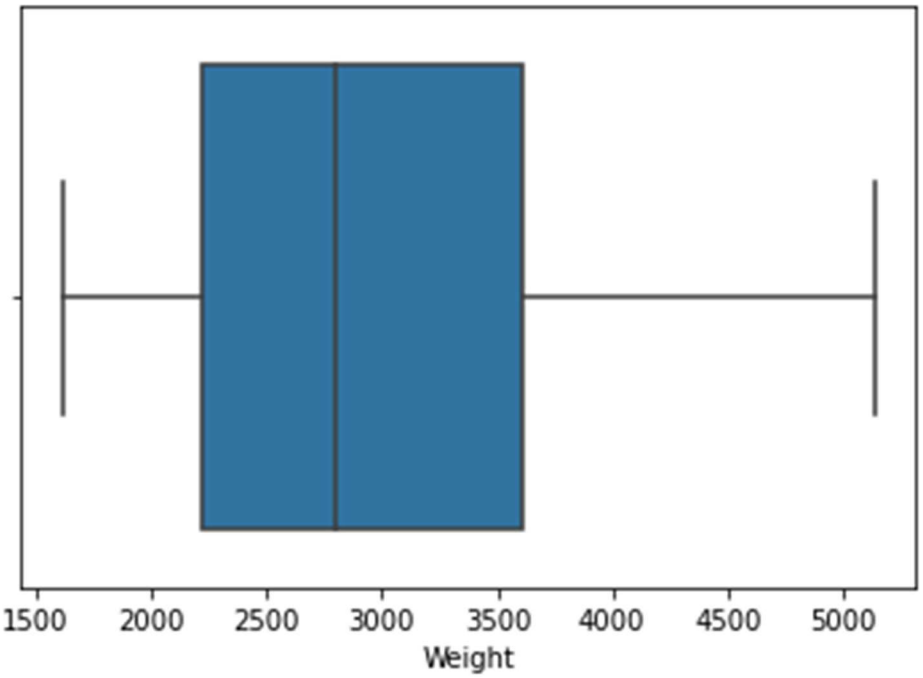


FIGURE 11

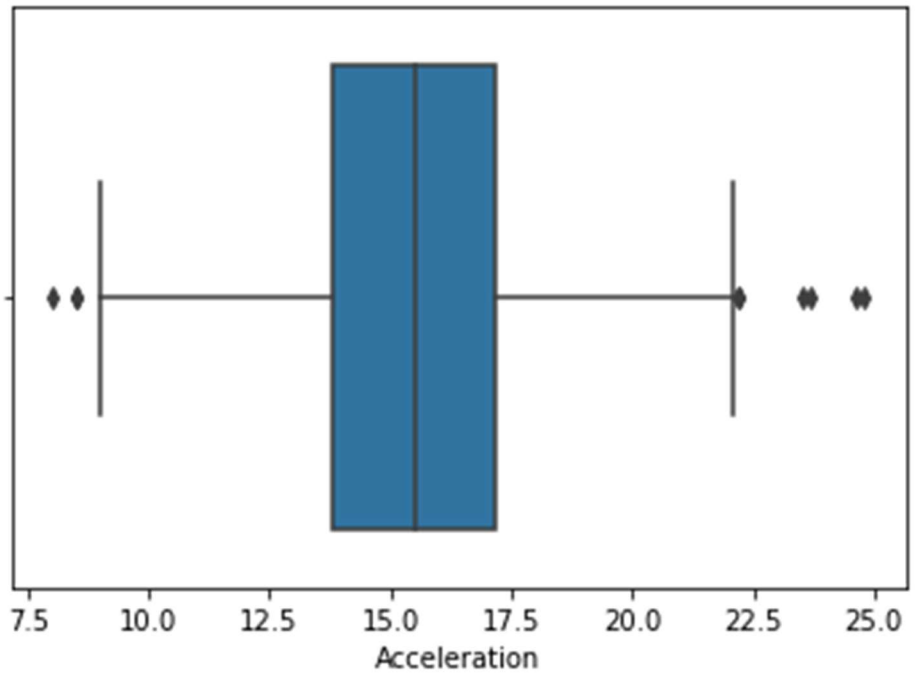


FIGURE 12

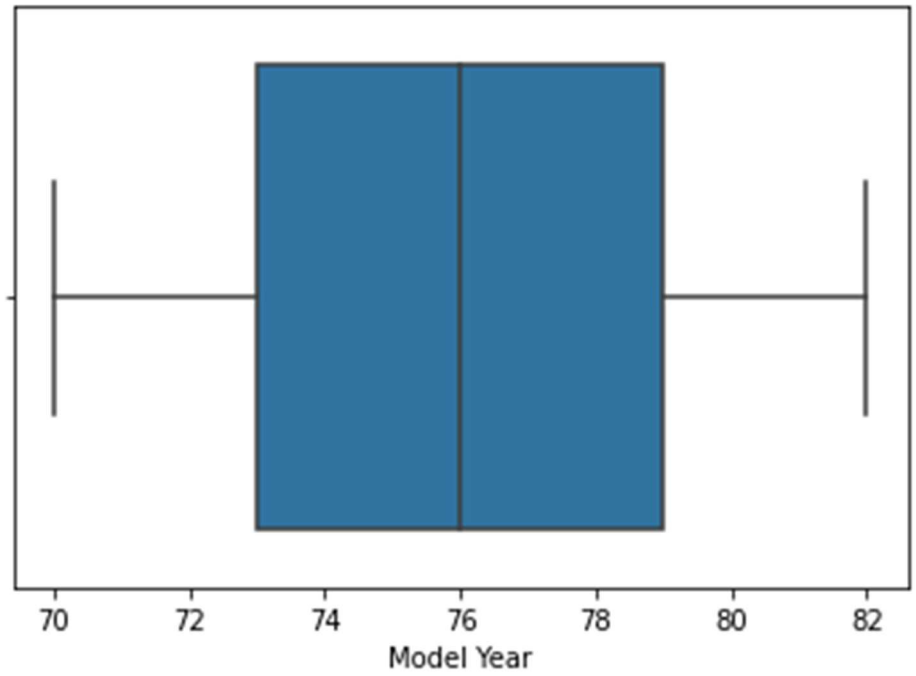


FIGURE 13

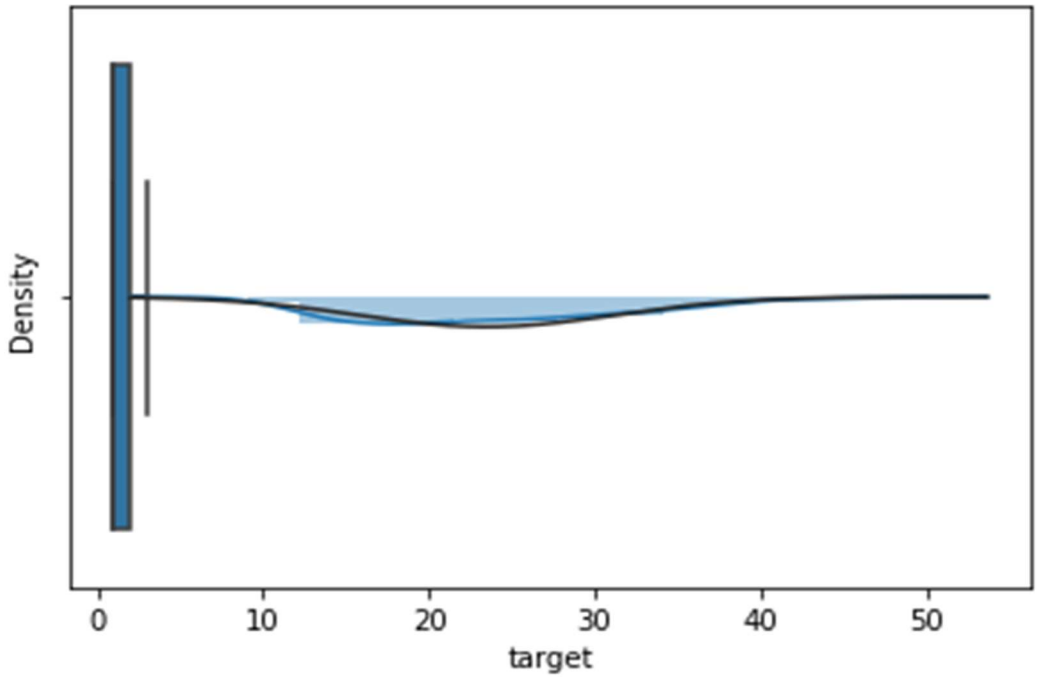


FIGURE 14

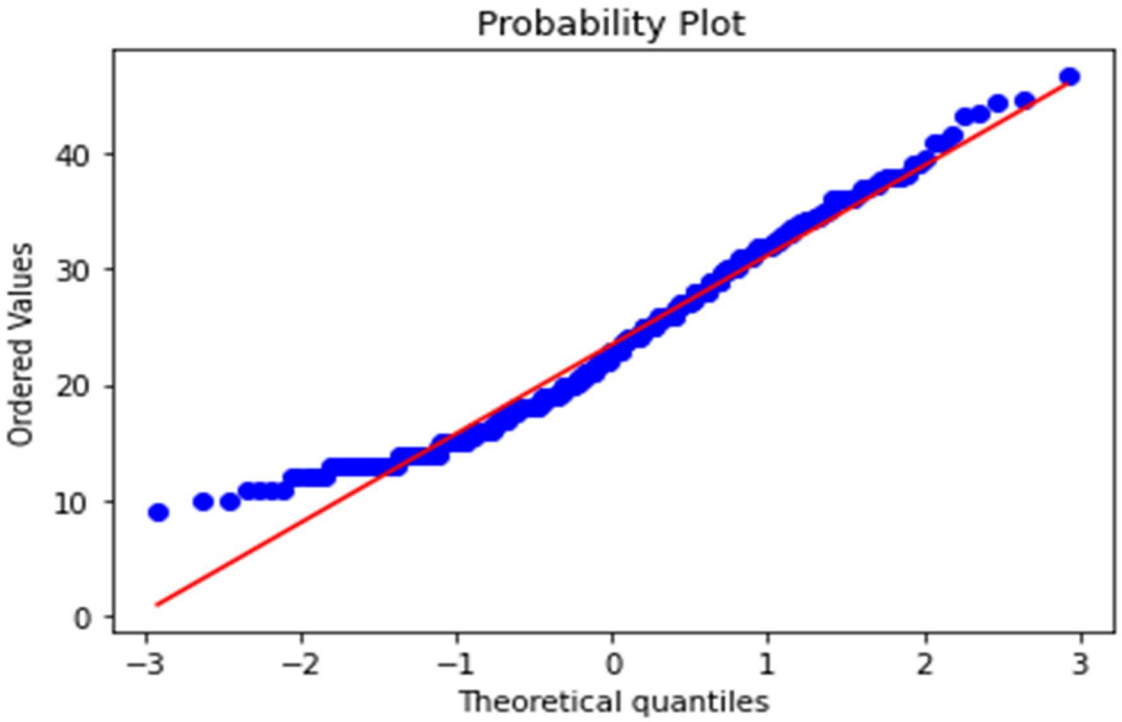


FIGURE 15

```
LR Coef:  [-1.52652234e-01 -1.10059003e-01 -3.51567523e-02 -7.81023601e-02
  7.76057988e-02 -2.71614700e-01  1.86827885e-01  1.64798730e-17
  5.74192760e-02  2.73675388e-02 -1.43431578e-02 -6.77254422e-02
  8.20686000e-02]
Linear Regression MSE:  0.020984711065869636
Ridge Coef:  [-0.07617499 -0.10434789 -0.0756786  -0.06362033  0.08849698 -0.17538752
  0.17840625  0.          0.02717488 -0.03019362 -0.02334111 -0.05192496
  0.07526607]
Ridge Best Estimator:  Ridge(alpha=0.31622776601683794, max_iter=10000, random_state=42)
Ridge MSE:  0.018839299330570596
```

```
------------------------------------------------------------
Lasso Coef:  [-0.01692687 -0.10976505 -0.11721736 -0.03064576  0.09866154 -0.01243765
  0.16495225  0.          0.00378698 -0.00505995 -0.         -0.
  0.07376033]
Lasso Best Estimator:  Lasso(alpha=0.0037065129109221566, max_iter=10000, random_state=42)
Lasso MSE:  0.016597127172690827
------------------------------------------------------------
ElasticNet Coef:  [-0.0518437  -0.10923982 -0.09538543 -0.03946691  0.09541227 -0.07791667
  0.14374215  0.          0.         -0.01852327 -0.         -0.00200172
  0.0793579 ]
ElasticNet Best Estimator:  ElasticNet(alpha=0.014873521072935119, l1_ratio=0.15000000000000002,
           max_iter=10000, random_state=42)
ElasticNet MSE:  0.017234676963922276
Fitting 5 folds for each of 18 candidates, totalling 90 fits
[01:06:21] WARNING: d:\bld\xgboost-split_1615294821523\work\src\objective\regression_obj.cu:170:
reg:linear is now deprecated in favor of reg:squarederror.
[01:06:21] WARNING: ..\src\learner.cc:541:
Parameters: { silent } might not be used.

  This may not be accurate due to some parameters are only used in language bindings but
  passed down to XGBoost core.  Or some parameters are not used but slip through this
  verification. Please open an issue if you find above cases.
```

```
Averaged Models MSE:  0.015753524420134283
```

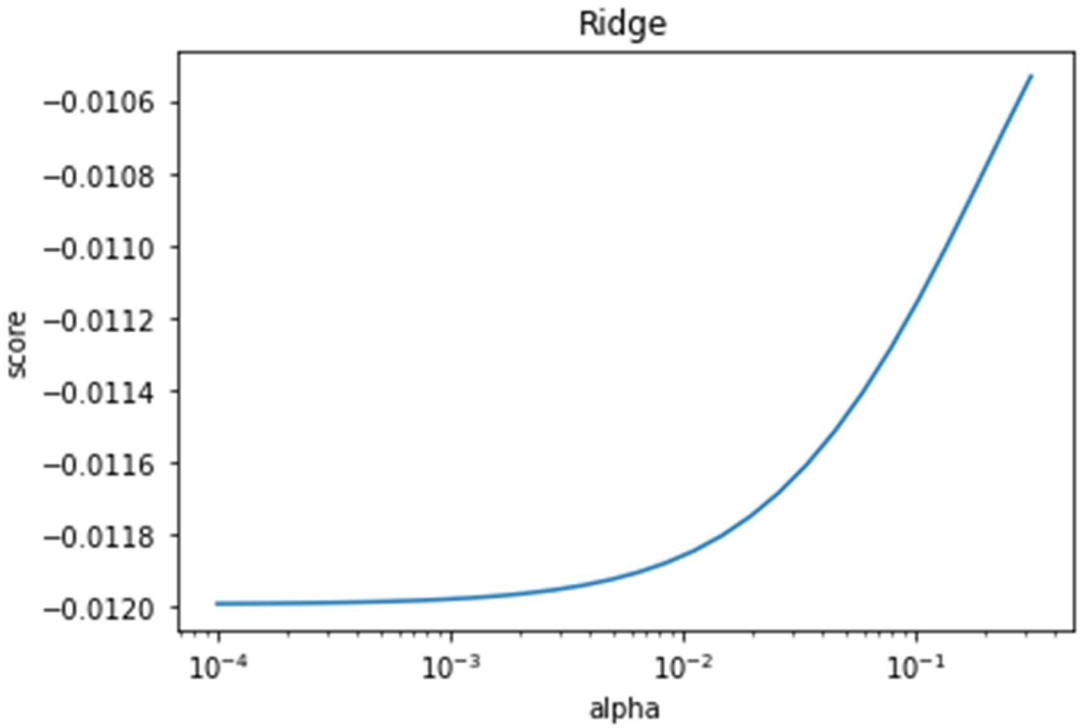


FIGURE 16

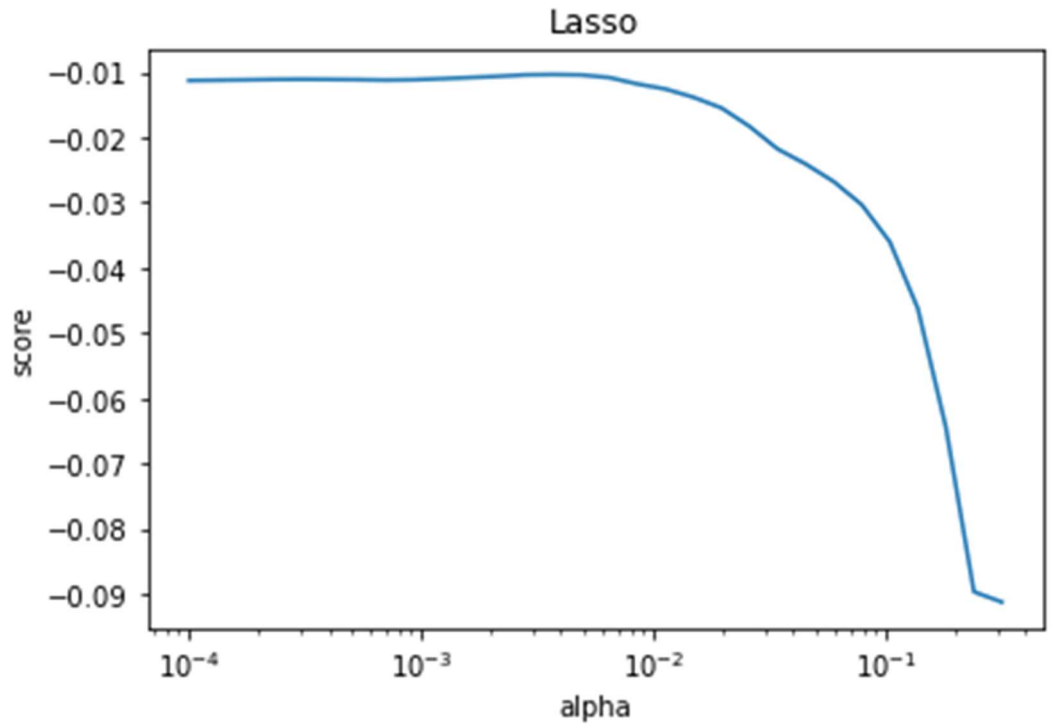


FIGURE 17

FINAL

StandardScaler:

Linear Regression MSE: 0.020632204780133015

Ridge MSE: 0.019725338010801216

Lasso MSE: 0.017521594770822522

ElasticNet MSE: 0.01749609249317252

XGBRegressor MSE: 0.017167257713690008

Averaged Models MSE: 0.016034769734972223

RobustScaler:

Linear Regression MSE: 0.020984711065869643

Ridge MSE: 0.018839299330570554

Lasso MSE: 0.016597127172690837

ElasticNet MSE: 0.017234676963922273

XGBRegressor MSE: 0.01753270469361755

Averaged Models MSE: 0.0156928574668921