Rental Price Prediction of four neighborhoods in Istanbul

Feda Abdulrazzaq Yahya Alashwal – 2018965

Omnia Mohamed Mahmoud Sedeek Elmenshawy - 2000007

Bahcesehir University
Artificial Intelligence Department
Istanbul, Turkey

Abstract

Rental prices differ based on some variables. In our project, we have tried to predict the prices of different types of properties, neighborhoods, sizes, status, and available rooms.
We have collected a (14x176) dataset manually from a Turkish Website in the name of Sahibinden [1] along with getting the latitude and longitude from the LatLong website [2]. This project studies multiple regression supervised machine learning models in a prediction task applied after analyzing and cleaning the collected dataset.

Keywords: Prediction, Regression, Prices, Analysis, Supervised Machine Learning

1. Introduction

This report explains the results of experimenting regression models on a property rental price dataset based in four different neighborhoods in Istanbul, Turkey, each neighborhood has its own prices based on location, and view.
Our project is based on four main tasks as follows:

- Data Collection and feature creation
- Data analysis and visualization, where we analyzed the data statistically to understand the prices, correlation, and distribution.
- Data pre-processing, where we manipulated the missing values, outliers, skewness, standardization, and encoding of the data.
- Data modeling, where we applied nine regression models to our dataset and experimented them with different test sizes and multiple parameters in order to get the best model evaluation.

2. Dataset Description

2.1. Data collection and feature creation

We have collected a dataset of 14 columns and 176 row, the information of the created features are as follows:

	Neighborhood	Latitude	Longitude	Type	Building Avg. Age	Rooms + Salon	Bathrooms	Sea View	Furnished	Net Area m^2	Balcony	Compund	Floor Num.	Price
0	uskudar	41.073	29.057	villa	0	6	4.0	yes	no	450	yes	NaN	0	390000
1	beylikduzu	40.969	28.599	dubleks	8	6	3.0	yes	no	180	yes	no	4	7500
2	beylikduzu	41.004	28.625	apartment	13	5	2.0	no	yes	174	yes	yes	16	10000

- Neighbourhood: It is a column that contains the names of the neighbourhoods which we decided to collect data from, the neighbourhoods are: Uskudar, Beylikduzu, Pendik, and Kagithane.
- Latitude: The Latitude coordinate based on the property given location.
- Longitude: The Longitude coordinate based on the property given location.
- Type: collected property types varies as follows: Residence, Villa, Apartment, Dubleks.
- Building Avg. Age: the average age of the collected property building.
- Rooms + Salon: The sum number of bedrooms and living rooms in the apartment.
- Bathrooms: the sum number of restrooms and bathrooms in the apartments.
- Sea View: A column clarifies whether the apartment has a sea view or not.
- Furnished: A column clarifies whether the apartment is furnished or not.
- Net Area m^2: The net area of the apartment in meters square.
- Balcony: whether the apartment has a balcony or not.
- Compund: whether the apartment is in a residential compound or not.
- Floor Num.: The floor number of the property.
- Price: The renting price of each property, and it is the prediction target column.

2.2. Data Analysis and visualization

The summary of the analytical description of the collected dataset shows that the price column is skewed and has outliers. Also, there are two missing cells in the Bathroom column, and the Compound column which forms 0.1% of the column attributes, so there is no need to fill them in the pre-processing task while we can easily drop them. Furthermore, we have no duplicated rows, and the correlating between some features and the price are strong such as the Rooms+Salon, bathrooms, Net Area, and Building average age, however, the floor number column has no effect on the price.

The full analysis are shown in the following graphs:

	count	mean	std	min	25%	50%	75%	max
Latitude	175.000000	20408.232840	35947.680524	40.873000	40.997000	41.080000	41010.500000	410217.000000
Longitude	175.000000	12985.140520	14490.025733	28.599000	28.960000	29.259000	29041.500000	29305.000000
Building Avg. Age	175.000000	10.200000	10.395622	0.000000	2.000000	7.000000	15.000000	35.000000
Rooms + Salon	175.000000	3.468571	1.465263	1.000000	2.500000	3.000000	4.000000	10.000000
Bathrooms	174.000000	1.500000	0.750722	1.000000	1.000000	1.000000	2.000000	4.000000
Net Area m^2	175.000000	110.062857	79.827195	20.000000	65.000000	86.000000	131.500000	800.000000
Floor Num.	175.000000	4.994286	4.921799	0.000000	2.000000	4.000000	6.000000	34.000000
Price	175.000000	10610.057143	29529.763117	2500.000000	4800.000000	7000.000000	9900.000000	390000.000000

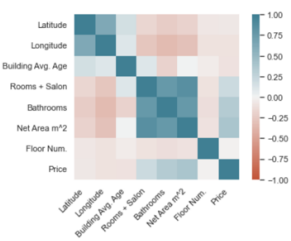


Figure 1: Statistical Analysis Summary and correlation matrix

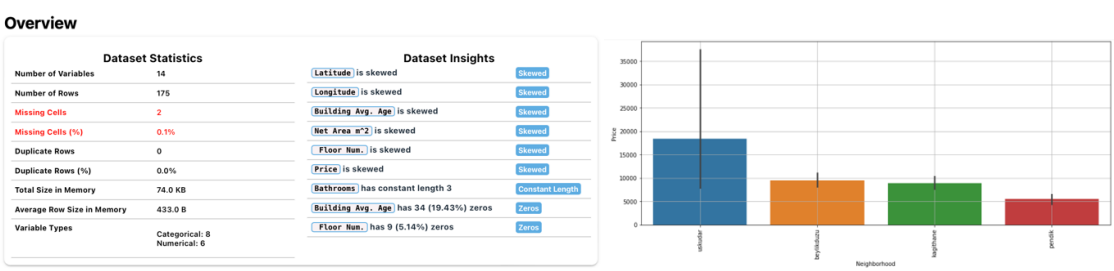


Figure 2: Data overview and Price Boxplots based on neighborhoods

Figure 3: target column analysis and missing values visualization

3. Methods

3.1. Data Pre-processing

In the pre-processing task, The missing values were dropped from the dataset as the statistical analysis show that they form 0.1% of it, and the outliers in the target columns were also dropped based on the following visualization, where all of the variables above 26K has been dropped:

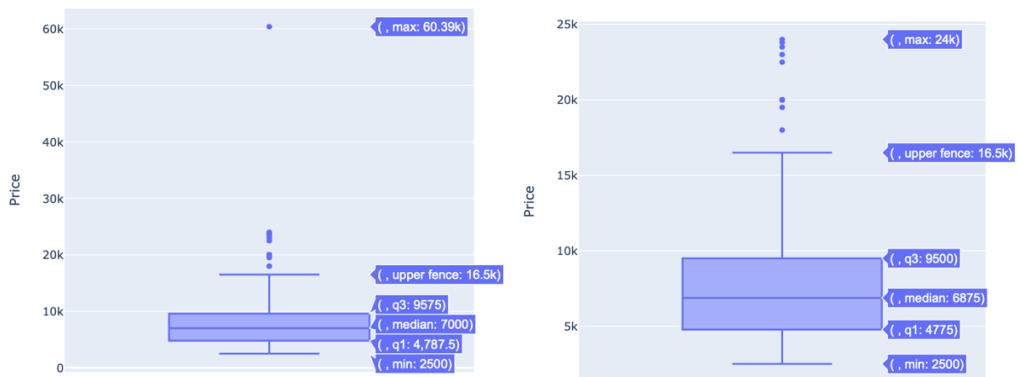


Figure 4: Price column boxplot before removing outliers and after removing them.

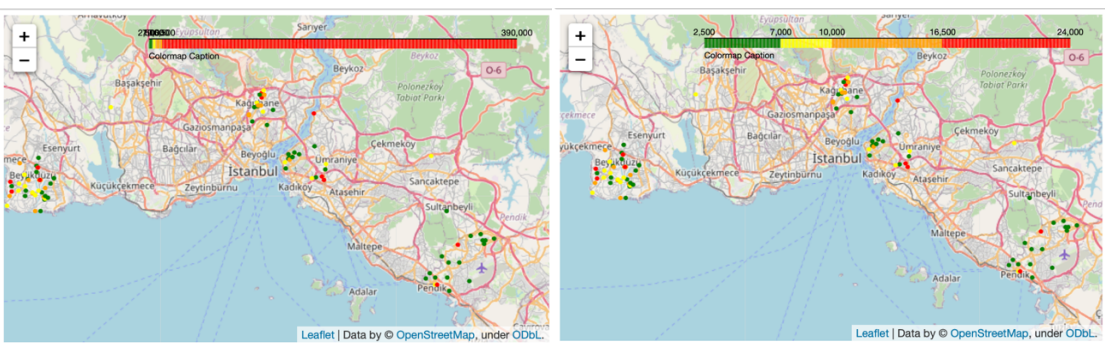
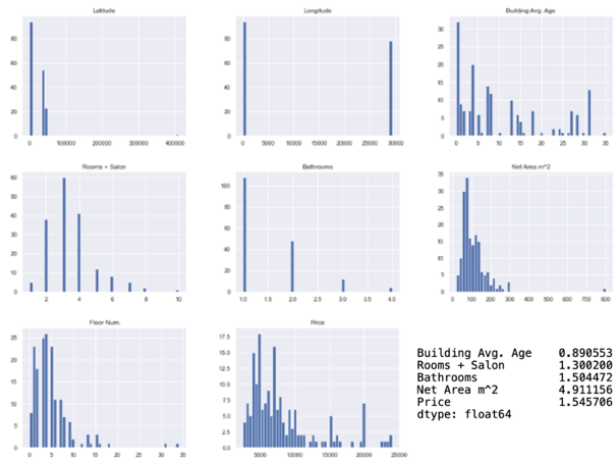


Figure 5: Istanbul price Map before and after dropping outliers

Furthermore, The skewness in the dataset is fixed according to the Boxcox method, only the Price, Rooms+Salon, Bathrooms, and Net Area columns skewness have been handled according their high skewness as follows:



```
Building Avg. Age    0.890553
Rooms + Salon        1.300200
Bathrooms            1.504472
Net Area m^2         4.911156
Price                1.545706
dtype: float64
```

Figure 6: Skewness numerical value and visualization.

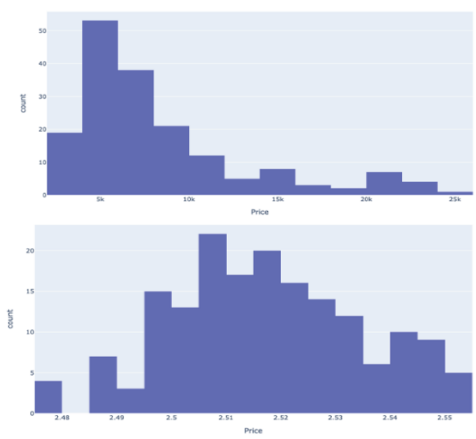


Figure 7: Price column Skewness before and after applying Boxcox. Method

Another task performed in the pre-processing is the encoding, we have applied Label Encoder to transform all categorical values into numerical ones in order to make our regression model able to read and learn from the data as follows:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 175 entries, 0 to 174
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Neighborhood       175 non-null    object
 1   Latitude           175 non-null    float64
 2   Longitude          175 non-null    float64
 3   Type               175 non-null    object
 4   Building Avg. Age  175 non-null    int64
 5   Rooms + Salon      175 non-null    int64
 6   Bathrooms          174 non-null    float64
 7   Sea View           175 non-null    object
 8   Furnished          175 non-null    object
 9   Net Area m^2       175 non-null    int64
 10  Balcony            175 non-null    object
 11  Compund            174 non-null    object
 12   Floor Num.        175 non-null    int64
 13  Price              175 non-null    int64
dtypes: float64(3), int64(5), object(6)
memory usage: 19.3+ KB
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 172 entries, 1 to 174
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Neighborhood       172 non-null    int64
 1   Latitude           172 non-null    float64
 2   Longitude          172 non-null    float64
 3   Type               172 non-null    int64
 4   Building Avg. Age  172 non-null    int64
 5   Rooms + Salon      172 non-null    float64
 6   Bathrooms          172 non-null    float64
 7   Furnished          172 non-null    int64
 8   Net Area m^2       172 non-null    float64
 9   Balcony            172 non-null    int64
 10  Compund            172 non-null    int64
 11   Floor Num.        172 non-null    int64
 12  Price              172 non-null    float64
 13  Sea_View           172 non-null    int64
dtypes: float64(6), int64(8)
memory usage: 20.2 KB
```

Figure 8: Data taype of each column before and after label encoder

In the last task of pre-processing, we applied Robust Scaler to rescale the data, then we selected the highly correlated features to include in the modelling based on the following correlation map as a filter method:

```
array([[-1.00000000e+00, -2.75811556e-06, -2.32257300e-05, ...,
         0.00000000e+00,  1.23834471e-01,  1.00000000e+00],
       [-1.00000000e+00, -1.90383198e-06, -2.23297819e-05, ...,
         3.00000000e+00,  5.03317112e-01,  0.00000000e+00],
       [-6.00000000e-01,  3.41713432e-07, -1.01311048e-05, ...,
        -5.00000000e-01,  1.25906551e+00,  0.00000000e+00],
       ...,
       [ 6.00000000e-01,  1.00016881e+00,  9.99386172e-01, ...,
        -5.00000000e-01, -3.36420869e-01,  0.00000000e+00],
       [ 6.00000000e-01,  9.99851501e-01,  1.00014428e+00, ...,
         2.50000000e-01, -2.01579822e-01,  0.00000000e+00],
       [ 6.00000000e-01,  1.00019321e+00,  9.98903739e-01, ...,
         0.00000000e+00,  3.69296636e-01,  0.00000000e+00]])
```

Figure 9: Data after Robust Scaller

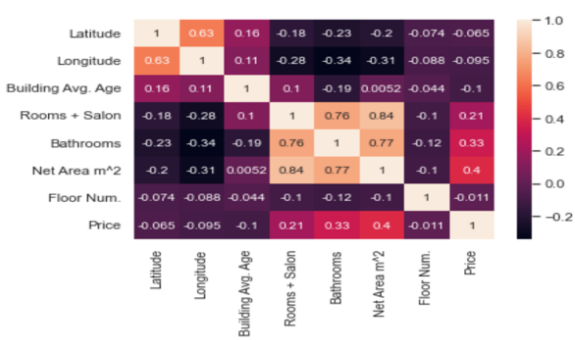


Figure 10: Numerical Correlation Matrix

So, the columns included in the modelling training and testing datasets are:

	Neighborhood	Type	Building Avg. Age	Rooms + Salon	Bathrooms	Furnished	Net Area m^2	Balcony	Compund	Price	Sea_View
1	0	1	8	2.018891	0.435286	0	3.300048	1	0	2.464551	1

3.2. Data modelling methods and algorithms

In the prediction task, we have applied nine regression models, tested them with two test sizes and different random state test, and changed the parameter of each model to get the best predictions possible, also, we have used multiple methods to evaluate our model prediction as follows:

- Models Used:
  - Linear regression
  - K Neighbors Regressor
  - Decision Tree Regressor
  - Random Forest regressor
  - Ada Boost Regressor
  - Gradient Boosting Regressor
  - XGBoost Regressor
  - Bayesian Ridge Regressor
  - Cat Boost Regressor

- Evaluation Methods Used:
  - Mean absolute percentage error
  - Mean Absolute Error
  - Mean squared error
  - Root mean square error
  - Coefficient of determination score
  - Actual vs Predicted values Graph
  - Linear Line Graph

4. Experimental Results

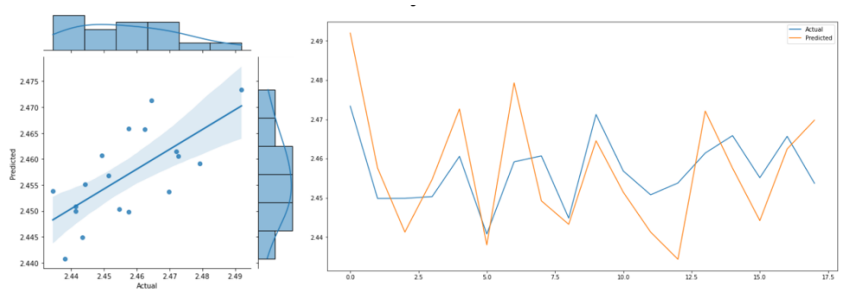
Here, we present each experiment done on each model:

- ***Experiment 1:***

  Random State = 40 and Test Size equals to 10% of the data, x_train shape is (154,10) and has a size of 1540, and x_test shape is (18,10) and has a size of 154.

  - **Model 1: Linear regression**

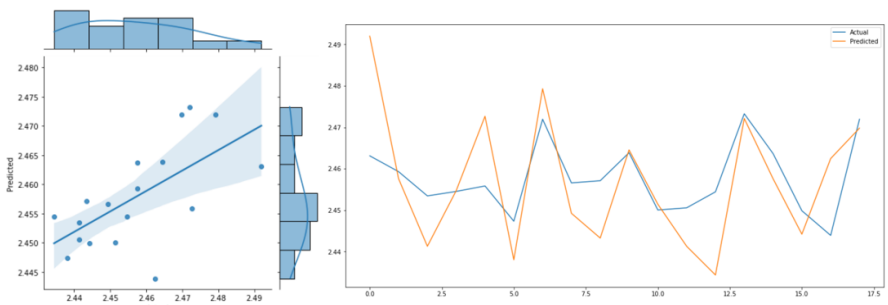
```
Mean Absolute Percentage Error is: 0.004007876915234344
MAE: 0.009861901227877655
MSE: 0.00012796992746877915
RMSE: 0.011312379390242319
Coefficient of determination score on training :  0.5406369066890198
Coefficient of determination score on testing :  0.45916534506161355
```



- **Model 2: K Neighbours Regressor**

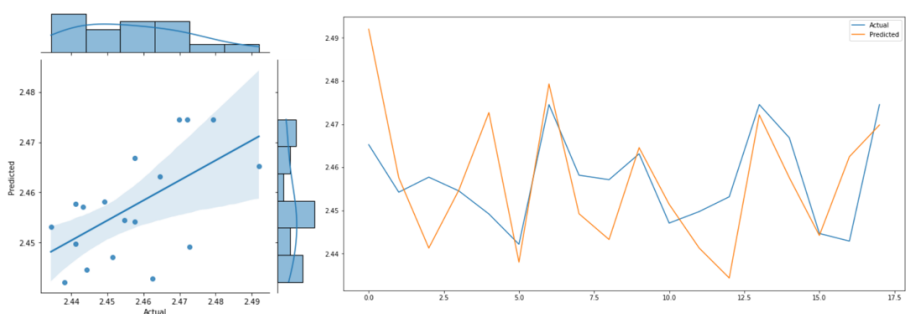
1) N_neighbors parameter = 7

```
Mean Absolute Percentage Error is: 0.0036723083675517768
MAE: 0.00902901577212366
MSE: 0.00014233347993110107
RMSE: 0.011930359589346042
Coefficient of determination score on training :  0.29888971398807274
Coefficient of determination score on testing :  0.39846118516010454
```



2) N_neighbors parameter = 5

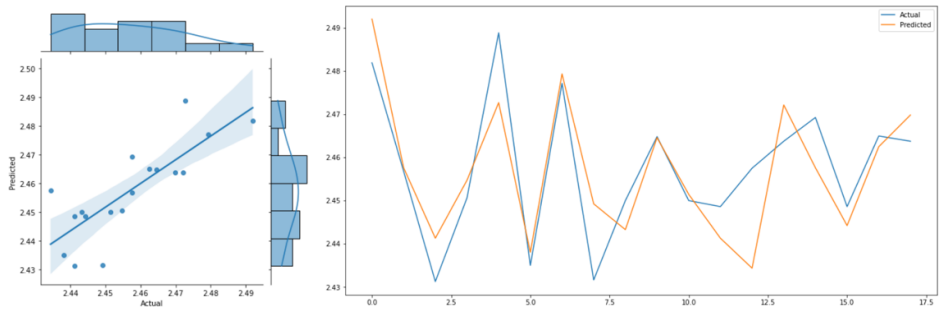
```
Mean Absolute Percentage Error is: 0.003868312090890391
MAE: 0.009513980575559645
MSE: 0.0001555292628290314
RMSE: 0.012471137190690807
Coefficient of determination score on training :  0.4226342492008873
Coefficient of determination score on testing :  0.34269232734008925
```



- **Model 3: Decision Tree Regressor**

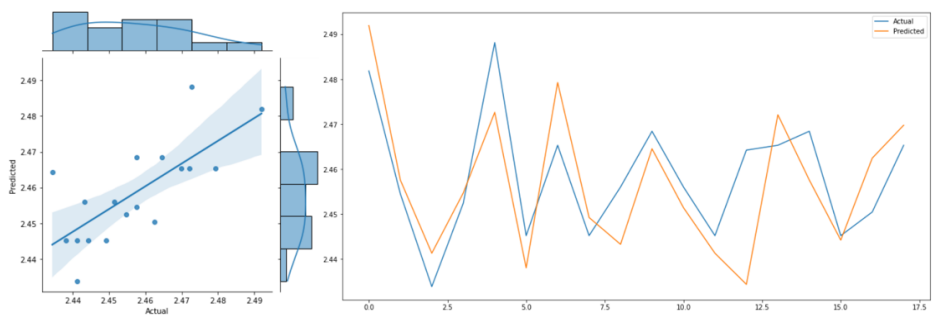
1) Max_depth parameter = 7

```
Mean Absolute Percentage Error is: 0.003073331407037265
MAE: 0.007544659799714859
MSE: 9.489690325902848e-05
RMSE: 0.009741504157933132
Coefficient of determination score on training :  0.9366589843085591
Coefficient of determination score on testing :  0.5989406656392806
```



2) Max_depth parameter = 3

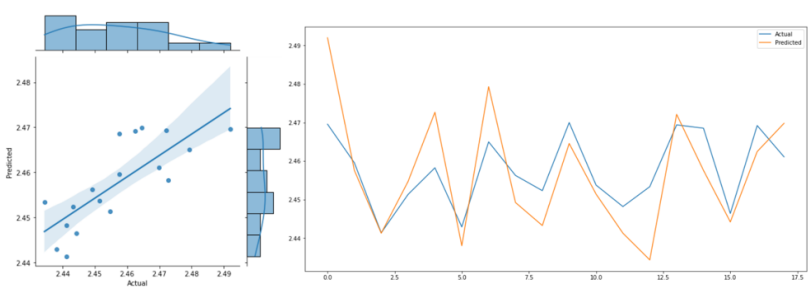
```
Mean Absolute Percentage Error is: 0.0034784029607937634
MAE: 0.008544441727797633
MSE: 0.00011739860663290257
RMSE: 0.010835063757675935
Coefficient of determination score on training :  0.8035668469645441
Coefficient of determination score on testing :  0.5038425342231778
```



- **Model 4: Random Forest regressor**

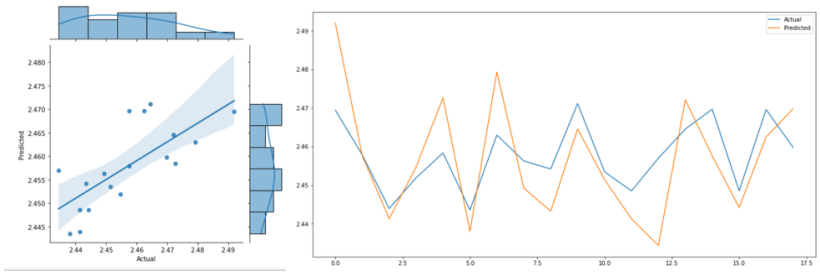
1) Max_depth parameter = 7

```
Mean Absolute Percentage Error is: 0.0032156551717823856
MAE: 0.007916354317539022
MSE: 9.895755439935271e-05
RMSE: 0.009947741170705675
Coefficient of determination score on training :  0.903053525736024
Coefficient of determination score on testing :  0.5817792832602977
```



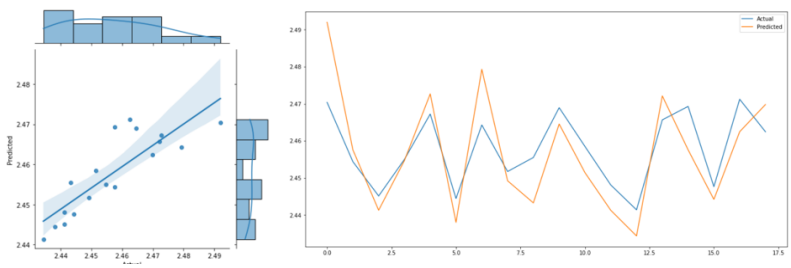
2) Max_depth parameter = 5

```
Mean Absolute Percentage Error is: 0.0036548890966542832
MAE: 0.008994549494292162
MSE: 0.00012094470104609246
RMSE: 0.010997486123932709
Coefficient of determination score on training :  0.808288862784702
Coefficient of determination score on testing :  0.4888558042447272
```



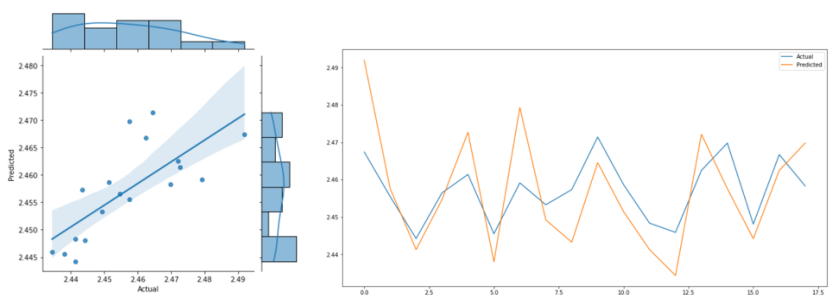
- **Model 5: Ada Boost Regressor**
  1) *learning_rate parameter = 1*

```
Mean Absolute Percentage Error is: 0.0030135023434476804
MAE: 0.007420375891693359
MSE: 7.944229216560225e-05
RMSE: 0.008913040567932037
Coefficient of determination score on training :  0.7806035456881547
Coefficient of determination score on testing :  0.6642559269920647
```



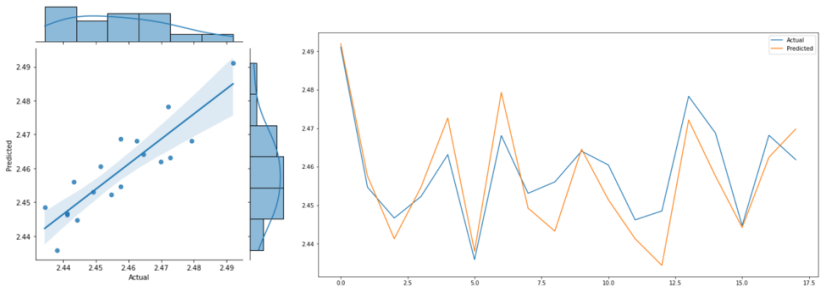
  2) *learning_rate parameter = .2*

```
Mean Absolute Percentage Error is: 0.003666809758566535
MAE: 0.009031017462530199
MSE: 0.00011734789025521937
RMSE: 0.01083272312279878
Coefficient of determination score on training :  0.7448533179978406
Coefficient of determination score on testing :  0.5040568750075054
```



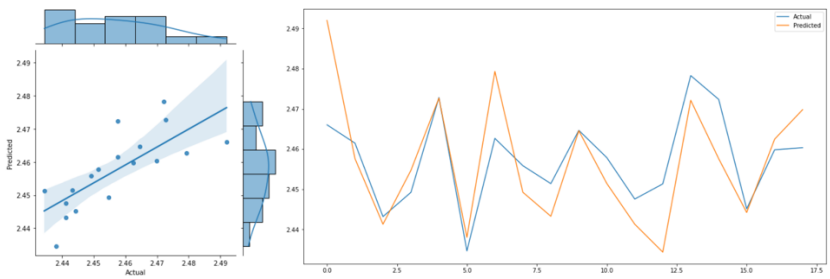
- **Model 6: Gradient Boosting Regressor**
  1) *learning_rate parameter = .2 , max depth = 2*

```
Mean Absolute Percentage Error is: 0.002517027976030031
MAE: 0.006181370032699363
MSE: 5.602094443645468e-05
RMSE: 0.007484714051749384
Coefficient of determination score on training :  0.8950549293763398
Coefficient of determination score on testing :  0.7632407179334824
```



2) learning_rate parameter = .2, max depth = 5

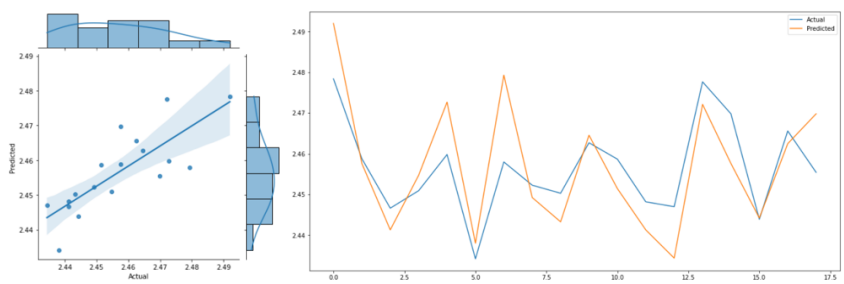
```
Mean Absolute Percentage Error is: 0.0030626082548832613
MAE: 0.007540671798947956
MSE: 0.0001023022184435061
RMSE: 0.01011445591435872
Coefficient of determination score on training :  0.999304406775328
Coefficient of determination score on testing :  0.5676438511319499
```



- **Model 7: XGBoost Regressor**

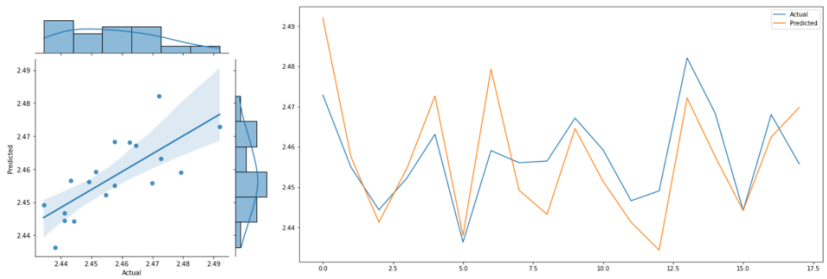
1) learning_rate parameter = .2 , max depth = 7

```
Mean Absolute Percentage Error is: 0.003071023496432224
MAE: 0.007561733216447066
MSE: 8.78334930698665e-05
RMSE: 0.009371952468395606
Coefficient of determination score on training :  0.993266222437032
Coefficient of determination score on testing :  0.6287924994873204
```



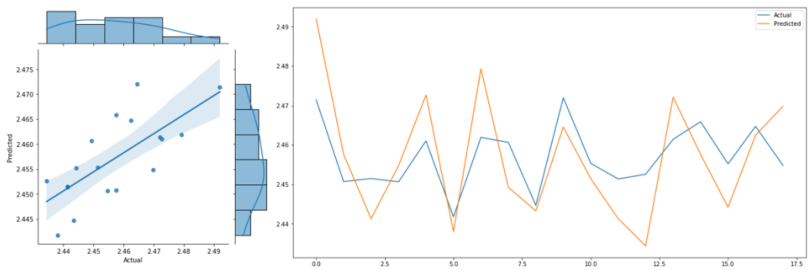
2) learning_rate parameter = .2, max depth = 3

```
Mean Absolute Percentage Error is: 0.0033726061275773916
MAE: 0.008305781120839942
MSE: 0.00010366638826497449
RMSE: 0.010181669227831676
Coefficient of determination score on training :  0.888531410622931
Coefficient of determination score on testing :  0.56187850977977
```



- **Model 8: Bayesian Ridge Regressor**

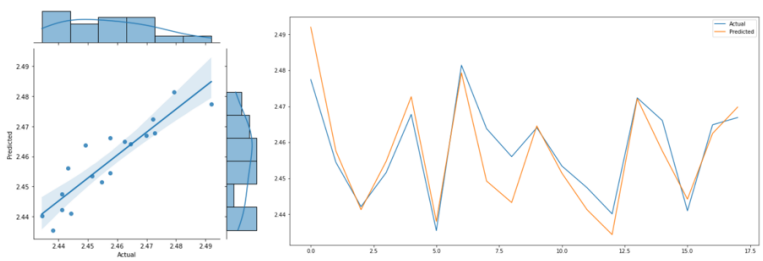
```
Mean Absolute Percentage Error is: 0.003929288619808761
MAE: 0.009668162024208306
MSE: 0.00012274649446841946
RMSE: 0.011079101699525078
Coefficient of determination score on training :  0.53620444079331
Coefficient of determination score on testing :  0.481240950168387
```



- **Model 9: Cat Boost Regressor**

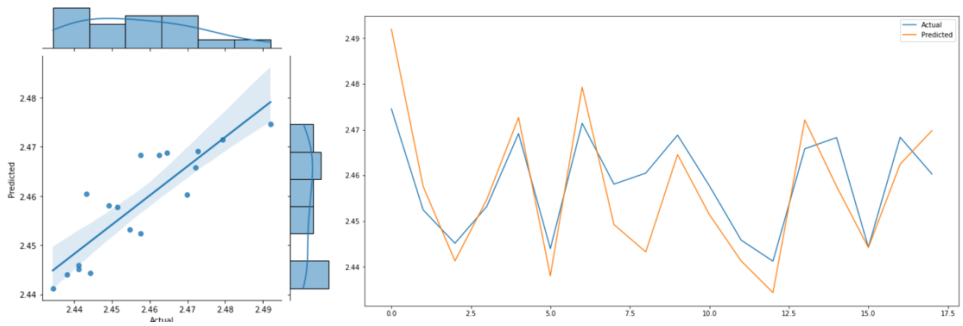
1) *learning_rate parameter = 1 , iterations = 10*

```
Mean Absolute Percentage Error is: 0.0020381778044328484
MAE: 0.005008949376018047
MSE: 4.5164057066763305e-05
RMSE: 0.006720420899524323
Coefficient of determination score on training :  0.9293989558485893
Coefficient of determination score on testing :  0.8091247865614379
```



2) *learning_rate parameter = 1, iterations= 4*

```
Mean Absolute Percentage Error is: 0.0028415132390409185
MAE: 0.006989952367246828
MSE: 6.869676631077182e-05
RMSE: 0.008288351242000535
Coefficient of determination score on training :  0.7861479247739179
Coefficient of determination score on testing :  0.7096693524958541
```

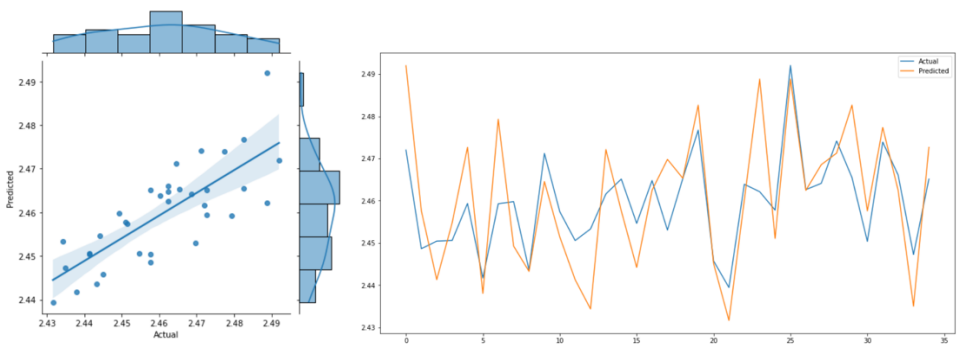


- ***Experiment 2:***

  Random State = 40 and Test Size equals to 20% of the data, x_train shape is (137,10) and has a size of 1370, and x_test shape is (35,10) and has a size of 35.

  - **Model 1: Linear regression**

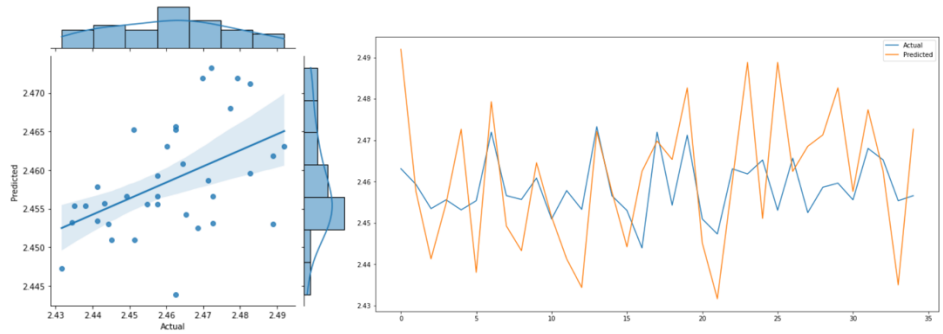
```
Mean Absolute Percentage Error is: 0.0033845638343474355
MAE: 0.008339699476390372
MSE: 0.00011007632488887511
RMSE: 0.010491726497042853
Coefficient of determination score on training :  0.5125661299338486
Coefficient of determination score on testing :  0.5952729035341477
```



  - **Model 2: K Neighbours Regressor**

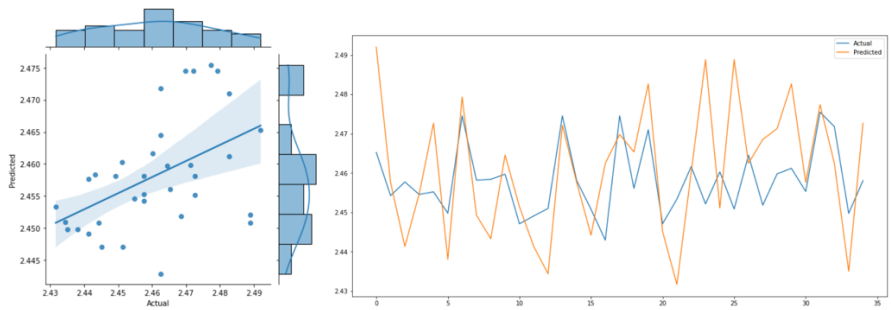
    1) N_neighbors parameter = 7

```
Mean Absolute Percentage Error is: 0.004730504353427449
MAE: 0.011656027515365507
MSE: 0.00021358545597918634
RMSE: 0.014614563147052544
Coefficient of determination score on training :  0.2724289767289565
Coefficient of determination score on testing :  0.2146919736541365
```



2) N_neighbors parameter = 5

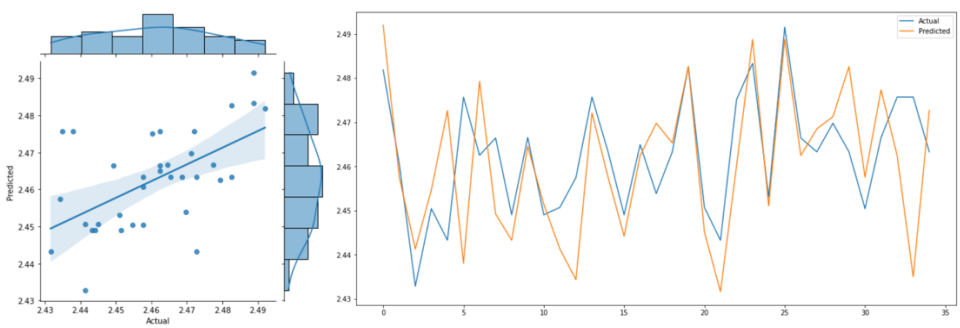
```
Mean Absolute Percentage Error is: 0.004595258302479799
MAE: 0.011327291530543678
MSE: 0.00021645593608542586
RMSE: 0.014712441540595016
Coefficient of determination score on training :  0.36887475315678675
Coefficient of determination score on testing :  0.20413783242498984
```



- **Model 3: Decision Tree Regressor**

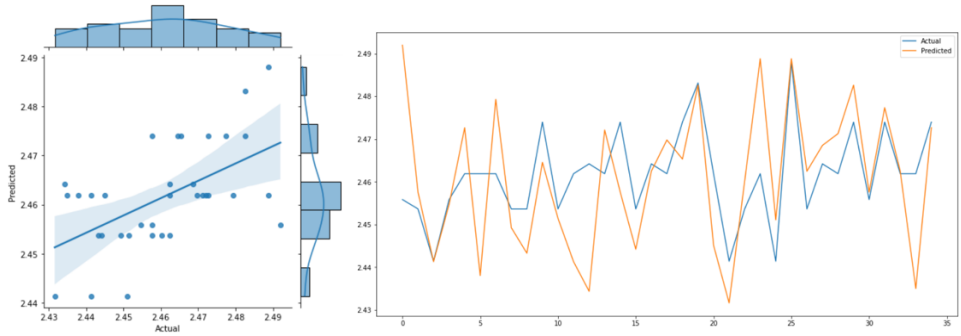
1) Max_depth parameter = 7

```
Mean Absolute Percentage Error is: 0.0041632980028409265
MAE: 0.01022588365934992
MSE: 0.00020002233894050734
RMSE: 0.014142925402493903
Coefficient of determination score on training :  0.9364724142584109
Coefficient of determination score on testing :  0.26456065326020906
```



2) Max_depth parameter = 3

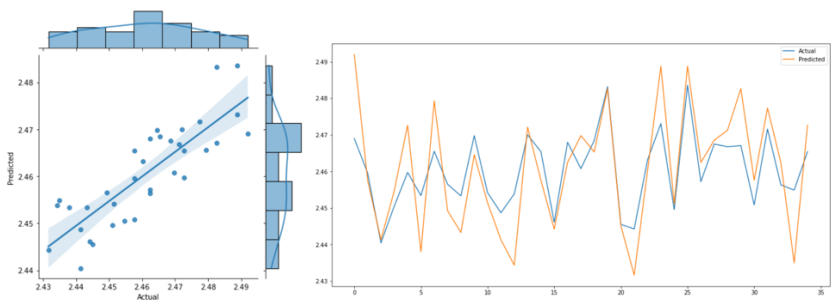
```
Mean Absolute Percentage Error is: 0.00418849476986167
MAE: 0.010300222035638105
MSE: 0.00019100984985611722
RMSE: 0.013820631311778677
Coefficient of determination score on training :  0.540000644941838
Coefficient of determination score on testing :  0.297697647457116
```



- **Model 4: Random Forest regressor**

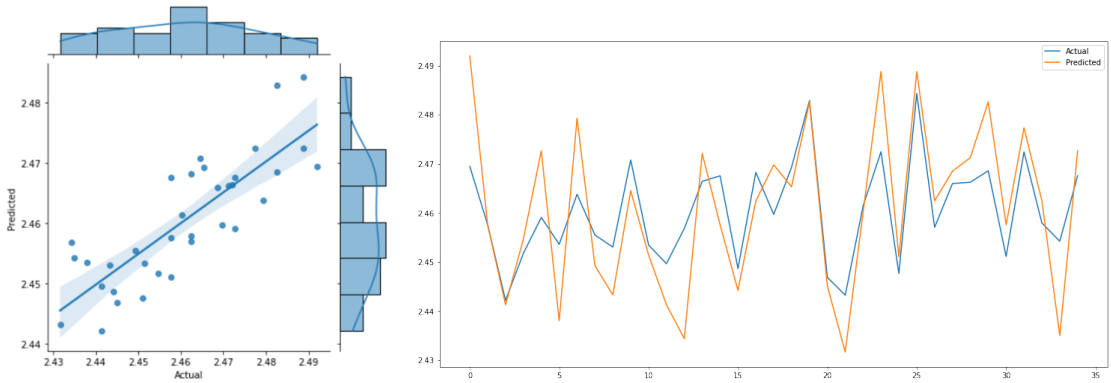
1) Max_depth parameter = 7

```
Mean Absolute Percentage Error is: 0.0030740979226245293
MAE: 0.007565870839924876
MSE: 9.300984834053094e-05
RMSE: 0.009644161360145885
Coefficient of determination score on training :  0.9023014853639911
Coefficient of determination score on testing :  0.6580226865346663
```



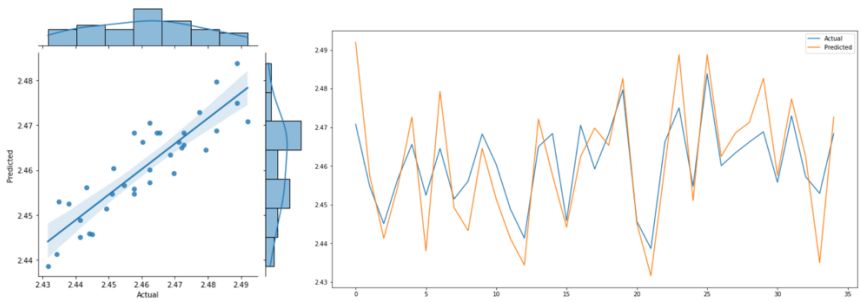
2) Max_depth parameter = 5

```
Mean Absolute Percentage Error is: 0.0031626864133774994
MAE: 0.007782821997392185
MSE: 9.757958719899162e-05
RMSE: 0.0098782380614658
Coefficient of determination score on training :  0.8109662285336341
Coefficient of determination score on testing :  0.6412207344195211
```



- **Model 5: Ada Boost regressor**
  3) *Learning_rate parameter = 1*

```
Mean Absolute Percentage Error is: 0.002883027917224329
MAE: 0.0070997279360416774
MSE: 7.531744271207511e-05
RMSE: 0.008678562249132923
Coefficient of determination score on training :  0.7707526010162792
Coefficient of determination score on testing :  0.7230738768495492
```

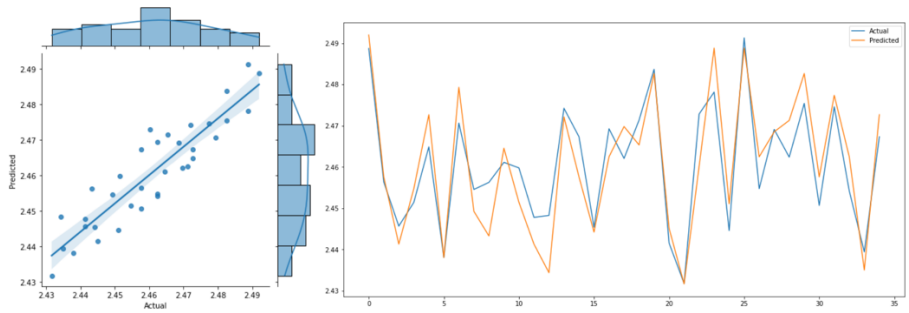


  1) *learning_rate parameter = .2*

```
Mean Absolute Percentage Error is: 0.0032180379734173562
MAE: 0.007923940549276231
MSE: 9.28962097870024e-05
RMSE: 0.00963826798688449
Coefficient of determination score on training :  0.7282485023671406
Coefficient of determination score on testing :  0.6584405111837238
```

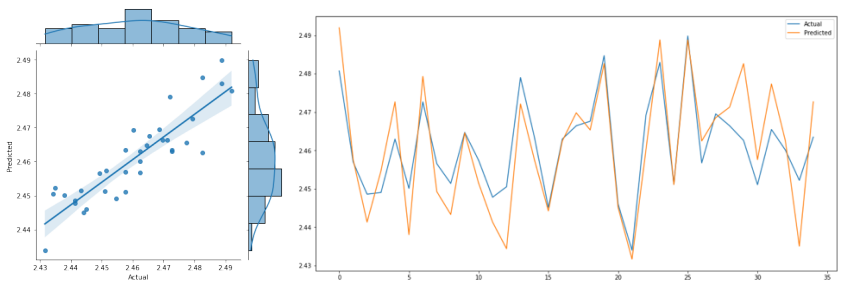
- **Model 6: Gradient Boosting Regressor**
  1) *learning_rate parameter = .2 , max depth = 2*

```
Mean Absolute Percentage Error is: 0.0023429231839951414
MAE: 0.00576519792295876
MSE: 4.6662751052815305e-05
RMSE: 0.00683101391103951
Coefficient of determination score on training :  0.8999104939057142
Coefficient of determination score on testing :  0.8284310475862845
```



2) learning_rate parameter = .2, max depth = 5

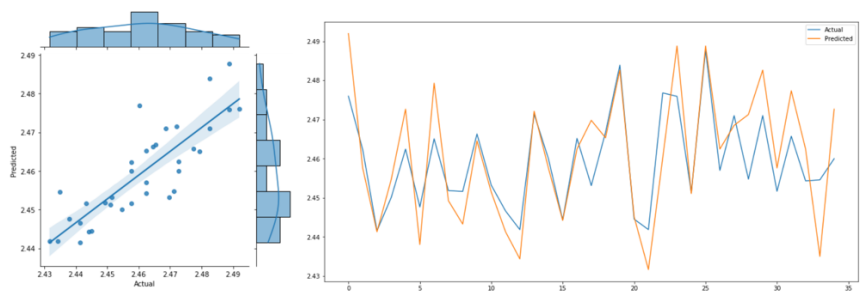
```
Mean Absolute Percentage Error is: 0.00315523560189827
MAE: 0.007772911701849061
MSE: 9.165203728166171e-05
RMSE: 0.009573507052363919
Coefficient of determination score on training :  0.7379719240349228
Coefficient of determination score on testing :  0.6630150673028357
```



- **Model 7: XGBoost Regressor**

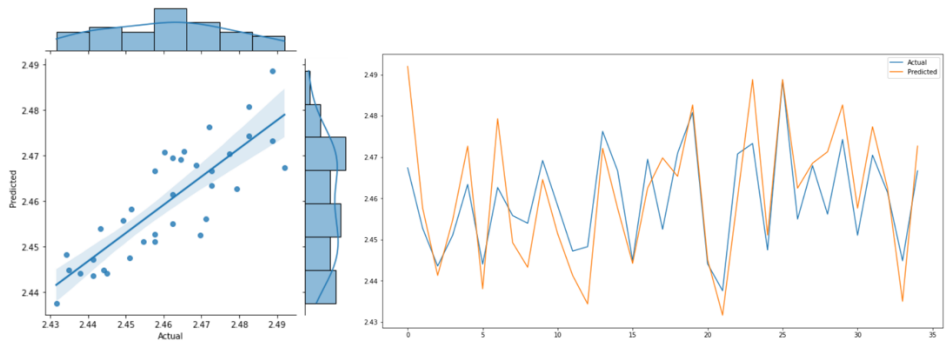
1) learning_rate parameter = .2 , max depth = 7

```
Mean Absolute Percentage Error is: 0.0028760345124156732
MAE: 0.0070847738857706104
MSE: 8.385810688952446e-05
RMSE: 0.00915740721435519
Coefficient of determination score on training :  0.9946992671196149
Coefficient of determination score on testing :  0.6916716818914377
```



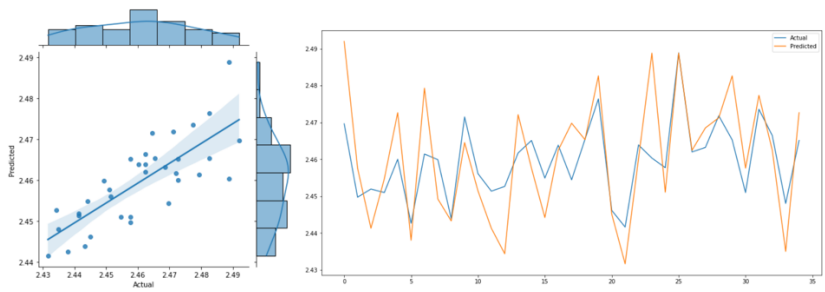
2) learning_rate parameter = .2, max depth = 3

```
Mean Absolute Percentage Error is: 0.0030158486458301587
MAE: 0.007431367610104278
MSE: 8.469220440313052e-05
RMSE: 0.009202836758474558
Coefficient of determination score on training :  0.883231168198469
Coefficient of determination score on testing :  0.6886048837839214
```



- **Model 8: Bayesian Ridge Regressor**

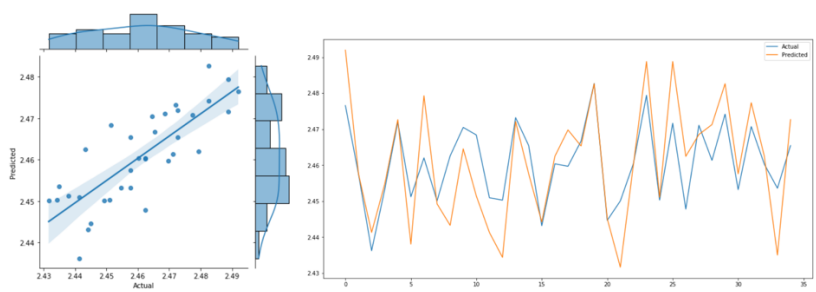
```
Mean Absolute Percentage Error is: 0.003375861789883962
MAE: 0.008316443048038698
MSE: 0.0001131913198500929
RMSE: 0.010639140935719054
Coefficient of determination score on training :  0.5082306696070706
Coefficient of determination score on testing :  0.5838197334952335
```



- **Model 9: Cat Boost Regressor**

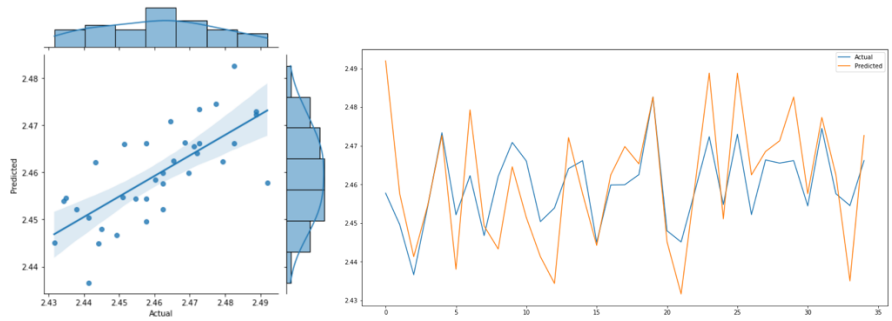
1) learning_rate parameter = 1 , iterations = 10

```
Mean Absolute Percentage Error is: 0.0030740979226245293
MAE: 0.007565870839924876
MSE: 9.300984834053094e-05
RMSE: 0.009644161360145885
Coefficient of determination score on training :  0.9133104823843498
Coefficient of determination score on testing :  0.6289065511433243
```



2) learning_rate parameter = .2, iterations= 4

```
Mean Absolute Percentage Error is: 0.003575351103555636
MAE: 0.008802807154608568
MSE: 0.00013380256743931331
RMSE: 0.01156730597154382
Coefficient of determination score on training :  0.759685738226371
Coefficient of determination score on testing :  0.5080365857588356
```

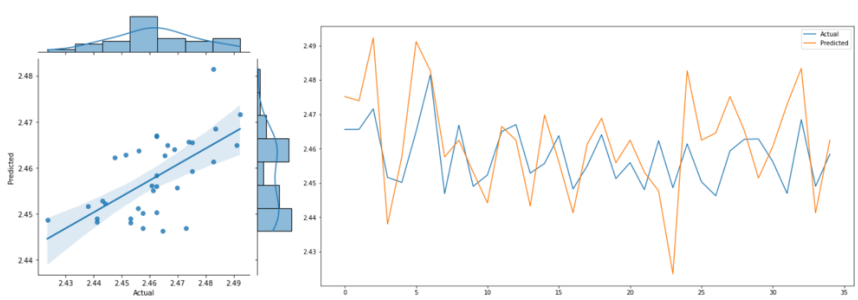


- ***Experiment 3:***

  Random State = 42 and Test Size equals to 20% of the data, x_train shape is (137,10) and has a size of 1370, and x_test shape is (35,10) and has a size of 35.

  - **Model 1: Linear regression**

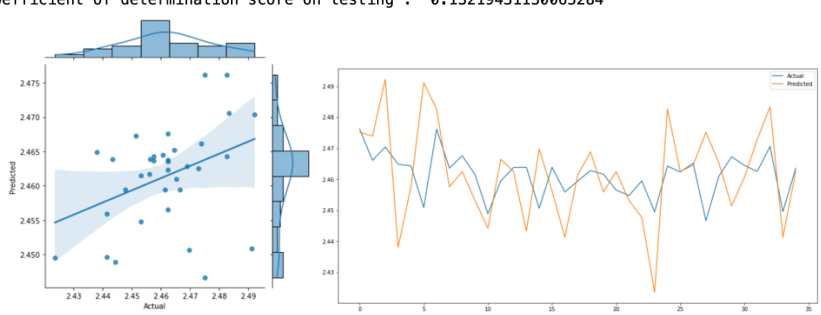
```
Mean Absolute Percentage Error is: 0.00422156904850834
MAE: 0.010399996152817234
MSE: 0.00015579246346775167
RMSE: 0.012481685121318822
Coefficient of determination score on training :  0.5830022959278136
Coefficient of determination score on testing :  0.31717627652551095
```



  - **Model 2: K Neighbours Regressor**

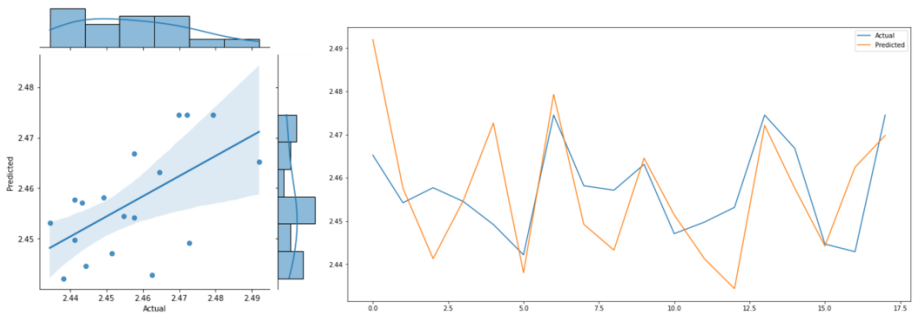
    3) N_neighbors parameter = 7

```
Mean Absolute Percentage Error is: 0.004276044336623497
MAE: 0.010530503243835008
MSE: 0.0001979977868458626
RMSE: 0.014071168638242617
Coefficient of determination score on training :  0.3386887996114223
Coefficient of determination score on testing :  0.13219431130065284
```



2) N_neighbors parameter = 5

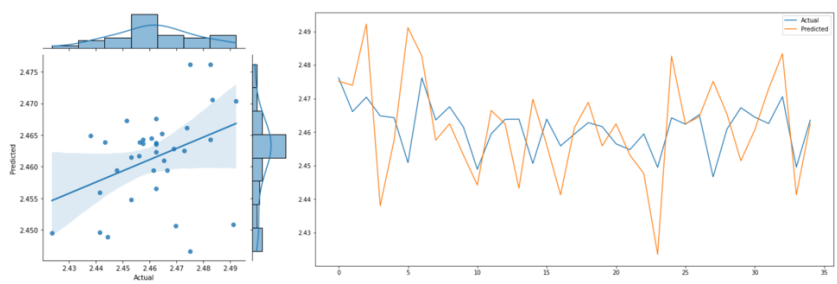
```
Mean Absolute Percentage Error is: 0.004485161271453648
MAE: 0.011044821389370378
MSE: 0.00021543378160024672
RMSE: 0.014677662674971337
Coefficient of determination score on training :  0.4170411772417437
Coefficient of determination score on testing :  0.055773985210009425
```



- **Model 3: Decision Tree Regressor**

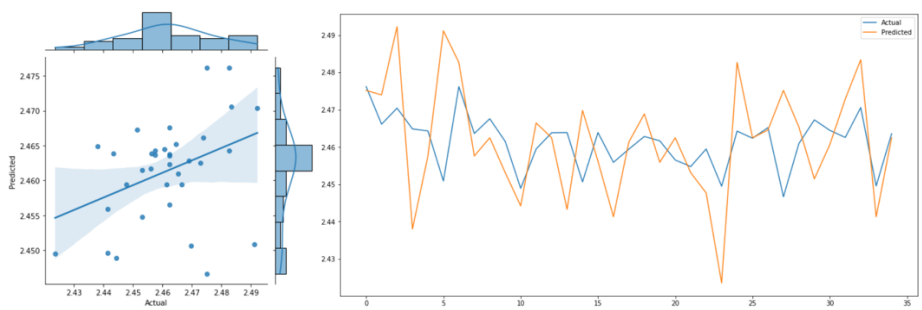
2) Max_depth parameter = 7

```
Mean Absolute Percentage Error is: 0.004852027709551641
MAE: 0.011956292863293009
MSE: 0.000246316608703886
RMSE: 0.01569447701275471
Coefficient of determination score on training :  0.903722250573327
Coefficient of determination score on testing :  -0.07958254311583524
```



3) Max_depth parameter = 3

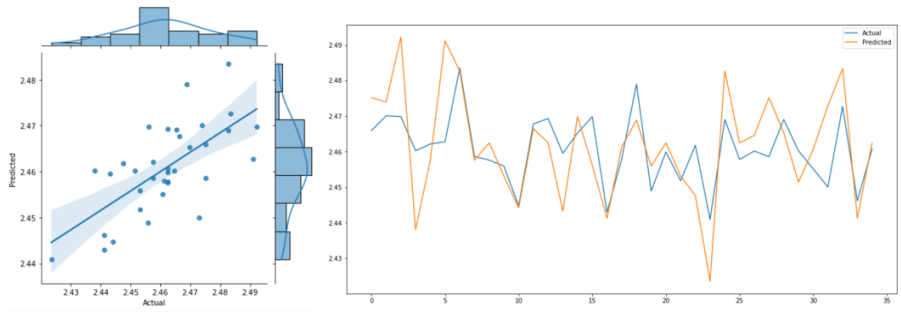
```
Mean Absolute Percentage Error is: 0.0034784029607937634
MAE: 0.008544441727797633
MSE: 0.00011739860663290257
RMSE: 0.010835063757675935
Coefficient of determination score on training :  0.8035668469645441
Coefficient of determination score on testing :  0.5038425342231778
```



- **Model 4: Random Forest regressor**

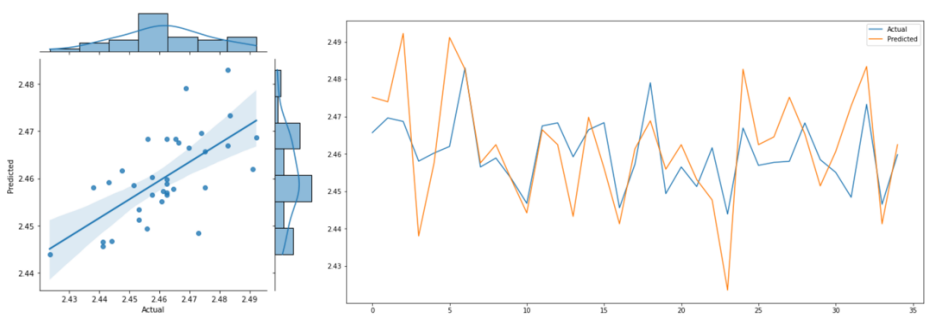
4) Max_depth parameter = 7

```
Mean Absolute Percentage Error is: 0.0034568638549379066
MAE: 0.008518054182152771
MSE: 0.0001265955315120951
RMSE: 0.01125146797142911
Coefficient of determination score on training :  0.9066725755707622
Coefficient of determination score on testing :  0.4451436848855401
```



5) Max_depth parameter = 5

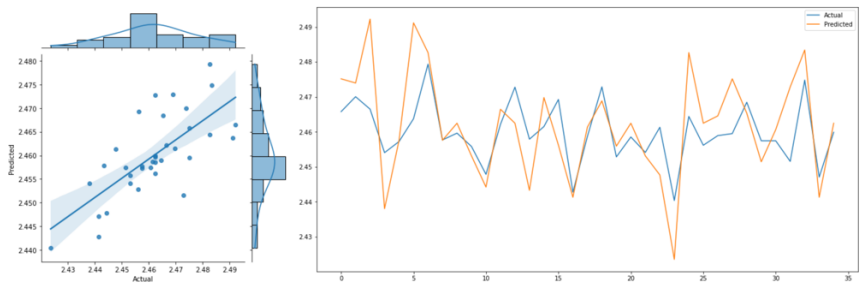
```
Mean Absolute Percentage Error is: 0.0035458811627419704
MAE: 0.008737310649089152
MSE: 0.00013319084584239755
RMSE: 0.011540833845194963
Coefficient of determination score on training :  0.8299226352059754
Coefficient of determination score on testing :  0.4162370421105257
```



- **Model 5: Ada Boost Regressor**

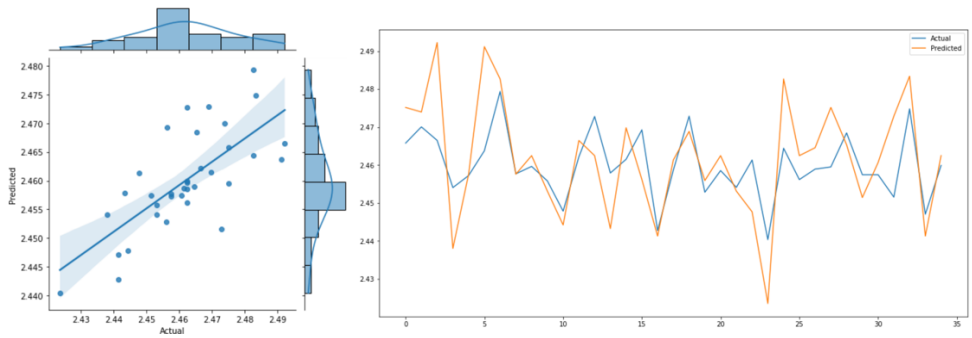
6) learning_rate parameter = 1

```
Mean Absolute Percentage Error is: 0.003243152947396894
MAE: 0.007997085565017744
MSE: 0.00011329207261582913
RMSE: 0.010643874887268694
Coefficient of determination score on training :  0.8042304322038119
Coefficient of determination score on testing :  0.5034514947528543
```



7) learning_rate parameter = .2

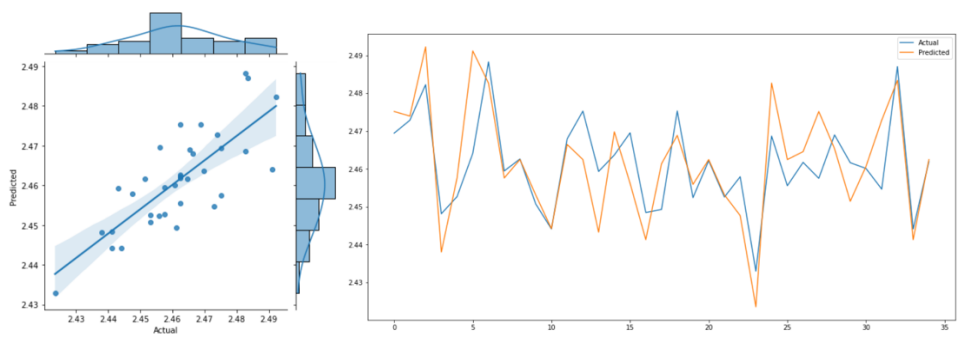
```
Mean Absolute Percentage Error is: 0.0033476486004849018
MAE: 0.008253540519247362
MSE: 0.00011950125010625208
RMSE: 0.010931662732917262
Coefficient of determination score on training :  0.7782714912165163
Coefficient of determination score on testing :  0.4762372534516234
```



- **Model 6: Gradient Boosting Regressor**

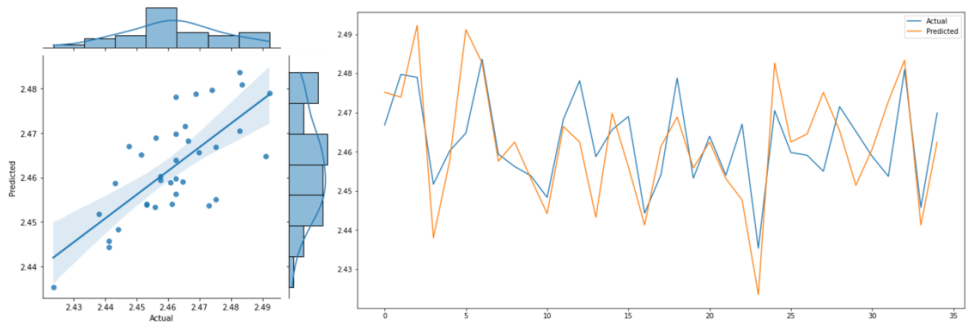
2) learning_rate parameter = .2 , max depth = 2

```
Mean Absolute Percentage Error is: 0.0028914271072212755
MAE: 0.00712629741816181
MSE: 8.997702111394488e-05
RMSE: 0.009485621809557077
Coefficient of determination score on training :  0.9218581265625796
Coefficient of determination score on testing :  0.605639173958604
```



2) learning_rate parameter = .2, max depth = 5

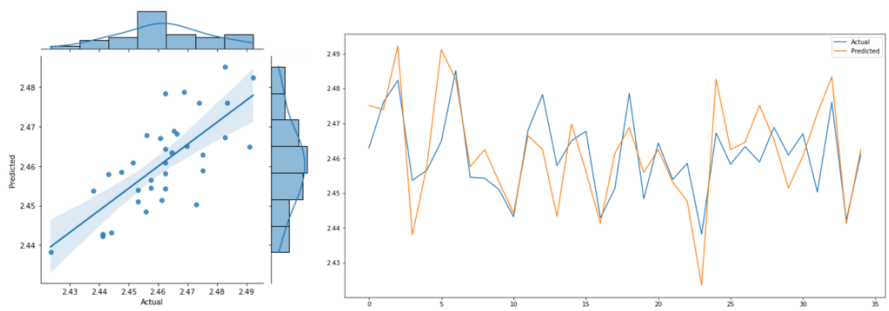
```
Mean Absolute Percentage Error is: 0.0033135482256331727
MAE: 0.008162729074961777
MSE: 0.00010997906103262921
RMSE: 0.010487090208090575
Coefficient of determination score on training :  0.99613002352936
Coefficient of determination score on testing :  0.5179721131114114
```



- **Model 7: XGBoost Regressor**

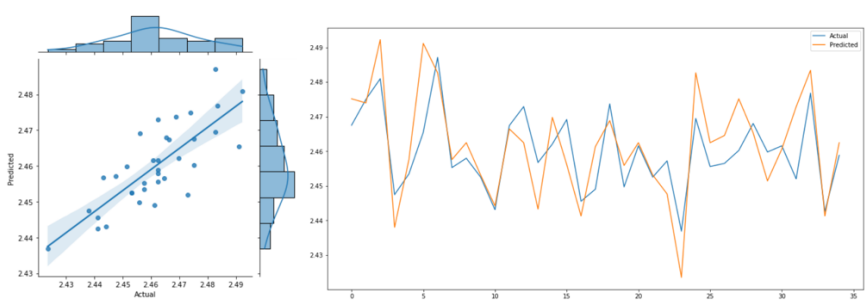
2) learning_rate parameter = .2 , max depth = 7

```
Mean Absolute Percentage Error is: 0.0032280873044831988
MAE: 0.007954885041426987
MSE: 0.00010630873216232232
RMSE: 0.01031061259878977
Coefficient of determination score on training :  0.9916143145831968
Coefficient of determination score on testing :  0.5340588195528795
```



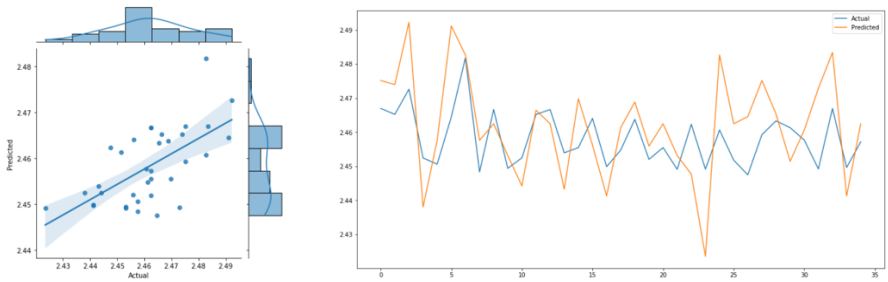
2) learning_rate parameter = .2, max depth = 3

```
Mean Absolute Percentage Error is: 0.0029794962163567063
MAE: 0.007343916078923637
MSE: 8.855129749958714e-05
RMSE: 0.009410169897487884
Coefficient of determination score on training :  0.9047065582512226
Coefficient of determination score on testing :  0.6118879865476852
```



- **Model 8: Bayesian Ridge Regressor**

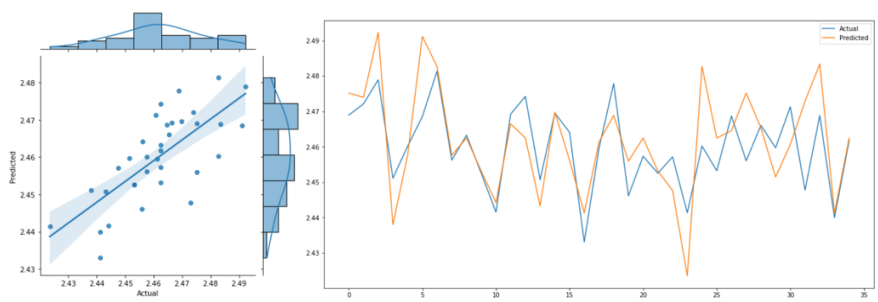
```
Mean Absolute Percentage Error is: 0.004160412498074075
MAE: 0.010248496920447967
MSE: 0.0001528929856357876
RMSE: 0.012364990320893405
Coefficient of determination score on training :  0.578814393349501
Coefficient of determination score on testing :  0.32988441532301604
```



- **Model 9: Cat Boost Regressor**

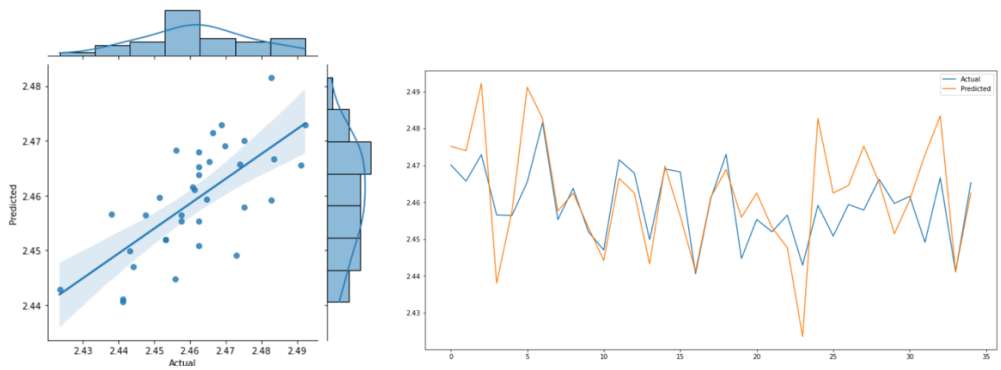
3) learning_rate parameter = 1 , iterations = 10

```
Mean Absolute Percentage Error is: 0.0031813069115030155
MAE: 0.007841680167628252
MSE: 0.00011029703556714082
RMSE: 0.010502239550074l1
Coefficient of determination score on training :  0.9397808051198466
Coefficient of determination score on testing :  0.5165784606150554
```



4) learning_rate parameter = .2, iterations= 4

```
Mean Absolute Percentage Error is: 0.003249054866832297
MAE: 0.008012203933383942
MSE: 0.00012321467967580808
RMSE: 0.011100210794206O3
Coefficient of determination score on training :  0.7970910132628615
Coefficient of determination score on testing :  0.4599616406966498
```



4.1. Conclusion

Regression models has different results when applied on the same data, also, each algorithm gives another prediction results based on the test size, random state, and parameters. In Addition, from the previous experiments, we conclude that the best possible results on the test size 10% and random state 40 is from the Catboost regressor when its learning rate equals to 1 and the number of iteration parameter is 10 as it has the minimum possible number applied errors and has the highest coefficient of determination on both the training data and the testing data (training score = 0.92, testing score: 0.809). On the other hand, the same model behaved in a different way when the test size changed to 20% of the data and its error percentage increased and the score decreased, as same as some other models which miss predicted a lot of values such as decision trees. Furthermore, when the test size is 20% and the random state is 40, Models like Gradient Boosting regressor and Ada Boost Regressor had better prediction than the others (training score = 0.89, testing score: 0.82) and better than their predictions when the test size was 10% with the best linear line fit. We also conclude that each model parameter changes the whole predictions such as the learning rate and the maximum depth of the model, along with changing the random state to 42 in the last experiment – when worst possible predictions happened-.

All in all, rental prices are so important to be predicted to avoid scams and to help homeowners pricing their properties, having a good prediction model is necessarily in such cases and also having more properly collected data is a huge plus.

5. Sources

[1] *Web Site for data Collection*. (2022). Sahibinden. https://www.sahibinden.com

[2] *Location coordinates finder Site*. (2022). Latlong. https://www.latlong.net

[3] *Pandas Documentation*. (2022). Pandas https://pandas.pydata.org/docs/user_guide/index.html

[4] *numpy Documentation*. (2022). Numpy https://numpy.org/doc/

[5] *Plotly Documentation*. (2022). Plotly https://plotly.com

[6] *Scikit-learn ML Documentation*. (2022). Plotly https://scikit-learn.org/stable/

[7] *Some code lines from our previous projects* . (2021). Github

https://github.com/omniaelmenshawy/California_housing

[8] *machine Learning Mastry*. (2022). *Regression metrics for machine learning*

https://machinelearningmastery.com/regression-metrics-for-machine-learning/

[9] *Towards data science*. (2022). *Best metrices to evaluate regression models*

https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b